Artificial Societies for Ambient Intelligence

The vision of Ambient Intelligence (AmI) is a society based on unobtrusive, often invisible interactions amongst people and computer-based services in a global computing environment. Services in AmI will be ubiquitous in that there will be no specific bearer or provider but, instead, they will be associated with a variety of objects and devices in the environment, which will not bear any resemblance to computers. People will interact with these services through intelligent and intuitive interfaces embedded in these objects and devices, which in turn will be sensitive to what people need.

For a large class of the envisaged AmI applications, the added value of these new services is likely to be for people in ordinary social contexts. Such applications beg for technologies that are transparent, so that their functional behaviour can be understood easily. Put simply, transparency should bring AmI interactions closer to the way people think rather than the way machines operate.

Another challenge posed by the AmI vision is that the electronic part of the ambience will often need to act intelligently on behalf of people. The conceptual components of ambience will need to be both reactive and proactive, behaving as if they were agents that act on behalf of people. It would be more natural, in other words, to use the agent metaphor in order to understand components of an intelligent ambience. An agent in this context can be a software (or hardware) entity that can sense and affect the environment, has knowledge of the environment and its own goals, and can proactively plan to achieve its goals or those of its user(s), so that the combined interactions of the electronic and physical environment provide a desirable outcome for one or more people.

If we assume that agents are abstractions for the interaction within an ambient intelligent environment, one aspect that we need to ensure is that their behaviour is regulated and coordinated, so that the system as a whole functions effectively. For this purpose, we need rules that take into consideration the social context in which these interactions take place, and the whole system begs for an organisation similar to that envisaged by artificial agent societies. The society is there not only to regulate behaviour but also to distribute responsibility amongst the member agents.

This symposium was proposed to help present current research and develop scenarios for the use of agent societies for AmI, and establish a body of knowledge and a theoretical framework in this context. We would like to thank the AISB 2007 organisers and the ASAMI 2007 Programme Committee members for their help and support.

Fariba Sadri and Kostas Stathis (Workshop Chairs)

Programme committee: Alexander Artikis (NCSR Demokritos, Greece); Juan Carlos Augusto (University of Ulster at Jordanstown, UK); Cristiano Castelfranchi (CNR, Italy); Oscar DeBruijn (University of Manchester, UK); Paul J. Feltovitch (IHMC, USA); Marie-Pierre Gleizes (IRIT, France); Gregory O'Hare (University College Dublin, Ireland); Andrea Omicin (University of Bologna, Italy); Paolo Pett (Medical Univ. of Vienna, Austria); Jeremy Pitt (Imperial College, UK); Eric Platon (NII, Japan); Harmut Raffler (Siemens AG, Germany); Alessandro Ricci (University of Bologna, Italy); Nicolas Sabouret (Laboratoire d'Informatique de Paris 6, France); Fariba Sadri (Imperial College, UK); Rob Saunders (University of Sydney, Australia); Daniel Shapiro (Stanford University, USA); Maarten Sierhuis (NASA-Ames, USA); Kostas Stathis (Royal Holloway - U. of London, UK); Francesca Toni (Imperial College, UK); George Vouros (University of Aegean, Greece); Pinar Yolum (Bogazici University, Turkey); Franco Zambonelli (Univ. of Modena, Italy)

Ambient Intelligence as a Never-Ending Self-Organizing Process: Analysis and Experiments

Jean-Pierre Georgé and **Valérie Camps** and **Marie-Pierre Gleizes** and **Pierre Glize**[**]

Abstract. Our team has been working for several years on building adaptive systems using self-organising mechanisms following a specific approach we called the AMAS[1] Theory. Its main originality is that it enables artificial systems to show relevant emergent behaviours by focusing on local cooperative interactions among the agents. This article aims at showing the relevance of this approach specifically, and more generally of any self-organisation approach, for Ambient Intelligence. For this we analyse and discuss AmI problems in the light of our experience in complex systems, as well as show encouraging results in a first experiment in a kind of AmI system (a service providing network of agents).

1. INTRODUCTION

AgentLink Roadmap [Luck, 2005] asserts that agents have their place in spreading fields such as Web Services, Semantic Web, Peer-to-Peer, Grid Computing, Ambient Intelligence, Self* Systems, etc..

"*If we assume that agents are abstractions for the interaction within an ambient intelligent environment, one aspect that we need to ensure is that their behaviour is regulated and coordinated, so that the system as a whole functions effectively. For this purpose, we need rules that take into consideration the social context in which these interactions take place, and the whole system begs for an organisation similar to that envisaged by artificial agent societies. The society is there not only to regulate behaviour but also to distribute responsibility amongst the member agents*"[2].

According to the previous citation, we consider that the central problem in Ambient Intelligence societies, which are highly open and dynamic, is to find generic local rules followed by the agents encapsulating numerous types of objects and devices in order to guarantee an efficient and relevant collective behaviour. To face the dynamics and the heterogeneity of AmI systems (such as workload, failures and interoperability of the devices, as well as their addition or suppression), these agents must enable the system to adapt in every context.

We already have proposed the AMAS (Adaptive Multi-Agent Systems) theory [Georgé, 2003b] to solve complex systems such as timetabling management [Picard, 2005] or aircraft design optimisation [Welcomme, 2006]… This way to design artificial systems follows a process defined in the ADELFE methodology [Bernon, 2005]. Our challenge is now to apply it to ambient systems. We propose in this paper to show how the AMAS theory can tackle these real-time adaptation problems to design system where each agent encapsulates a device. AMAS approach allows the design of complex systems that can be underspecified and for which an *a priori* known algorithmic solution does not exist.

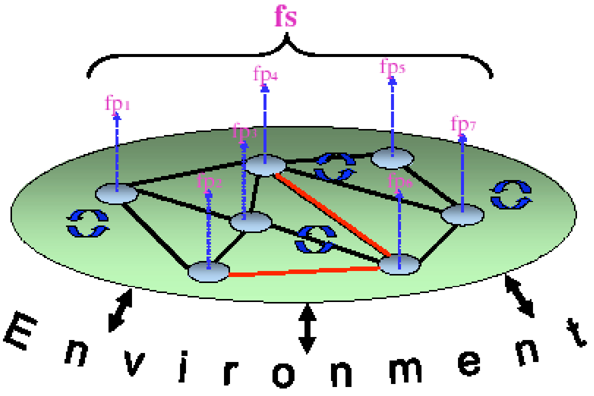


Figure 1: Adaptation - Changing the function of the system by changing the organization.

The paper is structured as follows. Section 2 shows how ambient systems can be studied as emergent systems. Then, the AMAS theory dedicated to design system with emergent functionality is presented and its use for ambient systems design is justified. Section 3 details preliminary experiments on a kind of AmI system composed of a large number of devices and some results are described. The paper ends with a discussion highlighting the advantages of self-organising systems for designing ambient systems.

2. AMBIENT INTELLIGENCE, A CLASS OF EMERGENT PHENOMENA

Let us consider for instance a room with a large number of electronic equipments controlled each by an autonomous microchip. This room has a goal: the satisfaction of the users living in it from day to day. The goal itself "user satisfaction" is really imprecise and incomplete, and the way to reach it even more. The specifications of this kind of software are incomplete and the system is often underspecified. Nevertheless, suppose that we are able to define a learning algorithm in order to assign correct behaviours to the objects and devices. These objects could move to other ambient environments such as an airport, a rescue service or a classroom. The interactions of a given device would be quite different with its new neighbours. Because these new situations are quite different from the prior, they need at least new learning algorithms and new cost

[1] Adaptive Multi-Agent Systems [Georgé, 2003b]

[2] Extract from this ASAmI Symposium call

** IRIT – Paul Sabatier University - 118, Route de Narbonne
31062 Toulouse, Cedex 9, France
email: {george, camps, gleizes, glize}@irit.fr

function for the evaluation of local device behaviours. The devices composing the society have their own behaviour, their own objectives and can be described by the result of the actions they can do such as: order coffee at a shop, make coffee, play music,… The humans evolving in the system are also an important factor. At the society level, we obtain a new function such as the harmony inside the house which can be qualified as emergent phenomena. An emergent phenomena is observed when there is at least two levels, the micro and the macro levels and when to describe the phenomena at the macro level, we cannot use the theory (or the vocabulary) used to described the parts at the micro levels [Müller, 2004]. More precisely in computer science, we have defined the behaviour of a MAS as emergent when the global function of the system (macro level) is not coded inside the agents (micro level) [Georgé, 2005].

Consequently, we fall in the limitations quoted by Wolpert and MacReady in their "No Free Lunch Theorems" because we must have a learning algorithm universally optimal on all fields. Under these conditions, Wolpert and Macready [Wolpert, 1997] proved that performances of all the optimisation methods using cost functions are equivalent, including the random[3]. This is very unsatisfactory!

A way to tackle the limitations of these theorems is to find a relevant learning algorithm which does not need a cost function derived from the global criteria to optimise and to design underspecified systems. We showed in previous works[4] [Picard, 2004], [Georgé, 2003a], [Gleizes, 2002], that algorithms, which do not directly depend on the global function to obtain, are a way to dynamically implement systems able to self-adapt to their contexts. In the case of AmI, these algorithms should use local adaptive behaviour and thus only take into account local knowledge resulting from representation of their neighbourhood. The concept of agent thus becomes natural for a local emergent solving. Because these local behaviour have, neither explicitly nor implicitly, information about the global goal to achieve, the collective behaviour of an ambient intelligence society can be qualified as an emergent phenomena [Di Marzo, 2005]. So, the AMAS theory dedicated to design systems with emergent functionality can be a good candidate for ambient system design.

2.1. Agents' Self-Organisation by AMAS Approach

We consider an AmI system as a multi-agent system S having a global function fs to achieve. Each part Pi realizes only a partial function fpi (Figure 1). fs is the result of the combination of the partial functions fpi, noted by the operator "●". The combination being determined by the current organization of the parts, we can deduce fs = fp1 ● fp2 ● ... ● fpn. As generally, fp1 ● fp2 ≠ fp2 ● fp1, by transforming the multi-agent organization, the combination of the partial functions is changed and therefore the global function fs changes. This is a powerful way to adapt the system to the environment. A pertinent technique to build this kind of systems is to use adaptive multi-agent systems. As in Wooldridge's definition of multi-agent systems [Wooldridge 2002], we will be referring to AmI systems constituted by several autonomous agents, plunged into a common environment and trying to solve a common task.

2.2. The Theorem of Functional Adequacy

Cooperation was extensively studied in computer science by [Axelrod, 1984] and [Huberman, 1991] for instance. "*Everybody will agree that co-operation is in general advantageous for the group of co-operators as a whole, even though it may curb some individual's freedom*" [Heylighen, 1992]. Relevant biological inspired approaches using cooperation are for instance Ants Algorithms [Dorigo, 1999] which give efficient results in many domains. In order to show the theoretical improvement coming from cooperation, we have developped the AMAS (Adaptive Multi-Agent System) theory [Georgé, 2003b] which is based upon the following theorem. This theorem describes the relation between cooperation in a system and the resulting functional adequacy[5] of the system.

Theorem. *For any functionally adequate system, there is at least a cooperative internal medium system that fulfills an equivalent function in the same environment.*

Definition. *A cooperative internal medium system is a system where no Non-Cooperative Situations exist.*

Definition. *An agent is in a Non-Cooperative Situation (NCS) when: (1) a perceived signal coming from the environment is not understood or is ambiguous; (2) perceived information does not produce any activity of the agent; (3) the conclusions are not useful to others.*

The cooperation failures are called "Non Cooperative Situations" (NCS) and can be assimilated to "exceptions" in traditional programming. Our definition of cooperation is based on three local meta-rules the designer has to instantiate according to the problem to solve:

- Meta-rule 1 (c_{per}): Every signal perceived by an agent must be understood without ambiguity.
- Meta-rule 2 (c_{dec}): Information coming from its perceptions has to be useful to its reasoning.
- Meta-rule 3 (c_{act}): This reasoning must lead the agent to make actions which have to be useful for other agents and the environment.

The theorem of functional adequacy means that we only have to use (and hence understand) a subset of particular systems (those with cooperative internal mediums) in order to obtain a functionally adequate system in a given environment. We concentrate on a particular class of such systems, those with the following properties [Gleizes, 2002]:

- The system is plunged into an environment.
- The system is composed of interacting parts called agents.
- The system is cooperative and functionally adequate with respect to its environment. The agents do not 'know' the global function the system has to achieve via adaptation.
- The system adapts itself to its environment. It has not an explicitly defined goal, rather it acts using its perceptions of

[3] It is not contradictory with many existing applications showing in practice that very good algorithms exist in specific limited contexts.

[4] For more details see also the website www.irit.fr/SMAC

[5] "Functional" refers to the "function" the system is producing, in a broad meaning, i.e. what the system is doing, what an observer would qualify as the behavior of a system. And "adequate" simply means that the system is doing the "right" thing, judged by an observer or the environment. So "functional adequacy" can be seen as "having the appropriate behavior for the task".

the environment as a feedback in order to adapt its behaviour that leads to have an adequate global function. The adaptation is realized at the agent level by their specific behaviour. It consists in trying and maintaining cooperation using their skills, representations of themselves, of the other agents and of the environment.

- Each agent only evaluates whether the changes taking place in the system are cooperative from its point of view - it does not have to know if these changes are a direct result of its own past actions.

2.3. The Engine for Self-organization

The designer provides the agents with local criterion to discern between cooperative and NCSs. The cooperative attitude between agents constitutes the engine of self-organization. The agents have to try to choose the more cooperative action when they can and also when NCSs occur to detect them and to remove them. Depending on the real-time interactions the multi-agent system has with its environment, the organization between its agents emerges and constitutes an answer to the difficulties of ambient intelligence problems (indeed, there is no global control of the system). In itself, the emergent organization is an observable organization that has not been given first by the designer of the system. Each agent computes a partial function fpi , but the combination of all the partial functions produces the global emergent function fs. Depending on the interactions between themselves and with the environment, the agents change their interactions i.e. their links. This is what we call self-organization. By principle, the emerging purpose of a system is not recognizable by the system itself, its only criterion must be of strictly local nature (relative to the activity of the parts which make it up). By respecting this, the AMAS theory aims at being a theory of emergence. So, our proposition to design ambient systems is to encapsulate the device in an agent with cooperative attitude that is transforming a device in a cooperative agent.

3. EXPERIMENTS OF SELF-ORGANIZING DEVICES

The first class of AmI systems we worked on in this paper are systems composed of a large number of distributed, heterogeneous and dynamic devices modelling specific services behaviours and implementing temporal resources, processes and tasks to be solved [Cabanis, 2006]. Basically, theses devices need to exchange relevant informations with each other. A cooperative agent encapsulates a device and the objective of each agent is to permanently maintain cooperative relations with agents which are relevant for it: this agent set constitutes its functional neighbourhood. Conversely, it tries to remove from this neighbourhood, agents having uninteresting skills. This is done by the detection and the treatment of NCSs. For each agent, the updating of the representations of its neighbourhood leads to the organisation changes.

3.1. NCSs Detection and Treatment

Listed NCSs were deduced from the meta-rules 1 and 3 presented in the section 2.2:

- NCS 1 ($\neg c_{per}$): An agent cannot associate a meaning to the received message. This NCS can be declined into three more specific NCSs :

  (1) Total incomprehension: the agent cannot associate a meaning to the received message. In this case, because the agent is cooperative, it does not ignore the message but it sends it, according to its representations, towards an agent it considers relevant for the resolution without changing the name of the sender (this action is called “restricted relaxation”).

  (2) Partial incomprehension: only one part of the received message has a meaning for the agent. In this case, the agent sends a partial answer corresponding to the understood part to the sender and it sends the remainder to an agent which it believes qualified (restricted relaxation).

  (3)Ambiguity: the received message has several meanings for the agent. In this case, the agent returns the message to the sender for clarification.

- NCS 3 ($\neg c_{act}$): Two agents want to reach a third one proposing a limited resource and their request exceeds the offer (they are faced with a conflict situation such as for storage capacity or computing performance). In this case, because the third agent is cooperative, it guides one of the former agents towards another having similar resource. All the agents being encapsulated by the same cooperative behaviour, an agent can thus recommend agents having similar competences when it is overloaded. It knows (because they are all cooperative) that it can later benefit from such advantages if a concurrent agent is overloaded. The treatment of these NCSs involves changes of organisation which leads to the creation, the update (reinforcement, reduction) or the removal of the representations (interaction links) possessed by each involved agent.

3.2. Simulations

We then made simulations[6] on an AmI network composed of a set of devices. Each device is dedicated to one of the four following tasks: Grid Calculus (distributed mathematical calculus), Grid Storage (distributed storage of data), P2P (specialised in data exchange) or Web Service (providing services to be composed with others). We have identified some important characteristics of the different tasks needed to implement the agents and used to specify the NCS management in this problem. They are split into the seven following classes:

1. A device can use specific standard. For example, communication protocols would not be the same between two geographically close devices. A restricted relaxation must be activated to ensure connexion.

[6] Developed in JavAct. See www.irit.fr/recherches/ISPR/IAM/JavAct.html

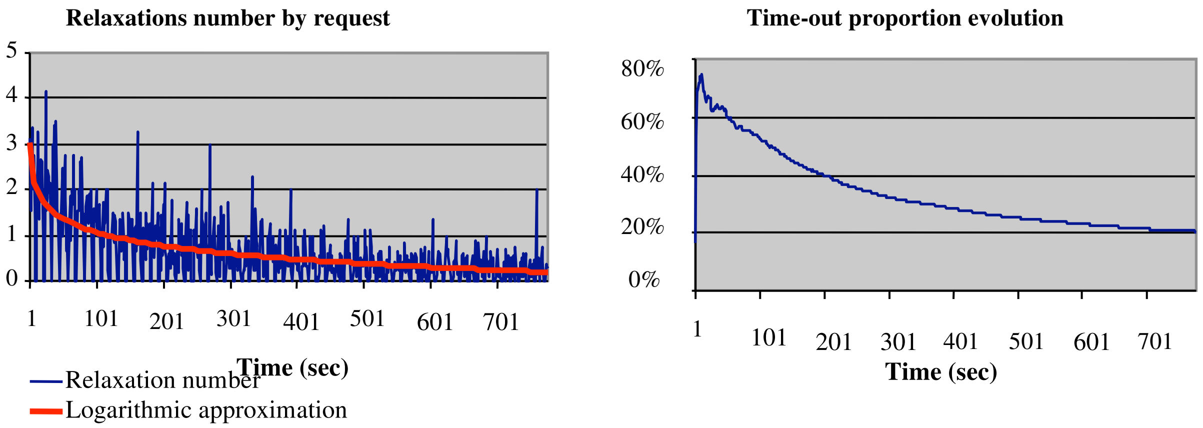
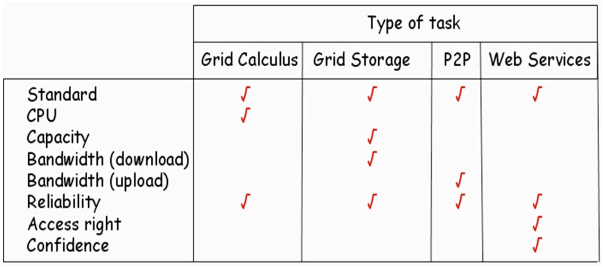


Figure **2**: Cooperation contribution in an AmI system

2. At a time, a device can have calculation capabilities useful for another device. This is this type of request sent during the simulation presented below.
3. A device can have high capacity storage useful for neighbours. Two neighbours can be in conflict when they want to use simultaneously the remaining available memory.
4. A device needs bandwidth (upload or download) when a transfer is time consuming.
5. In some critical AmI applications, reliability could be an important criterion to consider. In this case, a device chooses neighbours having this property.
6. For confidentiality reason, a device could not be allowed to associate a meaning to a received message (encryption). Nevertheless it could relax it to another device able and authorized to decode.
7. An AmI system is typically open, and we could not have an equal confidence (or trust) in all neighbours. This information is an important selection criterion when choosing a new partner or for restricted relaxation.

The adaptive behaviour previously presented through the different detection and treatment of NCSs has been instantiated to this context by taking into account the above characteristics for the four given types of tasks. As indicated in the table 1, only some characteristics are relevant for a given task.



	Type of task			
	Grid Calculus	Grid Storage	P2P	Web Services
Standard	√	√	√	√
CPU	√			
Capacity		√		
Bandwidth (download)		√		
Bandwidth (upload)			√	
Reliability	√	√	√	√
Access right				√
Confidence				√

Table **1**: Criteria taken into account in the simulation

The simulations realized consist of 100 agents, 80% of which are devoted to Grid Computing (GC) calculus. Initially, the AmI system is represented by a graph of agents randomly connected so their functional neighbourhood is composed at the beginning with relevant and not relevant agent. This graph evolves according to interactions between agents.

Randomly, a skill or a specific task to achieve (for example a light control device or a coffee machine) is given to each agent (which represents a device). Then, messages are sent to agents by the simulator to simulate the end-user's requests. The agents also communicate to each other by message sending. The messages can contain an explicit user request, a perception coming from sensors or even information exchanges between agents. In the simulation results (figure 2), 90 requests of GC calculus are submitted each second to different agents of the system. Each task can be relaxed a limited number of times (4 times in this simulation), i.e the number of times a task can be sent from an agent to another. Beyond this number, the task is removed and the sender agent considers its task as being without response after a given time limit (time-out). It then adjusts consequently its representations on the agent to which it has sent the request. Representations of the seven characteristics of the neighbours of an agents are expressed by using measurements of need (standards and access rights in Web Services...), measurements of probability (for the reliability of the services) and measurements of weighted averages for apparent performances (CPU, ...).

The results (figure 2) show a progressively decreasing number of relaxations (synonymous to NCSs) and a decreasing number of time-outs (unresolved requests/tasks) during the system functioning. These results mean that gradually each agent finds its right place in the organisation in spite of unforeseeable events that can occur during the system functioning. The right place means that the agent interacts with the right agents to achieve its goal. In the second curve an asymptotic limit to 20% of time-out can be seen. It is reached when all agents devoted to GC are busy; so the system tends towards its optimality. These preliminary results show that the AmI network, as a collective, adapts itself to the characteristics of each device, only by local perception of criteria and treatments which are independent of any global cost function knowledge.

4. DISCUSSION

The ISTAG (Information Society Technologies Advisory Group) has been engaged since 2000 [ISTAG, 2001], [ISTAG, 2003] in a scenario-planning exercise for European Community. The goal is to give ideas (see also the Philips Homelab [HomeLab] or visions and achievements in AmI from Lindwer and al. [Lindwer, 2003]) about what the daily life might be in an AmI environment in the year 2010. There will probably exist numerous approaches to these scenarios. Our approach considers them as emergent phenomena and we try to solve them by providing an approach enabling to build self-organising systems.

4.1. Ad-hoc versus Emergent Scenarios

The usual way to fit with these scenarios is to define ad-hoc middleware, infrastructure or protocols supporting them. From our point of view, this is not a complete relevant approach because in functioning, new scenarios always occur and don't match with these ad-hoc solutions. As quoted by Emiliano and Stephanidis [Emiliani, 2005] "*In such a context, the concepts of universal access and design for all acquire critical importance in facilitating the incorporation of accessibility in the new technological environment through generic solutions*".

Because the global behaviour of an AmI system evolves constantly, it cannot be a priori defined: it is a really emergent phenomenon resulting from the dynamic coupling between evolving human needs and mobile networked devices having limited capabilities. So, the design of ambient systems needs new models new tools which deals with their complexity, distribution, heterogeneity and openness. Systems able to provide emergent phenomenon represents one way but not the unique way to build them. In this paper, we focus on self-organising multi-agent systems to conceive ambient applications, where the self-organising mechanisms lead to the emergence of the goal of the applications.

4.2. Properties of Self-Organizing Systems

Self-organizing systems show a lot of properties needed in AmI systems. These properties are the following:

- Robustness. An AmI system is a very stressful environment for all the devices. Nevertheless, the previous experiments show that they are able to function correctly after a short delay of adaptation: this is basically a robustness property.
- Self-repair. When a part of a self-organizing system fails, its neighbours have to find new acquaintances related with their needs. This is exactly the goal of the cooperative self-organizing process shown in the previous parts. Thus, the global AmI system falls into a graceful degradation, according to the missing skills of the failing device.
- Scalability. An SoS has inherently the ability to grow incrementally without re-engineering the process because the adaptation process is self-contained in each autoomous device.
- Openness. Removing a device in a self-organizing system is a self-repair process for its old neighbours. Adding a new device is also a self-repair process from the new device point of view.
- Complexity reduction. The design of a self-organizing AmI system is bounded by the specification complexity of the cooperative behaviours of its isolated devices. Moreover, as we have done in the experiments, these specifications are for each generic class of devices only and not for the individuals. Consequently, the complexity of a self-organizing AmI is equal to the more complex device to design, and not to the scale of its global organization.

5. CONCLUSION AND PERSPECTIVES

Modern and future artificial systems, for which AmI systems are typical cases, show an extremely high dynamism, resulting in very complex and unpredictable interactions among their distributed components, making it impossible to normally reason about the global behaviour. This is the reason why the authors of the ambient intelligence roadmap consider that "*the traditional techniques for building distributed applications, are no more usable in such complex systems: they are only thought to operate in centralised and client server environments*" [Friedewald, 2003].

This is the main reason we search for a local adaptive approach based on Adaptive Multi-Agent Systems. It enables to build systems in which agents only pursue a local goal while trying to keep cooperative relations with other agents embedded in the system. This approach has been partially instantiated in a simulation of heterogeneous devices network. These first encouraging results convinced us of the need to apply it on a large scale real world AmI application in order to demonstrate thoroughly in the near future the properties enunciated in section 4.2.

In their roadmap for AmI, Friedewald and Da Costa claim that there is the need to use new paradigms in the design of such systems. They suggest for example that self-organising software will be available by 2006-2010. "*A key characteristic of a self-organizing system is that structure and function of the system "emerge" from interactions between the elements. The purpose should not be explicitly designed, programmed, or controlled. The components should interact freely with each other and with the environment, mutually adapting to reach an intrinsically "preferable" or "fit" configuration (attractor), thus defining an emergent purpose for the system*" [Gershenson, 2004].

We are not really sure that 2010 is the deadline for true self-organizing applications, nevertheless we agree that SoS are central for tackling the main problems of Ambient Intelligence systems. Our approach aims at providing a generic framework by using cooperative self-organisation rules to build these SoS.

6. REFERENCES

[Axelrod, 1984] R. Axelrod. The Evolution of Cooperation. Basic Books, New York, 1984

[Bernon, 2005] Bernon C., Camps V., Gleizes M-P, Picard G. – "Engineering Adaptive Multi-Agent Systems: The ADELFE Methodology" in "Agent-Oriented Methodologies", B. Henderson-Sellers, P. Giorgini (Eds.), Idea Group Pub, NY, USA, pp. 172-202, juin 2005.

[Cabanis, 2006] Cabanis V., "Etude de la dynamique auto-organisationnelle du Web fondée sur l'activité coopérative de ses composants", Master of research report of Paul Sabatier University, June 2006.

[Di Marzo, 2005] Di Marzo Serugendo G., Gleizes M-P., Karageorgos A.. Self-Organization in Multi-Agent Systems. Dans : The Knowledge Engineering Review, Cambridge University Press, Simon Parsons (Eds), Cambridge, UK, V. 20 N. 2, p. 165-189, juin 2005.

[Dorigo, 1999] M. Dorigo and G. Di Caro. The Ant Colony Optimization Meta-Heuristic. McGraw-Hill, 1999.

[Emiliani, 2005] P. L. Emiliani and C. Stephanidis, Universal access to ambient intelligence environments: Opportunities and challenges for people with disabilities, IBM Systms Journal, Volume 44 Number 3, 2005

[Friedewald, 2003] M. Friedewald, O. Da Costa, Science and Technology Roadmapping: Ambient Intelligence in Everyday Life, (AmI@Life) - JRC/IPTS - ESTO Study - Compiled and Edited by: Michael Friedewald Olivier Da Costa, 2003

[Georgé, 2005] Georgé J-P., Gleizes M-P., Experiments in Emergent Programming Using Self-organizing Multi-Agent Systems, In Multi-Agent Systems and Applications IV, Proc. of the 4th International Central and Eastern European Conference on Multi-Agent Systems (CEEMAS'05), Budapest, Hungary, 15-17 September 2005, Springer Verlag, LNAI 3690, pp. 450-459.

[Georgé, 2003a] Georgé J-P, Gleizes M-P, Glize P., Régis C., "Real-time Simulation for Flood Forecast: an Adaptive Multi-Agent System STAFF", dans AISB'03 symposium on Adaptive Agents and Multi-Agent Systems, University of Wales, Aberystwyth, Society for the Study of Artificial Intelligence and the Simulation of Behaviour, 2003.

[Georgé, 2003b] J-P. Georgé, The AMAS Theory for Complex Problem Solving based on Self-Organizing Cooperative Agents, First European Workshop on Multi-Agent Systems (EUMAS'03), Oxford, UK, 2003

[Gershenson, 2004] C. Gershenson and F. Heylighen, Protocol Requirements for Self-organizing Artifacts: Towards an Ambient Intelligence, In Proc. Int. Conf. on Complex Systems (New England Institute of Complex Systems), 2004

[Gleizes, 2002] Gleizes M-P, Glize P., "ABROSE: Multi Agent Systems for Adaptive Brokerage", in Fourth International Bi-Conference Workshop on Agent-Oriented Information Systems (AOIS-2002), Toronto, Ontario, Canada, mai 2002.

[Heylighen, 1992] F. Heylighen. Evolution, selfishness and cooperation; selfish memes and the evolution of cooperation. Journal of Ideas, 2(4):70–84, 1992.

[HomeLab] 365 days' Ambient Intelligence research in HomeLab, Philips, www.research.philips.com/technologies/misc/homelab/downloa ds /homelab_365.pdf.

[Huberman, 1991] B. Huberman. The performance of cooperative processes. MIT Press / North-Holland, 1991.

[ISTAG, 2001] Scenarios for Ambient Intelligence in 2010, ISTAG report (Information Society Technologies Advisory Group) of the European union, ftp://ftp.cordis.lu/pub/ist/docs /istagscenarios2010.pdf, 2001.

[ISTAG, 2003] Ambient Intelligence: from vision to reality, report of the Information Society Technologies Advisory Group European Union, ftp://ftp.cordis.lu/pub/ist/docs/istag-ist2003_draft_consolidated_report.pdf, 2003.

[Lindwer, 2003] M. Lindwer, D. Marculescu, T. Basten, R. Zimmermann, R. Marculescu, S. Jung, E. Cantatore, Ambient Intelligence Visions and Achievements: Linking Abstract Ideas to Real-World Concepts, Proc. Design Automation & TEst in Europe (DATE), 2003.

[Luck, 2005] Luck M., McBurney P., Shehory O., Willmott S. and the AgentLink Community "Agent Technology : Computing as Interaction – A Roadmap for Agent Based Computing", Compiled, ISBN 085432 845 9, http://www.agentlink.org/roadmap, 2005.

[Müller, 2004] Müller JP., " Emergence of Collective Behaviour and Problem Solving ", in Engineering Societies in the Agents World IV, Fourth International Workshop ESAW-2003, Revised Selected and Invited Papers, pages 311-327, LNAI 3071 Springer Verlag 2004

[Picard, 2005] Picard G., Bernon C., Gleizes M-P., ETTO : Emergent Timetabling Organization, In Multi-Agent Systems and Applications IV, Proc. of the 4th International Central and Eastern European Conference on Multi-Agent Systems (CEEMAS'05), Budapest, Hungary, 15-17 September 2005, Springer Verlag, LNAI 3690, pp. 440-449.

[Picard, 2004] Picard G., "Agent Model Instantiation to Collective Robotics in ADELFE", in Fifth International Workshop on Engineering Societies in the Agents World (ESAW'04), Toulouse, France, Springer Verlag, LNCA 3451, p. 209-221, october 2004.

[Welcomme, 2006] Welcomme J-B., Gleizes M-P., Redon R., Druot T., in Self-Regulating Multi-Agent System for Multi-Disciplinary Optimisation Process, Proceedings of the 4th European Workshop on Multi-Agent Systems (EUMAS'06), Lisbon, Portugal, December 14-15, 2006. B. Dunin-Keplicz, A. Omicini & J. Padget Eds. ISSN 1613-0073.

[Wolpert, 1997] Wolpert D.H, Macready W.G., "No Free Lunch Theorems for Optimization", IEEE Transactions on Evolutionary Computation, Vol.1, N°.1, 1997.

[Woolridge, 2002] [96] M. Wooldridge. An introduction to multi-agent systems. John Wiley & Sons, 2002.

Open Responsive Environments using Software Agents

Frank Guerin and **Wamberto Vasconcelos**[1]

Abstract. Flexible, robust and scalable solutions for responsive environments must be open: physical components in the environment, *i.e.* devices, objects and people, come and go during the environment lifetime. Openness in responsive environments can be naturally achieved by associating each component with a software agent; the software agent is responsible for managing the components' resources and representing its interests while cooperating with other components/agents. The software agents can be endowed with arbitrary functionalities, including, for instance, reasoning and negotiation abilities. However, to achieve real openness, we advocate that software agents must possess a high degree of *transparency*, whereby they can inspect one another and assess their suitability for delegation and cooperation. This means that agents must be able to inspect each others' protocols; and furthermore new agents joining the system must be able to add new protocols describing their operation. We introduce a declarative approach to describing software agents' functionalities enabling their inspection via simulation: software agents can then "put themselves in each other's shoes".

1 Introduction

Responsive environments are physical surroundings with objects and devices that are able to change their state or behaviour to accommodate the presence of people as well as other devices/objects [6]. Technologies and techniques stemming from ubiquitous computing [20, 21], such as RFID tags [15] and Bluetooth[2] [4, 7], provide the building blocks for responsive environments. However, designers and engineers face a difficult challenge when putting together a solution for a responsive environment, as these are inherently dynamic and open: components, that is, people, objects and devices, come and go during the lifetime of an environment. Moreover, any solution must account for changes in individual components' properties and behaviours (*e.g.*, devices that malfunction, people who change preferences or needs) as well as global constraints (*e.g.*, changes in best practices or in health and safety recommendations).

One can create high-quality bespoke solutions to responsive environments, whereby designers, engineers and programmers get together and decide on the various devices, objects and participants to inhabit the physical space in order to achieve certain functionalities. Following this decision, purpose-built software (centralised or distributed) is developed to harness the capabilities of the components and allow their coordinated interaction. This is the case, for instance, of the successful solution developed at Philips as part of the PHENOM project[3] aimed at helping people retrieve, share, and re-live recollections of past events.

A major disadvantage of such an approach is its closedness: even though exceptions and malfunctionings may have been carefully considered, these still are limited to the ones known at the time. Changes can happen that have not been anticipated – for instance, the government may enforce new health and safety regulations that impact a responsive environment solution. In the light of unexpected changes of this kind, a custom-built solution has to be carefully redesigned.

In this paper we propose means to achieve true openness in agent-based responsive environments. Our proposal hinges on software agents that have a *public* declarative description of how they work. New agents, associated with new physical components, publish their description; other agents can peruse these descriptions and make informed decisions as to whether the new agents are suitable to have tasks delegated to. The agents' functionality descriptions must also reflect the components' capabilities: an agent will publish, for instance, that it can commit itself to providing pictures every 10 seconds (if the agent is associated with a camera) or that it can provide information on the temperature of a room (if the agent is associated with a temperature sensor).

This paper is organised as follows. In the next section we explain aspects of agent-based responsive environments. In Section 3 we describe our proposal for achieving openness in agent-based solutions for responsive environments. In Section 4 we illustrate our proposal with an agent-based solution for patient care in the home. We survey related work in Section 5; we draw conclusions and give directions for future work in Section 6.

2 Agent-Based Responsive Environments

Software agents [23] have been used in responsive environments solutions (*e.g.*, [2, 10, 12] and [18]). The association of distributed threads of execution with physical components allows for arbitrary functionalities to be used in the management of resources and coordination of activities. These functionalities are combined with the desirable features of software agents such as proactiveness and social abilities (communication) [23]. For instance, a digital camera able to take pictures can be associated with a software agent that will manage any requests from other components for pictures, but the agent will also store the last n pictures taken. Even though the camera itself may not have provisions for storing more than one picture, by associating an independent thread of execution with it, we are able to extend its functionalities.

The same physical components can be associated with different software agents at different times, thus allowing for hassle-free *versioning*. In such case, engineers and programmers devise new versions of software agents to replace previous ones, fixing any bugs, improving on existing features or adding new functionalities to take advantage of new components. The new software agents can take over from their previous counterparts without the need to redesign

[1] Dept. of Computing Science, Univ. of Aberdeen, Aberdeen AB24 3UE, UK, email: {fguerin, wvasconc}@csd.abdn.ac.uk

[2] http://www.bluetooth.com

[3] http://www.research.philips.com/technologies/syst_softw/phenom/overview.html

the whole solution from scratch.

We show in Figure 1 a simplified responsive environment scenario. The diagram shows a physical environment with components $c_1, \ldots, c_8$ and associated agents $a_1, \ldots, a_8$ executing in a "digital cloud". The components are objects, devices and people: some of these have communication and/or sensing capabilities, for instance, an RFID antenna that reads RFID tags, a Bluetooth-enabled PDA carried by a person, a plasma screen or a radio-controlled light switch.

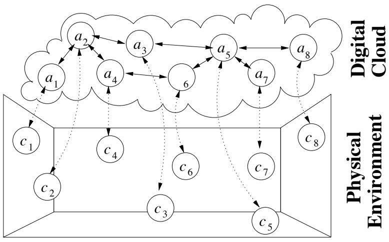


Figure 1. Physical Environment and Software Agents

A simple (and scalable) bootstrapping mechanism for agent-based solutions to responsive environments relies on an initial sensor to detect all other components. This sensor will have its associated software agent started manually: this agent will continuously process the events flagged by its sensor (*e.g.*, signal detected within range or signal left range). When the sensor detects the presence of another component within its range then its agent starts up another agent to be associated with the newly detected component.

The software agents inhabit a "digital cloud". This is a metaphor for the underlying infrastructure working like the operating system of a stand-alone computer, integrating via software a set of hardware components. The digital cloud can be implemented in various ways. In [2, 6], for instance, a blackboard system provides support for ad-hoc message-passing among the agents; whereas the agents themselves are lightweight threads running on a single Java Virtual Machine (JVM) [16].

Components have disparate capabilities. For instance, RFID-tagged decorative items, such as vases books, can only be sensed: they do not have any behaviours to be controlled by their agents. Other components, such as plasma screens and loudspeakers, have means to interact with the environment, but the environment cannot interact back with them. Bluetooth-enabled PDAs and mobile phones can interact with the environment and also allow for the environment to directly interact with them. The software agent must be aware of the capabilities of the component it is associated with: the component's unique identification should allow agents to find out about its capabilities. Ontologies and taxonomies for describing devices have been proposed (*e.g.*, [14, 8] and [19]) with a view to enabling automatic interoperability and discovery.

3 Open Responsive Environments

Our approach to achieving openness in responsive environments involves making all the specifications governing agents' interactions open to inspection and modification. "Open to inspection" means that agents' interaction protocols can be inspected by other agents who may wish to: (i) determine how to interact with another agent in order to achieve some desired goal (i.e. an agent can find the appropriate path through another agent's protocol, which will give the desired outcome); (ii) foresee the consequences of taking on certain roles in an interaction (because the protocols' specifications are open to inspection, agents can mentally step forward through the protocol and see what consequences could arise from taking on certain roles). In order to make it feasible for agents to analyse open specifications in this way, the specifications will need to be in a declarative language. "Open to modification" means that agents can propose new protocols, or modify existing ones, as required by the scenarios that arise (when new devices are added or removed, or new humans with different requirements appear etc.). In order to allow seamless reconfigurations of roles, duties etc., we want to be able to add interaction rules or protocols at run time without needing to restart the system. This means that there must always be a protocol available which allows an agent to propose new rules.

As we are dealing with roles, duties and interaction rules, we are naturally lead to the metaphors and ideas which have been used in work in electronic institutions [9]. We use the idea of an *institution* to provide a framework within which we describe the rules governing agent interactions, as well as the current status of the agents within our responsive environments; i.e. the specification of the institution includes all the interaction protocols and normative relations that hold for agents. The institution keeps track of public information about the state of affairs in the environment as well as the rules which determine how environmental events (including agents' actions) can change this state. This information is described by the *institutional facts* F, which consists of the *rule* type of facts R and the *state of affairs* type of facts A, so $F = \langle R, A \rangle$. The state of affairs facts A can include such things as the roles occupied by various agents, and the permissions, powers and obligations associated with these roles (i.e. the norms). For example, an agent may have the role of safety officer, and may have permission to contact the emergency services, while being obliged to regularly check the smoke and fire sensors. Through these types of normative facts the institution can regulate the activities of the agents in the environment. The rules facts R describe how institutional facts can be created or modified. Rules can have preconditions which depend on the physical world and/or on other institutional facts. For example, the detection of smoke and the absence of cooking activity could be necessary preconditions to being permitted to contact the emergency services; occupying the role of safety officer (a purely institutional fact) could be another precondition. Concrete examples of rules and facts appear in the next subsection.

The environmental sensors are responsible for transforming any relevant change in the world's state into an event (typically by sending a message), therefore we can say (without loss of generality) that the institutional facts change only in response to events in the world (we do not allow rules to refer to states of the world directly - instead the reference to the world's state happens via events). Typical events include messages being transmitted, timer events and possibly other non communicative actions of agents, or events such as agent death. Let E be the set of possible events and let $\mathcal{F}$ be the set of possible institutional facts. Let $update$ be a function which updates the institutional facts in response to an event; $update : E \times 2^{\mathcal{F}} \rightarrow 2^{\mathcal{F}}$. Now in an institution $\mathcal{I}$, it is the institutional rules R which indirectly define this update function. The institution interprets the rules in order to define the update function. Let the interpreter function be I, where I maps R to some update function. An institution $\mathcal{I}$ can then be fully specified by specifying the interpreter I and the facts $F \subset \mathcal{F}$. Recall that F is itself composed of the *rule* type of facts R and the *state of affairs* type of facts A, so $F = \langle R, A \rangle$. Therefore institution $\mathcal{I}$ can be represented by a tuple $\langle I, F \rangle$. Note that to describe an institution by $\langle I, F \rangle$ is to describe it in its current state (thus we will not speak of the *state* of institution $\langle I, F \rangle$). Given an institution described by $\langle I, F_0 \rangle$ at some instant, and a subsequent sequence of events

$e_1, e_2, e_3 \ldots$, we can calculate the description of the institutional facts after each event by repeately applying the $update$ function, obtaining a sequence of facts descriptions: $F_0, F_1, F_2, \ldots$, where each F_i is related to F_{i-1} as follows: $F_i = update_{i-1}(e_i, F_{i-1})$ where $update_{i-1} = I(R_{i-1})$ (and $F_i = \langle R_i, A_i \rangle$ for all i). The interpreter I remains fixed throughout all runs. Note that the institutional facts being modified by a rule could be rules themselves. This is how our framework will allow protocols to be added or modified as the system runs.

It is worth noting that the update rule we have been describing needs to have access to all events in the system in order to build a complete picture of the social facts. We envisage a central computer in the responsive environment which maintains the record of the state of the institution. All events must be sent to this centre, and agents can query the centre for the definitive copy of the institutional facts. This centralised solution may not be desirable in some scenarios, but there are ways it could be more distributed, for example by having more than one centre, each maintaining duplicate records in case of failure. It would also be possible to allow subgroups of devices to have their own set of institutional facts during certain interactions, and to coordinate with the centre for more important business; this solution is not explored here however.

3.1 A simple Prolog Interpreter

The rule interpreter I mentioned above is the immutable part of an institution. A poor choice of I could place limits on what is possible with that institution; we now specify an I which does not restrict the types of rules which can be added to the institution as the system runs. We make use of Prolog as the logic programming paradigm is particularly appropriate for agent communication, there is also evidence that Prolog already enjoys considerable popularity in the agent communication semantics community [1, 13, 11].

```
interpretEvent(F,Event,NewF):-
  F=[Rules,Asserts],
  Event=..EventAsList,
  append(EventAsList,[F,NewF],NewEventAsList),
  Pred=..NewEventAsList,
  copy_term(Rules,Rules2),
  member([_|[Pred|Tail]],Rules2),
  callPred(Tail,Rules).

callPred([],_).

callPred([HeadPred|Tail],Rules):-
  copy_term(Rules,Rules2),
  member([_|[HeadPred|NestTail]],Rules2),
  callPred(NestTail,Rules),
  callPred(Tail,Rules).

callPred([HeadPred|Tail],Rules):-
  call(HeadPred),
  callPred(Tail,Rules).
```

This interpreter takes facts `F` and event `Event` as input and returns the updated facts `NewF`. Its operation is quite simple: it searches through the rules part of `F` until it finds a rule matching the head of `Event` and then it invokes this rule using `callPred`. Note that the *interpretEvent* predicate invokes *member* to find the appropriate predicate to match the event (i.e. find it in R). This is important so that agents are unable to directly invoke Prolog predicates with their messages; their messages are interpreted first. Without this precaution our institution could never limit agents from getting access to the underlying Prolog and hence having the power to make arbitrary modifications to the institution, as it would always accept Prolog predicates, which could be used to reprogram it. Rules stored in R are written in the form of lists, with an index number at the head of each rule. A Prolog clause of the form `pred1(A,B):-pred2(A),pred3(B).` becomes `[1,pred1(A,B),pred2(A),pred3(B)]`. This corresponds to the Horn clause $pred2(A) \wedge pred3(B) \rightarrow pred1(A, B)$.

```
[
 [ 1,
   addRule(Rule,[R1,A1],[NewR1,A1]),
   append(R1,[Rule],NewR1)               ],
 [ 2,
   deleteRule(Index,[R2,A2],[NewR2,A2]),
   delete(R2,[Index|_],NewR2)            ]
]
```

Let the above program be called *prog*. Let the interpreter machine $I = \langle prog, Prolog \rangle$. Let the assertions A be initially empty and the rules R containing only the two rules above.

This is the core of our institution. Let us briefly illustrate how a new rule could be added to it which would change the interpretation of subsequent events. The following is an example of an event which would add our our new rule:

```
addrule([3,assert(Fact,[R,A],[R,[Fact|A]])])
```

After interpreting this event, the rules R will be updated so that subsequent `assert` events cause the addition of an element to the assertions A. For example, a subsequent event

```
assert(alive(agent1))
```

would add "`alive(agent1)`" to A. Note that this is invoking our rule 3 and not Prolog's built-in *assert* predicate.

We now add some basic "housekeeping" rules. We will have a *timer* predicate in A, which records the current time, e.g. *timer(524)*. We will assume that our agent platform generates timer events at regular intervals. Whenever a timer event happens we want to update the clock and execute a set of housekeeping rules. These rules perform housekeeping checks, for example to see if an agent has failed to meet a deadline. The following rule (in R) handles the timer event:

```
[  3,
   timer(Time,[R,A],[NewR,NewA]),
   replace(A,timer(Time),UpdatedA),
   protocolHandler(UpdatedA,
                timer(Time),postProtocolA),
   housekeeping([R,postProtocolA],[NewR,NewA])
]
```

Here we have assumed the existence of a *replace* predicate which replaces a named predicate in A with a new version. The initial *housekeeping* predicate simply preserves the institutional facts F; subsequent components will modify the predicate, adding their own rules. The protocol handler will see if any protocol is currently active, and if so it will execute the specific rules for that protocol; it is often the case that timer events may trigger off state transitions in a protocol. Protocols will be described below.

It is desirable to add another layer for the interpretation of agent communications. We create a *speechAct* rule for this purpose. Agents communicate by sending messages (events) of the form *speechAct(sender, receiver, performative, content)*. We must rely on the platform's message handling layer to check that messages do indeed have the correct *sender* parameter (and possibly to discard any messages where the parameter is falsified); there is no way to do this once the event is processed by the interpreter. We also rely on the platform to distribute the message to the intended recipients. The message event is then handled by our *speechAct* rule. This rule has a section of Prolog code for each possible act: inform, request, propose, etc. The *speechAct* predicate is particularly useful to gather together all those operations which need to be done during the processing of any message (e.g. check roles, permissions and empowerments). This is described below. With this in place we protect the lower level operations from direct access by the agents. We do not want agents to be able to directly invoke the timer event or the rule changing events; however, we can still create speech acts which allow the modification of rules by suitably empowered agents. For example we will include an *addProtocol* speech act which can be sent by a suitably empowered agent to the central computer. The content of this speech act will be a list of rules to be added in order to implement the new protocol. The Prolog code to handle the *addProtocol* speech act simply calls the *addrule* event for each rule of the protocol.

3.2 Implementing Norms

The normative relations we implement are defined by predicates stored in the assertions A. Relations can apply to agents directly or via *roles*; an agent occupies one or more roles (also stored in the assertions A). There are four types of normative predicate: *power, permitted, obliged* and *sanction*. Sanctions are defined for actions which agents should not do. Permitted or obliged actions are treated as exemptions to these sanctions, i.e. the sanction applies unless the agent was permitted or obliged. Power and permission have arity 3: the first parameter is the agent, the second is the performative of the speech act he is empowered/permitted to do, and the third is a condition. For example

```
power(tempAgent,setTemp,[Content=[C],C<27])
```

means that any agent in the role of tempAgent (temperature controller) is empowered to send a *setTemp* speech act provided it complies with the following conditions: the content of the act must be a list containing a single number whose value is less than 27. If a condition is the empty list then it is always true. Sanctions and obligations add a further (fourth) parameter, which is the "sanction code" Following [13] we will associate a 3-figure "sanction code" with each norm violation (in fact in our example we simply always use code 101). We will not go into details of how to handle sanctions here, but in the scenarios considered it might be appropriate to retire sanctioned agents, or to require that some engineer is requested to inspect the devices these agents are managing, as faulty devices could cause the agent to fail to fulfil its duties. Finally the obligation adds a fifth parameter which is the deadline by which the specified speech act must be sent.

The following algorithm (in pseudocode) is added to the *speechAct* rule to handle the normative relations, it effectively defines an operational semantics for the normative relations:

Algorithm HANDLE-NORMS

```
1. input: a speech act with Sender, Receiver, Performative, Content
2. Check if there is an obligation which requires that Sender (or one
   of the roles he occupies) send this speech act. If so remove the
   obligation from $A$ and jump to 5.
3. Check if there is a sanction for Sender (or one of the roles he
   occupies) sending this speech act: If not, go to the next step; If so,
   ◦ check if there is a permission for Sender (or one of the roles he
     occupies) to send this speech act: If so, go to the next step; If
     not, apply the specified sanction.
4. Check if Sender (or one of the roles he occupies) is empowered to
   send this speech act: If not, discard the act and exit this algorithm.
5. Process the act as normal.
```

With this implementation we make obligation imply permission and power.

We also need to add the following to the *housekeeping* rule (recall that the housekeeping rule is invoked on every timer event):

- For each obligation check if it has timed out. If so, apply the sanction to the agent and remove the obligation from A.

3.3 Implementing Protocols

Protocols are encoded via their own rules in R. Each protocol has a unique name and may be represented by a number of clauses in R. Protocols essentially determine what actions are to be taken next, given the current state and an event that happens (events are either messages or timer events in our scenarios). They do this by consulting the current state and modifying the normative relations according to the event that has just happened; typically a protocol will specify some actions as obligatory or permitted at each stage of the protocol. Agents initiate protocols by using the special speech act *initProtocol*; this firstly causes a new fact to be inserted in A which names the protocol currently being enacted. The *speechAct* predicate then passes control to the protocol on initiation. Each protocol has an initiation section which sets up relevant roles and variables as well as the initial normative relations. After initiation, the protocol handler code is invoked on every event, and it can check what protocols are being enacted currently, and call the protocol's rules. The protocol's specification is organised in sections, the first one being the initiator, and subsequent ones being invoked either on timer events or message events (this can be seen in the example below). Those sections which handle message events can access the sender, receiver, performative and content of the speech act. Sending a *exitProtocol* message terminates the protocol.

4 Example Scenario: A Visitor With Special Needs

The environment is an apartment inhabited by humans. A visitor who is an elderly patient with special needs arrives to spend some time at the apartment. We will focus on one of the visitor's special needs: he needs to take medication regularly, and needs to be reminded to do so. The visitor also brings an electronic medical cabinet which is wireless enabled, and which is able to report to the central computer when drugs have been taken from it. The visitor's RFID tag is recognised by the central computer; the tag includes information about the visitor's special needs; the central computer can then connect to the internet to find agent code to handle these needs, and a new agent is spawned to manage the environment in order to satisfy the newcomer's special needs. We assume that the environment is already being controlled by some agents prior to the visitor's arrival. There may be other humans present in the environment and existing agents may be controlling the environment in accordance with their preferences; we do not want to upset this activity unnecessarily. For example, there could be an agent in charge of the entertainment system, an agent in charge of fire safety, and one in charge of air conditioning; each has control of some of the environment's devices and we want these agents to keep doing their jobs as normal. The newly spawned agent (healthcare agent) will need to interact with the entertainment system agent as it has the capability to deliver messages (by audio and video) to humans in the environment. The healthcare agent will need to add new protocols to the institution to describe how it interacts with the other agents in the system, so that the other agents can inspect them and know what obligations will follow from accepting a request to give notifications, for example. The healthcare agent must be empowered to add new protocols, and will send an *addProtocol* speech act to the central computer to do so; we will not describe this in detail now, but will focus on the protocol itself. The protocol below is initialised by a speech act from the healthcare agent to the entertainment system, of the following form:

```
speechAct(agentA, agentB, initProtocol,
            [medNotification,agentC])
```

The actual names *agentA* and *agentB* are irrelevant, because as soon as the protocol is initiated these agents will be assigned roles, and subsequent messages will be sent to these roles. The unique name of this protocol is *medNotification*. The medical cabinet agent is named *agentC* and is included in the content of the initiating message. (We do not include the description of another necessary protocol by which the healthcare agent will instruct the cabinet agent to send notifications directly to the entertainment system.) The initiation of the *medNotification* protocol amounts to the healthcare agent asking the entertainment agent to take on the role of notifier for a single instance of the medicine taking notification (i.e. we assume the healthcare agent takes responsibility for making this request every time medication is needed). Taking on the notifier role entails carrying out a notification firstly at standard audio volume. If after five minutes there is no event indicating that medication has been removed from the cabinet, then the notifier must carry out high volume notifications. If there is still no event after a further five minutes, the notifier agent must contact the emergency services to get outside help. This simple *medNotification* protocol is now presented in pseudocode:

```
on initiation:
 add role(Sender,healthAgent)
 add role(Receiver,notifierAgent)
 Content=[Protocol,AgentName]
 add role(AgentName,cabinetAgent)
 retrieve timer(Time)
 ReplyTime:=Time+00:00:20
 NotifyTime:=Time+00:00:30
 add permit(notifierAgent,accept,
    [Receiver=healthAgent])

%no reply is assumed to mean rejection
on timer variable Time>ReplyTime:
 add obliged(healthAgent,exitProtocol,
    [Receiver=notifierAgent],101,NotifyTime)

on accept(notifierAgent,healthAgent):
 ReplyTime:=0
 add obliged(notifierAgent,inform,
    [Receiver=audiovisual,Content=[medVol,
    "Take Medication"]],101,NotifyTime)

on inform(notifierAgent,audiovisual):
 retrieve timer(Time)
 TimeLimit1:=Time+00:05:00
 TimeLimit2:=Time+00:10:00

%patient has taken medicine so end protocol
on inform(cabinet,notifierAgent):
 add obliged(notifierAgent,exitProtocol,
    [Receiver=healthAgent],101,TimeLimit2)

on timer variable Time>TimeLimit1:
 add obliged(notifierAgent,inform,
    [Receiver=audiovisual,Content=[highVol,
    "Take Medication"]],101,NotifyTime)

on timer variable Time>TimeLimit2:
 EmergTime:=TimeLimit2+00:00:05
 add obliged(notifierAgent,inform,
    [Receiver=emergencyServ,Content=[
    "No Patient Response"]],101,EmergTime)
 add obliged(notifierAgent,exitProtocol,
    [Receiver=healthAgent],101,EmergTime)
```

Of course there are serious security issues which would need to be considered in practical scenarios, and safeguards would need to be in place to ensure that a rogue agent would not arbitrarily change the institution's rules. For example we could require that the central computer check through the protocol to ensure it is safe, or alternatively it might only accept new protocols which have been certified by a trusted third party source. However, in this work we were merely concerned investigating what type of communication infrastructure would allow new protocols to be added (or existing ones modified) at run-time, without disrupting the ongoing interactions in the environment.

5 Related Work

In [12] we have the notion of "ambient mediated" interaction between humans and software agents, which is akin to our approach. Human inhabitants of a responsive space are generally interested in the services that the space can provide to it, and would rather have access to those services automatically. It is expected, therefore, that humans consider irrelevant whether those services are the outcome of coordinated actions of a society of software agents. The interactions among software agents, between agents and the environment are, in that project, carefully constrained to tightly controlled protocols. Our work extends their results by giving room to more sophisticated patterns of interaction.

The Intelligent Inhabited Environments research group at the University of Essex explicitly proposes, as we do, the construction of intelligent responsive environments through the coupling of the physical world and virtual worlds inhabited by software agents. Their test bed – the iDorm experiment, which is a student dormitory facility to serve a single student, equipped with a host of sensors and effectors that can monitor the activities in it and respond accordingly – only allows for single-occupant scenarios and hence theirs is a restricted form of interaction among the software agents. This contrasts with our approach and proposed test cases, in which we necessarily must take into account multiple occupants of a single space, complex interactions among their proxy agents and even among proxy agents and other agents that may exist in the informational space without having necessarily a physical counterpart, thus leading to more complex patterns of interaction among agents and among agents, human and the environment. Some recent results from the Intelligent Inhabited Environments group can be found in [10].

Another research work related to ours is that reported in [5]. Similar to our work, those researchers propose a framework in which software agents can negotiate based on their capabilities and goals and on context dependent information – *e.g.*, information arising from sensors. Those authors, however, also prefer not to consider the possibility of multiple human users sharing the same output – in which case this output should be of use to all concurrent users – or more convoluted negotiation protocols (their approach is basically first-come-first-served). The test bed proposed by those authors, nonetheless, is noteworthy: an active, context aware arts museum (`http://peach.itc.it/`). Our proposal should be applicable to their scenario with great effectiveness, since active museums match perfectly with the abstract setting we have taken into account.

In our work we have focused specifically on intelligent responsive environments which require sophisticated intelligent behaviour and interaction among the software agents that constitute them. A similar approach is explored in [18], but our work differs in the way we account for spatial interaction: in that work physical location rules the interactions among intelligent software agents (*viz.*, the "situatedness" of agents, [22]), whereas in our work physical location is an attribute of "spaceless" information agents, who interact through a tuple space.

In the research on agent institutions, and interaction protocols, there are numerous closely related works which have inspired our approach. Firstly, our desire to achieve openness lead us to the idea of agents having public descriptions of how they work; this idea appears in [17], where detailed operating instructions of an artifact are exposed. We try to take this idea to a higher level of abstraction where we do not consider the internal mental states of an agent, but only those things which the agent commits to doing via public normative relations (e.g. permissions, obligations). Secondly there is the idea of changing the rules governing these normative relations. In [13] the possibility of agents modifying the rules of the institution is mentioned; it is stated that this would require "interpretable code or some form of dynamic compilation". In [3] the event calculus formalism has been implemented to animate a specification of a rule governed agent society, but it is also stated that features of the underlying programming language could be made accessible to complement the event calculus formalism; this comes closer to the flavour of our proposal. We forego the logical formalism and make use of the programming language directly as our specification language. In [9] normative relations are implemented in the Jess production rule system. The authors mention the possibility of "societal change", where societies may "evolve over time by altering, eliminating or incorporating rules". This societal change facility is not actually implemented in [9], but the authors do specify norms in a computationally grounded language based on observable phenomena.

To the best of our knowledge this is the first approach to agent institutions which is completely modifiable by the agents at run-time; i.e. any protocols/interaction rules can be added by the agents. Of course it is also possible within the framework to restrict the ability of some agents to change the rules. Complete freedom to change the institution might not be necessary in many scenarios, but the framework proposed here can easily be refined for the special cases of particular applications.

6 Conclusions, Discussion and Future Work

We have been concerned with providing an infrastructure which would allow agent societies to reconfigure themselves dynamically. Aspects of this have already been handled very well in the literature on coordinating agent teams, for example the well known contract net protocol which allows tasks to be allocated. We have chosen to focus our attention on a more neglected area, and that is openness for specifications; i.e. allowing both their inspection and modification. The need for this is motivated by the dynamic nature of ambient intelligent environments; not only are low level sensor/controller devices added and removed, but high level task specifications might also change. A good example of complex change is when the "best practices" for patient care might change; for example in the scenario above a new "best practice" recommendation might introduce the need to log significant events such as the patient taking longer than expected to take medication by adding an entry to a record on the Internet. In our future work we intend to look at developing tools which will allow protocols to be analysed to determine the consequences of protocol design decisions, for example if there are any executions which do not satisfy some required design goals. This work will bring together existing work on protocols for patient care with work on specifying and verifying electronic institutions.

REFERENCES

[1] R. Agerri and E. Alonso, 'Semantics and pragmatics for agent communication', in *Progress in Artificial Intelligence, Lecture Notes on Artificial Intelligence Series (LNAI) 3808*, ed., Bento et al., 524–536, Springer-Verlag, (2005).

[2] C. Aitken. Designing Software Agents to Manage Responsive Environments, May 2006. B.Sc. Hons. Report, Dept. of Comp. Science, Univ. of Aberdeen, UK, available at `http://www.csd.abdn.ac.uk/~wvasconc/aitken_chris.pdf`.

[3] A. Artikis, M. Sergot, and J. V. Pitt, 'Specifying norm-governed computational societies', in *Imperial College Department of Computing Technical Report DTR06-5*, (2005).

[4] J. Bray and C. Sturman, *Bluetooth: Connect Without Cables*, Prentice Hall PTR, USA, 2000.

[5] P. Busetta, T. Kuflik, M. Merzi, and S. Rossi, 'Service Delivery in Smart Environments by Implicit Organizations', in *Procs. 1st Int'l Conf. on Mobile & Ubiquitous Systems (MobiQuitous'04)*, Boston, U.S.A., (August 2004). IEEE Comp. Society.

[6] F. S. Correa da Silva and W. W. Vasconcelos, 'Managing Responsive Environments with Software Agents', *Applied A. I.*, (to appear 2007).

[7] M. Dideles, 'Bluetooth: a Technical Overview', *Crossroads*, **9**(4), 11–18, (2003).

[8] FIPA. Device Ontology Specification. Foundation for Physical Agents document available at `http://www.fipa.org/specs/fipa00091/XC00091C.pdf`, 2001.

[9] A. Garcia-Camino, J.A. Rodriguez-Aguilar, and P. Noriega, 'Implementing norms in electronic institutions', in *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pp. 667 – 673, (2005).

[10] H. Hagras, V. Callaghan, M. Colley, G. Clarke, A. Pounds-Cornish, and H. Duman, 'Creating an Ambient Intelligence Environment Using Embedded Agents', *IEEE Intell. Systs.*, **19**(6), 12–20, (2004).

[11] Y. Labrou, *Semantics for an agent communication language*, Ph.D. dissertation, University of Maryland Graduate School, 1996.

[12] J. M. V. Misker, C. J. Veenman, and L. J. M. Rothkrantz, 'Groups of Collaborating Users and Agents in Ambient Intelligent Environments', in *Proc. 3rd Int'l Joint Conf. on Autonomous Agents & Multi-Agent Systems (AAMAS'04)*, volume 3, New York, USA, (2004). ACM Press.

[13] J. Pitt, L. Kamara, M. Sergot, and A. Artikis, 'Formalization of a voting protocol for virtual organizations', in *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'05), Utrecht, July 2005*. ACM Press, (2005).

[14] D. J. Russomanno, C. Kothari, and O. Thomas, 'Building a Sensor Ontology: A Practical Approach Leveraging ISO and OGC Models', in *Proc. Int'l Conf. on AI*, Las Vegas, USA, (2005).

[15] J. R. Smith, K. P. Fishkin, B. Jiang, A. Mamishev, M. Philipose, A. D. Rea, S. Roy, and K. Sundara-Rajan, 'RFID-Based Techniques for Human-Activity Detection', *Comm. ACM*, **48**(9), 39–44, (2005).

[16] Spell, B., *Professional Java Programming*, Wrox Press Inc, 2000.

[17] M. Viroli, A. Omicini, and A. Ricci, 'Engineering mas environment with artifacts', in *Proceedings of the Second International Workshop on Environments for Multiagent Systems E4MAS 2005, Utrecht*, (2005).

[18] G. Vizzari, *Dynamic Interaction Spaces and Situated Multi-Agent Systems: from a Multi-Layered Model to a Distributed Architecture*, Ph.D. dissertation, Universita degli Studi di Milano-Bicocca, Italy, 2004.

[19] W3C. Composite Capabilities/Preferences Profile Public Home Page. World Wide Web Consortium document available at `http://www.w3c.org/Mobile/CCPP/`, 2004.

[20] M. Weiser, 'The Computer for the Twenty-First Century', *Sci. Amer.*, 94–104, (September 1991).

[21] M. Weiser, 'Some Computer Science Issues in Ubiquitous Computing', *Comm. ACM*, **36**(7), 75–84, (1993).

[22] D. Weyns and T. Holvoet, 'A Formal Model for Situated Multi-Agent Systems', *Fund. Informaticae*, **63**, 1–34, (2004).

[23] M. Wooldridge, *An Introduction to MultiAgent Systems*, John Wiley & Sons Ltd., England, U.K., 2002.

Conviviality for Ambient Intelligence

Patrice Caire [1]

Abstract. Conviviality is usually considered a positive concept related to sociability, however, further analysis reveals a negative side related to regulations. In this survey paper, we examine the multifaceted concept of conviviality and raise the question: Which definition of conviviality can be used and made operational for ambient intelligence? We propose a two-fold definition of conviviality as a condition for social interactions and an instrument for the internal regulation of social systems. We, then, propose to use conviviality for ambient intelligence as a mechanism to reduce mis-coordinations between individuals, groups and institutions, and as a tool to reinforce social cohesion. Intelligent interfaces, for example, allow instant interactions and thereby create strong needs for coordination and regulation mechanisms that have to be addressed to ensure the safeguard of individuals against abuses, such as privacy intrusions and identity manipulations. It is therefore crucial to take into account social and cognitive factors and to address the ethical issues raised by the large scale development of ambient intelligent systems.

1 INTRODUCTION

Generally speaking, a convivial place or group is one in which individuals are welcome and feel at ease [1] [40] [39], but definitions in literature spread from individual freedom realized in personal interdependence [18], to rational and cooperative behavior [38], to normative instrument [45]. In the context of digital communities and institutions, conviviality refers to qualities such as trust, identity and privacy. One of the four themes of the European Community 5th framework program titled the *Societe de l'Information Conviviale* (User-Friendly Information Society) promoted user empowerment, human interactions, ambient intelligence and distributed services. The Convivio Net Consortium (2003-2005) fostered *convivial technologies* designed to be people centered, support communication and interaction, bridge the digital divide and increase social cohesion and community identity. Figure 1, adapted from [6], illustrates, with key reference dates, the conviviality theme, ambient intelligence vision and development of digital cities [11]; Their goal being to "transform, modernize and improve the level and quality of life of the population at both individual and community levels" [19]. In [6], we identified the need for survey on the use of conviviality.

In this paper, conviviality for ambient intelligence, we raise the question: Which definition of conviviality can be used and made operational for ambient intelligence? This breaks down into the following research sub-questions: What kinds of notion of conviviality exist? How can the positive aspects of conviviality be used for ambient intelligence? How should the negative aspects be taken into account?

In [38], conviviality is defined as "the essential and global characteristic of services ... it emerges from the intelligence of the system and not from a set of local characteristics ... that vary depending upon the application context and the types of users"; Consequently a list of criteria will by itself not suffice. Additional critical factors to consider are: on the one side, the relations that bind the criteria together and on the other side, the way these relations are perceived by individuals.

Ambient technologies, foresaw in 1991 as *ubiquitous computing* by Mark Weiser [46], rely upon transparent, unobtrusive and intuitive interfaces, closer to the way people think and feel than to the way machines operate.The term ambient intelligence, used in 1999 by the European Union's Information Society Technologies Program Advisory Group (ISTAG) [20], describes a vision where "people will be surrounded by intelligent and intuitive interfaces embedded in everyday objects around us and an environment recognizing and responding to the presence of individuals in an invisible way". One of its goals is to give individuals the possibility to express themselves more efficiently, accurately and effortlessly [20], by invisibly capturing and tracking their preferences into profiles [22]. Hence, the need for context aware applications to take into consideration notions such as privacy, identity and conviviality [44] [9].

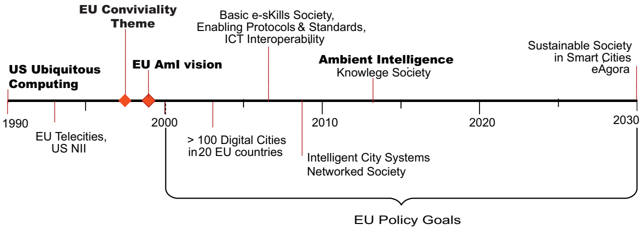


Figure 1. Reference dates for EU digital cities programs, conviviality theme and ambient intelligence

In this paper, we raise ethical issues, such as privacy threats, surveillance of users and identity theft, but do not review them in detail and leave as future work more in-depth analysis. Also out of scope, is how to provide a crisp and usable means of evaluation and measurement of conviviality.

Our methodology is a literature review. The layout is as follows: In each section, we first give an overview on the kinds of notions of conviviality that exist in the field and then suggest how these notions can be used for ambient intelligence. In Section 2 we focus on socio-cognitive approaches, in Section 3 on computer science, agent theory and multi-agent systems, in Section 4 on Human Computer Interaction and in Section 6 we discuss results and summarize our findings.

[1] University of Luxembourg, Luxembourg, email: patrice.caire@uni.lu

2 SOCIO-COGNITIVE APPROACHES

2.1 Definitions of conviviality

Looking at some definitions shows that the meaning of conviviality depends on the context of use (table 1): In sociology, conviviality typically describes a relation between individuals and emphasizes positive values such as equality and community life. However, with power shifting between individuals and groups, conviviality relations change: Minority and majority groups form, outsiders are excluded, others force their way in. This process dynamic and temporal process raises questions such as: How is conviviality created? how does it evolve? What makes it fail?

Table 1. Definitions of conviviality

Etymological and domain specific definitions of conviviality
Origin: 15th century ”convivial”, from latin, convivere ”to live together with, to eat together with”. (French Academy Dictionary)
Adj. Convivial: (of an atmosphere, society, relations or event) friendly and lively, (of a person) cheerfully sociable. (English Oxford Dictionary)
Technology: Quality pertaining to a software or hardware easy and pleasant to use and understand even for a beginner.(Adj.)User friendly, (Noun) Usability. By extension also reliable and efficient. (Grand Dictionnaire Terminologique)
Sociology: Set of positive relations between the people and the groups that form a society, with an emphasis on community life and equality rather than hierarchical functions. (Grand Dictionnaire Terminologique)

2.2 The role of conviviality in social systems

A less common view of conviviality, that pertains to sociology, is when it becomes an instrument to exercise power and enforce one point of view over another [45]. Conviviality is then experienced as a negative force by the loosing side. We summarized, from different sources, positive and negative aspects of conviviality and present, as examples, some excerpts (table 2): The emphasis is on sharing of common grounds and inclusiveness for positive side, on division and coercive behaviors for negative side.

Table 2. Conviviality: Positive and negative aspects

Positive aspects (enabler)
Share knowledge and skills
Deal with conflict
Inclusiveness
Equality
Trust

Negative aspects (threat)
Crush outsiders
Fragmentation
Totalitarism
Reductionism
Deception

2.2.1 Individuals vs. groups

In 1958, Polanyi [34] is the first to use conviviality in a scientific and philosophical context; He describes it as synonymous with empathy ”which alone can establish knowledge of other minds”. By allowing individuals to identify with each other, empathy provides a way to acquire personal knowledge by experiencing the feelings, thoughts and attitudes of an individual. In 1974, Polanyi further describes a community as convivial when it aims at sharing knowledge: members trust each others, share commitments and interests and make mutual efforts to build conviviality and preserve it [35].

In his 1971 critical discourse on education, Deschooling Society [17], Illich defines a convivial learning experience as one based on role swapping, teacher role alternates with learner role, to emphasize the concept of reciprocity as key component to conviviality. In 1973, Illich's Tools for Conviviality [18] brings a new dimension to the concept defined as ”an intrinsic ethical value”. Indeed, for Illich, conviviality means ”individual freedom realized in personal interdependence”, it is the foundation of a new society, one that gives its members the means, referred to as tools, for achieving their personal goals: ”A convivial society would be the result of social arrangements that guarantee for each member the most ample and free access to the tools of the community and limit this freedom only in favor of another member's equal freedom”.

In the 1980's, Putnam and his colleagues further extend the concept of conviviality as an enhancement to social capital. In 1988, they refer to conviviality as a ”condition for civil society” [36], and in 2000, argue that in a civil society ”communities are characterized by political equality, civic engagement, solidarity, trust, tolerance and strong associative life” [37], stressing the strong link between the performance of political institutions and the character of civil life.

Building on Illich learning webs, skill exchange networks and peer-matching communication concepts, Papert and the Constructionists, emphasize in 1991 ”learning-by-making” [32], and in 2001, Sipitakiat develops digital technologies for conviviality, stressing the notion of equilibrium” [40].

In a 2004 semiotics symposium on conviviality,, Schechter takes another look at the concept: ”in a basic sense, conviviality is a social form of human interaction” [39]. The author binds interaction to physical experience and recognizes the social dimension of conviviality, as a way to reinforce group cohesion through the recognition of common values. ”Thus the sharing of a certain kind of food and/or drink can be seen as a way to create and reinforce a societal group through a positive feeling of togetherness (being included in/or part of the group), on which the community's awareness of its identity is based.” Schechter transforms the physical experience of conviviality into a learning and knowledge sharing experience. ”To know is to understand in a certain manner that can be shared by others who form with you a community of understanding”.

It is worth noting that the conviviality values from socio-cognitive context, such as social cohesion, inclusiveness and participation, by putting individuals at the center of change, coincide with the very values praised by the ambient intelligence vision.

2.2.2 The darker side of conviviality

A negative side of conviviality can however emerge, when it becomes an instrument in the hand of power relations: ”Conviviality is achieved for the majority, but only through a process by which nonconviviality is reinforced for the minority” states Ashby [2], who further denounces the instrumentalization of conviviality when one group is favored at the expense of another, ”truth realities about minorities are built from the perspective of the majority via template token instances in which conflict is highlighted and resolution is achieved through minority assimilation to majority norms”.

”Conviviality masks the power relationships and social structures that govern communities” argues Taylor [45] who then, explores the contradiction between institution and conviviality, asking ”whether it is possible for convivial institutions to exist, other than by simply creating another set of power relationships and social orders that, during the moment of involvement, appear to allow free rein to individual expression … *Community members* may experience a sense of conviviality which is deceptive and which disappears as soon as the members return to the alienation of their fragmented lives." These issues raise important ethical questions that must be addressed in the new world of ambient intelligence, for example, with guidelines and best practices, that include all parties point of views, and new coordination theories [27] and mechanisms that manage dependencies among activities.

"Until now, there has been no reasonably comprehensive survey of AmI research projects in Europe, the USA and Canada focused on privacy, security, identity and trust issues" states Wright in his *Safeguards in a World of Ambient Intelligent* project report [49]. No one has considered the range of safeguards needed to protect individuals. The negative sides of conviviality, by revealing these mechanisms, indicate what is to be avoided and point to the mix of different safeguards that have to be put in place to adequately protect individuals, groups and institutions.

2.2.3 From groups to institutions

While Lomosits recommends that conviviality be achieved through consensus and not imposed [26], Hofkirchner identifies the normative idea of unity-through-diversity as deserving attention "when applying conviviality to the level of world society" [14]. The author examines the unity-diversity relation, equates the terms unity-diversity with identity-difference and then describes the four resulting scenarios: (1) "establish identity by eliminating difference at the cost of the differentiated side" yielding reductionism and universalism or (2) "of the undifferentiated side yielding unity without diversity", that is particularism, totalitarism and homogenization; (3) "establish difference by eliminating identity yielding diversity without unity", that is fragmentation and (4) "establish identity in line with difference yielding unity and diversity". The achievement of conviviality is in this integration of difference and differentiation of identity, yielding for example, transculturalism.

"Conviviality (just like conflicts) is based on agreements or contradictions" states Somov [41] who further explains the normative aspect of conviviality with the idea that conviviality belongs to the area of regulation of human interrelations. This regulation aspect of conviviality makes it particularly relevant to future large scale developments of ambient intelligence devices.

2.3 The use of conviviality for ambient intelligence

In ambient intelligence applications, such as the mass-scale annotation system GeoNotes, users "annotate physical locations with virtual notes, which are then pushed to or accessed by other users when in the vicinity" [33]. Groups of users are hence formed by region. What happens afterward between these users would seem to be what is important. With the set up of convivial relations and spaces users are encourage to share knowledge and cooperate with each other, and discouraged to abuse other users. Another ambient intelligence application, Collaborative Capture [8] allows, for example, "a group of people taking pictures at an event to merge their captures and provide a complete collection" . This raises, of course, privacy issues as you may not want to share all your pictures. In the context of spontaneous interactions, traditional security, with authorizations, is difficult to apply and innovative approaches, based on more dynamic notions such as conviviality, have to be investigated. "The very notion of ubiquitous capture can be frightening: the potential capture activity of anyone, anywhere may change social relations between people". In an overall computing environment, focus must be on people and their social situations [42]. Conviviality reinforces common shared ground between the members of a group and can thereby create protection barriers between and for its peers.

3 COMPUTER SCIENCE APPROACHES

3.1 The role of conviviality in Multi-agent systems

In multi-agent systems an agent is defined as "a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives . . . Agents are capable of flexible (reactive, proactive, social) behavior" [48]. This capability is particularly crucial for ambient intelligence since it allows agents to cooperate, coordinate their actions and negotiate with each other.

3.1.1 The use of conviviality for Intelligent Tutoring Systems

The system proposed by Gomes et al. [12] provides a recommendation service of student tutors for computational learning environments. "Each agent pupil represents a pupil logged onto the system. One of the functions of the system is to be the client for an instant message service. Through its agent pupil, any pupil can communicate with other pupils in the system.Another function of the agent pupil is to pass information on the affective states of the pupil. This information can be inferred by the agent or be adjusted by the pupil itself."

The authors' claim that "convivial social relationships are based on mutual acceptance through interaction" hence on reciprocity and in this case students helping each other. A utility function takes as input a student's social profile and computes the student's affective states indicating if the student needs help; if s/he does then the system recommends a tutor. Remaining challenges are with defining utility function inputs to compute recommendations, presently a set of random values, and to automate inferences of students requiring help. This exposes the urgent need for further research in evaluation methods and measures for concepts such as mood, sociability and conviviality.

However, these critical challenges of a technical nature, pointed out so far, are pale in comparison with the ethical issues raised by the possible development of such a system: Preserving pupils'privacy, securing the information gathered to create their social profiles, deterring possible misuse of pupils' affective states and system errors concerning the data. In fact, it is imperative that designers of such systems use guidelines, for instance, the European Privacy Design Guidelines for the Disappearing Computer [21] in order to "implement privacy within the core of ubiquitous computing systems" [22].

3.1.2 The use of conviviality for Conversational Agent

"All service offerings must integrate conviviality to the interaction between user and system as an essential preoccupation" [38]. To fulfill this goal, Sadek et al. define a convivial agent as rational and cooperative. An interaction is convivial "if the agent presents, jointly and at all times, one or all of the following characteristics: Capacity for negotiation, contextual interpretation, flexibility of the entry language, flexibility of interaction, production of co-operative reactions and finally of adequate response forms." These communicative capacities and social intelligence based on emotional intelligence are crucial to enhance agents'ability to interact with users.

Indeed, building on this work, Ochs et al. [31] distinguish felt emotions from expressed emotions noting that ”a person may decide to express an emotion different from the one she actually felt because she has to follow some socio-cultural norms”. We believe this direction to be very relevant to the evaluation of conviviality as it dissociates personal feeling from social expression.

3.1.3 The use of conviviality for reputation systems

Reputation is defined as ”the overall quality or character as seen or judged by people in general and the recognition by other people of some characteristic or ability” [29]. When Casare and Sichman state that ”reputation is an indispensable condition for the social conviviality in human societies” [7], they emphasize that reputation provides transparency quality of the information provided with reputation, throughout the group about its member, this transparency insures the conviviality of the group, as all group members receive the same information about their peers. The authors' system insures that everyone is aware of anyone's behavior, that is anyone's compliance or not to the rules of the group. Casare and Sichman define a functional ontology of reputation for multi-agent systems whereby ”roles are played by entities involved in reputative processes such as reputation evaluation and reputation propagation.”

The authors' claim that ”concepts of the legal world can be used to model the social world, through the extension of the concept of legal rule to social norm and the internalization of social mechanisms in the agent's mind, so far externalized in legal institutions”. In their system, the agents actual behaviors are compared to the social norms observed in their world. The process, however, presupposes an initial reputation profile of users that agents can then update in real time. Reputation acts as a communication tool, ensuring complete social transparency throughout the system. The strict application of norms to reputation however may be difficult and suffer from rigidity. Of course, the same holds for conviviality.

By its very definitions, ”the vision of ambient intelligence has the potential to create an invisible and comprehensive surveillance network, covering an unprecedented share of our public and private life . . . Besides the obvious risk of accidental leaks of information, profiles also threaten universal equality, a concept central to many constitutions, basic laws, and human rights, where *all men are created equal*. Even though an extensively customized ambient-intelligence future where I only get the information that is relevant to my profile holds great promise, the fact that at the same time a large amount of information might be deliberately withheld from me because I am not considered a valued recipient of such information, would constitute a severe violation of privacy for many people” [5].

3.2 The role of norms in multi-agents systems and how it applies to conviviality and ambient intelligence

The role of norms is increasingly getting attention specifically in multi-agents systems (MAS) where the most common view is that ”norms are constraints on behavior via social laws” [3]. In their introduction to normative multi-agent systems, Boella et al. give the following definition: ”A normative multi-agent system is a multi-agent system together with normative systems in which agents on the one hand can decide whether to follow the explicitly represented norms, and on the other the normative systems specify how and in which extent the agents can modify the norms.” [3]. Agents therefore decide how to interact with each other, following conviviality conventions or not, they can, also, modify these conventions and thereby contribute to their evolution. Furthermore, the role of norms for conviviality is an instrument for the internal regulation of social systems [6]: For example, in digital cities ”government regulations extend laws with specific guidance to corporate and public actions” [24].

Several kinds of norms are usually distinguished in normative systems. Within the structure of normative multi-agent systems Boella et al distinguish “between regulative norms that describe obligations, prohibitions and permissions, and constitutive norms that regulate the creation of institutional facts as well as the modification of the normative system itself” [4]. A third kind of norms, procedural norms, can also be distinguished “procedural norms have long been considered a major component of political systems, particularly democratic systems”, states Lawrence, who further define procedural norms as “rules governing the way in which political decisions are made; they are not concerned with the content of any decision except one which alters decision-making procedures” [25].

Boella et al. further describe action models where ”agents are goal directed and try to maximize their choice of means to obtain a goal”. It is assumed that an agent belongs to a group and must follow the norms like all members of that group. In such a system, conviviality maximizes benefits for a group, for instance, by standardizing the conventions of the groups'communications, conviviality contributes to the efficiency of processes and the achievement of the group's common goals.

The role of norms for conviviality reinforces social cohesion by reflecting the group's core values internally as well as externally. By making the rules explicit the role of norms for conviviality contribute to reducing conflicts, to optimize members' performances within communities as well as between communities and improve coordination throughout; All crucial for the development of ambient intelligence applications and coordination. Finally, the social warranty and protection mechanisms of conviviality are achieved through the expression of its group member's feelings toward each other: praise and encouragements for members who conform to the rules, anger and blame for the ones who do not. Such behavior coordination and regulation mechanisms are the very ones that underlie future ambient intelligent society and can therefore greatly gain by explicit conviviality specifications.

4 HUMAN COMPUTER INTERACTION (HCI) APPROACHES

According to Lamizet, conviviality was elaborated to describe both ”institutional structures that facilitate social relations and technological processes that are easy to control and pleasurable to use” [23]. On one hand conviviality allows individual expression facilitated by personalized interface and customized content while on the other hand it contributes to the standardization of media and the uniformization of representation systems. In her study of animated toys, Ackermann, looking at the relational qualities of playthings notes that beyond humanoid traits, it is an AniMate's manners of interaction that matter: ”Beyond smarts, it is its conviviality. Beyond obedience or bossiness, it is an AniMate's relative autonomy and ability to share control” [1]. Building on Illich's notion of conviviality based on individual freedom and role swapping [18], Ackermann explores partial and shared control as critical quality of conviviality.

4.1 Toward social intelligence

Markopoulos et al. identify four critical challenges to human computer interaction research for ambient intelligence components: ”Designing ambient intelligence systems and environments so that they can be perceived as socially intelligent . . . Designing intelligence that will support human-to-human cooperation and social interactions. . . How to evaluate social intelligence? . . . What are the benefits of social intelligence?” [28]. Answer to the last question would appear to be a requirement for the evaluation of social intelligence and for designing intelligence that will support social interactions. Markopoulos et al. experimenting with the iCat, a research platform that exhibits a rich set of human-like behaviors for studying social robotic user-interfaces, further state that for the ambient intelligence research community, the challenge ahead is: ”the need to make systems capable of understanding and relating to people at a social level, timing, and cuing their interactions in a socially adept manner” [28]. This are some of the challenges social intelligence design aims to address with ”methods of establishing the social context, embodied conversational agents, collaboration design, public discourse, theoretical aspects of social intelligence design, and evaluation of social intelligence” [30]. We note with interest the relation between social intelligence and conviviality particularly in application domains such as: collaborative environment, e-learning, community support systems, symbiosis of humans and artifacts and digital democracy. However, in our opinion, as the pervasiveness and role of technology increases in our society, so does the role of conviviality, whereas social intelligence seem to remain focus on general intelligence applied to social situations.

4.2 Artificial companions and Mixed-Initiative Interaction

The Companions that Wilks envisions [47] are persistent software agents attached to single users. They act as intermediaries for all information sources that users cannot manage. For instance, Companions for seniors provide company to senior citizens who feel lonely, they act as technical task assistant to search the web for travels or keep track of events their owners forget. Conversely, Companions for juniors provide assistance with teaching, explanations-on-demand and advices.

In a rather new area of research called mixed-initiative interaction ”people and computers take initiatives to contribute to solving a problem, achieving a goal, or coming to a joint understanding” [16]. A critical element is how users focus their attention: ”Attentional cues are central in decisions about when to initiate or to make an effective contribution to a conversation or project” [15]. Mixed-initiative research aims at developing software that filters appropriately incoming information to shield users from incoming disturbances such as emails and phone calls. The filtering of incoming information is achieved through measuring user’s keystrokes and scrolling activities, recording the number of opened windows, analyzing content, checking events in calendars, location and time of day and so on.

4.3 Conviviality as user experience for ambient intelligence scenarios

The goal, to design interfaces that are closer to the way human think than the way machine operate, raises questions such as: ”What is, at this very moment, the user’s state? What does s/he want, like, need, wish? Is s/he alone, at home, in family, with friends, at work [13]? In the context of such spontaneous interactions, innovative approaches based on dynamic notions such as conviviality, trust and behavior are required. Furthermore, in the area of the disappearing computer, ”the shift from information worlds to experience worlds” [43] is particularly significant. As stated by de Ruyter and Aarts, user experience for ambient intelligence must be based on: ”(i) safeguarding the privacy of the home environment, (ii) minimizing the shift of user attention away from the actual content being consumed and (iii) creating the feeling of being connected when consuming content over different locations” [10]. From individual social assistants to communications facilitators, numerous research directions in HCI exemplify the interest for cognitive and social input to address issues as wide apart as information clutter and digital divide. We believe that conviviality can be an important concept to help address the broad challenges of ambient intelligence, by providing mechanisms for adaptive user interactions, while preserving the granularity of human experience.

5 CONCLUSION

We summarize by first noting that conviviality is usually considered a positive concept related to sociability, however, further analysis reveals a negative side related to regulations. In this survey paper, we examine the multi-faceted concept of conviviality and raise the question: Which definition of conviviality can be used and made operational for ambient intelligence? We propose a two-fold definition of conviviality as a condition for social interactions and an instrument for the internal regulation of social systems. We then raise the questions: How can positive sides of conviviality be used for ambient intelligence and can negative sides be taken into account?

Ambient intelligence applications can greatly benefit from the positive aspects of conviviality: Sharing knowledge and skills, dealing with conflict, enabling inclusiveness and encouraging equality and trust among parties. However, conviviality has first to be expressed explicitly and formalized before it can be used, efficiently, as coordination mechanism between individuals, groups and institutions, and as a tool to reinforce social cohesion.

It is crucial to build into ambient intelligence applications designs, the necessary protections against the potentially negative sides of conviviality, such as deception, group fragmentation and reductionism. Intelligent interfaces, for example, allow instant interactions and thereby create strong needs for coordination and regulation mechanisms. These needs have to be addressed to ensure the safeguard of individuals against abuses, such as privacy intrusions and identity manipulations. Best practices and guidelines for designing ambient intelligence systems, must include aspects such as ensuring each party’s point of view, in order to avoid the crushing of one side by another. The concept of conviviality, because it allows to take into account the social and cognitive factors and ethical issues raised by large scale development of ambient intelligence systems and also points out the negative sides to be prevailed over, plays a crucial role for ambient intelligence.

ACKNOWLEDGEMENTS

We thank the City of Luxembourg for their financial support and Matthias Nickles, Joris Hulstijn, Guido Boella, Leon van der Torre, Emil Weydert, Pascal Bouvry, Catherine Pelachaud and the referees of this conference for useful suggestions.

REFERENCES

[1] Edith K. Ackermann, 'Playthings that do things: a young kid's "incredibles"!', in *IDC '05: Proceeding of the 2005 conference on Interaction design and children*, pp. 1–8, New York, NY, USA, (2005). ACM Press.

[2] Wendy Ashby, 'Unmasking narrative: A semiotic perspective on the conviviality/non-conviviality dichotomy in storytelling about the german other', *Trans, Internet journal for cultural sciences*, **1**(15), (2004).

[3] Guido Boella, Leendert van der Torre, and Harko Verhagen, 'Introduction to normative multiagent systems', *Computational & Mathematical Organization Theory*, **12**(2-3), 71–79, (October 2006).

[4] Guido Boella and Leendert W. N. van der Torre, 'Regulative and constitutive norms in normative multiagent systems.', in *KR*, eds., Didier Dubois, Christopher A. Welty, and Mary-Anne Williams, pp. 255–266. AAAI Press, (2004).

[5] Jürgen Bohn, Vlad Coroama, Marc Langheinrich, Friedemann Mattern, and Michael Rohs, 'Social, economic, and ethical implications of ambient intelligence and ubiquitous computing', in *Ambient Intelligence*, eds., W. Weber, J. Rabaey, and E. Aarts, 5–29, Springer-Verlag, (2005). Springer-Verlag.

[6] P. Caire, 'A critical discussion on the use of the notion of conviviality for digital cities', in *Proceedings of Web Communities 2007*, (2007).

[7] Sara Casare and Jaime Sichman, 'Towards a functional ontology of reputation', in *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pp. 505–511, New York, NY, USA, (2005). ACM Press.

[8] Paul Couderc, 'Collaborative capture: A new perspective for sensor networks', *ERCIM News*, **Embedded Intelligence**(67), (October 2006).

[9] Joelle Coutaz, James L. Crowley, Simon Dobson, and David Garlan, 'Context is key', *Commun. ACM*, **48**(3), 49–53, (2005).

[10] Boris de Ruyter and Emile Aarts, 'Ambient intelligence: visualizing the future', in *AVI '04: Proceedings of the working conference on Advanced visual interfaces*, pp. 203–208, New York, NY, USA, (2004). ACM Press.

[11] Ingrid Goetzl, 'Telecities - digital cities network', in *Revised Papers from the Second Kyoto Workshop on Digital Cities II, Computational and Sociological Approaches*, pp. 101–109, London, UK, (2002). Springer-Verlag.

[12] Eduardo Rodrigues Gomes, Elisa Boff, and Rosa Maria Vicari, 'Social, affective and pedagogical agents for the recommendation of student tutors', in *Proceedings of Intelligent Tutoring Systems 2004*, (2004).

[13] Tom Gross, 'Ambient interfaces for distributed work groups', *ERCIM News*, **Ambient Intelligence**(47), (October 2001).

[14] Wolfgang Hofkirchner, 'Unity through diversity.dialectics - systems thinking - semiotics', *Trans, Internet journal for cultural sciences*, **1**(15), (2004).

[15] Eric Horvitz, Carl Myers Kadie, Tim Paek, and David Hovel, 'Models of attention in computing and communication: from principles to applications.', *Commun. ACM*, **46**(3), 52–59, (2003).

[16] Eric Horvitz, Paul Koch, and Johnson Apacible, 'Busybody: creating and fielding personalized models of the cost of interruption.', in *Computer Supported Cooperative Work*, eds., James D. Herbsleb and Gary M. Olson, pp. 507–510. ACM, (2004).

[17] Ivan Illich, *Deschooling Society*, Marion Boyars Publishers, Ltd., 1971.

[18] Ivan Illich, *Tools for Conviviality*, Marion Boyars Publishers, August 1974.

[19] Toru Ishida, 'Understanding digital cities.', in *Digital Cities*, eds., Toru Ishida and Katherine Isbister, volume 1765 of *Lecture Notes in Computer Science*, pp. 7–17. Springer, (2000).

[20] Achilles Kameas and Irene Mavrommati, 'Extrovert gadgets', *Commun. ACM*, **48**(3), 69, (2005).

[21] Saadi Lahlou and Franois Jegou, 'European disappearing computer privacy design guidelines v1.0', Ambient agoras report d15.4., Disappearing Computer Initiative, (October 2003).

[22] Saadi Lahlou, Marc Langheinrich, and Carsten Roecker, 'Privacy and trust issues with invisible computers', *Commun. ACM*, **48**(3), 59–60, (2005).

[23] Bernard Lamizet, 'Culture - commonness of the common?', *Trans, Internet journal for cultural sciences*, **1**(15), (2004).

[24] Gloria T. Lau, Kincho H. Law, and Gio Wiederhold, 'Analyzing government regulations using structural and domain information.', *IEEE Computer*, **38**(12), 70–76, (2005).

[25] David G. Lawrence, 'Procedural norms and tolerance: A reassessment', *The American Political Science Review*, (1976).

[26] Helgo Lomosits, 'Future is not a tense', *Trans, Internet journal for cultural sciences*, **1**(15), (2004).

[27] Thomas W. Malone, Kevin Crowston, Jintae Lee, Brian Pentland, Chrysanthos Dellarocas, George Wyner, John Quimby, Charles S. Osborn, Abraham Bernstein, George Herman, Mark Klein, and Elissa O'Donnell, 'Tools for inventing organizations: Toward a handbook of organizational processes', *Management Science*, **45**(3), 425–443, (1999).

[28] Panos Markopoulos, Boris de Ruyter, Saini Privender, and Albert van Breemen, 'Case study: bringing social intelligence into home dialogue systems', *interactions*, **12**(4), 37–44, (2005).

[29] Incorporated Merriam-Webster, *Merriam Webster OnLine Dictionary*, Merriam-Webster, 2006.

[30] Toyoaki Nishida, 'Social intelligence design - an overview.', in *JSAI Workshops*, eds., Takao Terano, Toyoaki Nishida, Akira Namatame, Shusaku Tsumoto, Yukio Ohsawa, and Takashi Washio, volume 2253 of *Lecture Notes in Computer Science*, pp. 3–10. Springer, (2001).

[31] Magalie Ochs, Radoslaw Niewiadomski, Catherine Pelachaud, and David Sadek, 'Intelligent expressions of emotions.', in *Affective Computing and Intelligent Interaction*, eds., Jianhua Tao, Tieniu Tan, and Rosalind W. Picard, volume 3784 of *Lecture Notes in Computer Science*, pp. 707–714. Springer, (2005).

[32] Seymour Papert and Idit Harel, *Constructionism*, chapter 1, Cambridge, MA: MIT Press., 1991.

[33] Per Persson and Fredrik Espinoza, 'Geonotes: Social enhancement of physical space', *ERCIM News*, **Ambient Intelligence**(47), (October 2001).

[34] Michael Polanyi, *Personal Knowledge: Towards a Post-Critical Philosophy*, Routledge & Kegan Paul Ltd, London, 1958.

[35] Michael Polanyi, *Personal Knowledge : Towards a Post-Critical Philosophy*, University Of Chicago Press, August 1974.

[36] Robert D. Putnam, 'Diplomacy and domestic politics: The logic of two-level games', *International Organization*, **42**(3), 427–460, (1988).

[37] Robert D. Putnam, 'Bowling alone: the collapse and revival of american community.', in *Computer Supported Cooperative Work*, p. 357, (2000).

[38] M. David Sadek, Philippe Bretier, and E. Panaget, 'ARTIMIS: Natural dialogue meets rational agency', in *International Joint Conferences on Artificial Intelligence (2)*, pp. 1030–1035, (1997).

[39] Madeleine Schechter, 'Conviviality, gender and love stories: Plato's symposium and isak dinesen's (k. blixen's) babette's feast', *Trans, Internet journal for cultural sciences*, **1**(15), (2004).

[40] Arnan Sipitakiat, *Digital Technology for Conviviality: Making the Most of Students' Energy and Imagination in Learning Environments*, Master's thesis, MIT, Cambridge, MA,USA, 2001.

[41] Georgij Yu. Somov, 'Conviviality problems in the structure of semiotic objects', *Trans, Internet journal for cultural sciences*, **1**(15), (2004).

[42] Constantine Stephanidis, 'A european ambient intelligence research facility at ics-forth', *ERCIM News*, **Embedded Intelligence**(67), (October 2006).

[43] Norbert Streitz, Carsten Magerkurth, Thorsten Prante, and Carsten Roecker, 'From information design to experience design: smart artefacts and the disappearing computer', *interactions*, **12**(4), 21–25, (2005).

[44] Norbert Streitz and Paddy Nixon, 'Introduction', *Commun. ACM*, **48**(3), 32–35, (2005).

[45] Millie Taylor, 'Oh no it isn't: Audience participation and community identity', *Trans, Internet journal for cultural sciences*, **1**(15), (2004).

[46] Mark Weiser, 'The computer for the 21st century', *Scientific American*, 66–75, (September 1991).

[47] Yorick Wilks, 'Artificial companions.', in *Machine Learning for Multimodal Interaction*, eds., Samy Bengio and Hervé Bourlard, volume 3361 of *Lecture Notes in Computer Science*, pp. 36–45. Springer, (2004).

[48] Michael Wooldridge, 'An introduction to multi-agent systems.', *J. Artificial Societies and Social Simulation*, **7**(3), 16–23, (2004).

[49] David Wright, 'The dark side of ambient intelligence', *The journal of policy, regulation and strategy for telecommunications*, **7**(6), 33–51, (2005).

Abstraction as a Tool for Multi-Agent Policy Evaluation

Krysia Broda and **Christopher John Hogger** [1]

Abstract. Abstraction is a valuable tool for dealing with scalability in large state space contexts. This paper addresses the design and evaluation, using abstraction, of good policies for minimal autonomous agents applied within a situation-graph-framework (SGF). In this framework an agent's policy is some function that maps perceptual inputs to actions deterministically. A good policy disposes the agent towards achieving one or more designated goal situations, and the design process aims to identify such policies. The agents to which the framework applies are assumed to have only partial observability, and in particular may not be able to perceive fully a goal situation. A further assumption is that the environment may influence an agent's situation by unpredictable exogenous events, so that a policy cannot take advantage of a reliable history of previous actions.

Communicable knowledge of other agents is modelled in SGF by enriched perceptions, allowing co-operation despite the agents' minimality. Two different abstractions are described in this paper, situation abstractions and group abstractions. The latter are appropriate when several agents are operating. The Bellman discount measure provides a means of evaluating situations and hence the overall value of a policy and this paper describes experiments that test the accuracy of the method in the presence of each kind of abstraction. A modification for situation abstractions is described and its power is demonstrated through comparison with simulation results.

1 Introduction

Our interest is in designing good policies for particularly simple autonomous agents. A good policy disposes the agent towards achieving one or more designated goal situations, and the design process aims to identify such policies. The simplest case is a memoryless reactive agent whose policy consists solely of some function that maps perceptual inputs to actions deterministically. We also include modest extensions such as inclusion of finite memory, wireless communication and nondeterministic (relational) policies. The term *minimal agent* will be used loosely here to cover both the simplest case and these near-minimal extensions. The focus on minimal agents anticipates application contexts where physical or economic constraints make it impractical to deploy more sophisticated cognitive agents embodied in correspondingly sophisticated hardware. Examples include remote exploration and medical nano-robotics where the desired goals may be achievable by a large community of physically small and inexpensive primitive agents among which occasional losses and dysfunctionalities can be readily tolerated.

Our design method is based on discounted-reward analysis [11] applied to a directed policy graph whose arcs represent the transitions that occur under the policy being considered. Each transition takes the agent from some situation – a $(state, perception)$ pair – to some successor situation. The analysis yields an overall policy value that will depend upon, *inter alia*, the designated goal situation(s) and whatever rewards and probabilities are assigned to the graph's arcs by the designer. The method is called the *situation-graph-framework* (SGF) to reflect its reliance upon explicit situation graphs, and was first reported in [1]. The SGF framework can be distinguished from both the standard MDP and POMDP frameworks [7, 4]. It is non-Markovian, since an agent's next perception is conditional upon more than its current perception and action. The design process makes use of the full state, through the use of situations. The agents to which the framework applies are assumed to have only partial observability, and in particular may not be able to perceive fully a goal situation. This feature distinguishes the framework from other design methods that rely upon complete goal observation. The formulation in [6], where an agent's perception is treated as the state, is a special case of SGF. The POMDP framework uses an estimation of the distribution of the full state, called a belief state, to guide the planning process [4]. This can result in policies in which the action taken when perceiving p may implicitly depend on the route taken to p – that is, an agent may follow a policy expressed by a graph. Agents in SGF are not equipped to follow such policies because they are designed for use in communities of agents, where unexpected events are the norm. The core assumption in using "belief states" is that remembrance of the past is a reliable basis on which to estimate an agent's current situation, which is a safe assumption in the specific circumstance that the environment can be impacted only by that agent. This assumption will not hold if the environment can be additionally impacted by exogenous events, including the actions of other agents. SGF represents such events by so-called x-arcs in special situation graphs called viewpoint graphs and employs a particular elaboration of the discounted-reward analysis to deal with them. Experimental evaluation of policies designed in this way for communities of non-cooperating *cloned* agents was first presented in [2] and demonstrated strong empirical evidence for the efficacy of the design process. Some initial theoretical results pertaining to approximations of group abstractions were given in [3]. Designing policies for groups of reactive agents by quantitative evaluation of their situation graphs stands in contrast with frameworks like the Go! formalism [9] in which rational co-operating agents are freely programmed with arbitrarily complex reasoning capabilities and whose overall efficacy is not subjected to quantitative evaluation.

Whether dealing with single agents or communities, SGF – like the other frameworks – must in general confront the issue of scalability. The key to this is appropriate use of *abstraction*. Broadly speaking, abstraction amounts to ignoring many minor distinctions – such as between states, perceptions or agents – that are considered unlikely to have a significant bearing upon outcomes. It achieves this by collecting similar concrete entities (e.g. states) into single generic entities which then become the first-class elements of the formulations used

[1] Department of Computing, Imperial College London, email: {kb,cjh}@doc.ic.ac.uk

in the design process. The SGF viewpoint treatment just mentioned is one such kind of abstraction, in that an x-arc in a viewpoint graph signifies an event instigated by some other agent but without specifying which particular agent, and so avoids the explicit and cumbersome multi-agent vectors that some MDP designers [8] have resorted to in order to deal with communities. The other kind of abstraction we use is *situation-abstraction*, that is, the abstraction of both states and perceptions. This can be applied irrespective of whether one is dealing with one agent or many. Its first benefit is to reduce the size of the situation-graphs being dealt with and hence to ameliorate the burden of estimating the probabilities on their arcs. Its second benefit is that abstraction of perceptions reduces the size of the policy-space over which optimization is pursued, since the number of possible policies depends upon the number of possible perceptions.

Section 2 outlines the features of SGF and two kinds of abstraction for dealing with scalability. Section 3 describes situation-abstraction and explains how its use produces inaccurate predictions of policy value if one relies upon the standard discounted-reward procedure. Section 3.2 describes a new modification of that procedure to reduce the inaccuracies and Section 3.3 gives simulation results for two case studies to show the resulting improved predictive power. Section 4 describes group abstraction and gives simulation results to illustrate the accuracy when applied to co-operating agents. Section 5 concludes with an assessment of the method and future work.

2 Basic Features of SGF

A simple example serves to illustrate the basic features of SGF. It assumes a world consisting of planks and agents. A plank may be held at either end by one agent or at both ends by two different agents. Therefore each state is representable by a triple $[r, t, f]$ in which r is the number of (raised) planks held at both ends, t is the number of (tilted) planks held only at one end, and f is the number of (flat) planks held at neither end. An agent is equipped to perceive in any state that it is holding an end (h) or not (nh), and that it is seeing one of the following: an unheld end (su), a held end (sh) or no end ($s0$).[2] For 2 planks and 2 agents this gives a total of 16 situations, each being a pair (o, p) where o is a state and p is a perception that the agent may have of o.

	$o = [r, t, f]$
0	[0, 0, 0]
1	[0, 0, 1]
2	[0, 1, 0]
4	[0, 0, 2]
5	[0, 1, 1]

	p	$A(p)$
a	s0, nh	{w}
c	su, nh	{li, w}
f	sh, h	{dr, di}

Figure 1. States, perceptions and actions for Planks World

A formulation of this kind induces a *situation graph*. To convey the essential features of such a graph it is convenient to use a simplified example which considers a world with 1 agent and 2 planks. Figure 1 shows the possible states and perceptions. For compact presentation we use just these labels to denote situations: $0a$, $1a$, $1c$, $2f$, $4a$, $4c$, $5f$. Figure 1 also shows for each perception p the set $A(p)$ of actions the agent might perform when perceiving p. Altogether there are 4 kinds of action: `lift`, `drop`, `dispose` and `wander`. The actions `lift` and `drop` refer to an agent raising and lowering one end of a plank. The action `dispose` refers to an agent trying to discard a plank one of whose ends it is holding. The w (wander) action refers to the agent updating its perception of the state; this includes the reflexive case of maintaining its current perception. The w action is the only action to leave the state invariant.

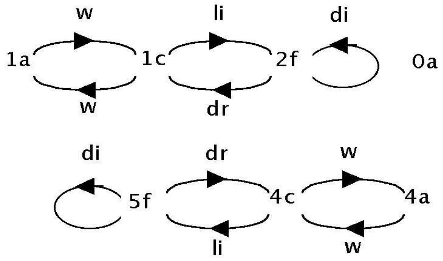


Figure 2. Complete transition graph

Figure 2 shows the corresponding situation graph. If only one end of a plank is being held (situations $2f$ and $5f$) and a `dispose` action is performed, then the situation remains unchanged. A policy for the agent is a total function from perceptions to actions. The total number of possible policies is the product of the cardinalities of all the $A(p)$ action sets, which in the present case is 4. They include, for instance, the policy $\{a \rightarrow \texttt{w}, c \rightarrow \texttt{li}, f \rightarrow \texttt{di}\}$. Each one corresponds to a restriction of the complete graph whereby the arcs emerging from all situations sharing a common perception all bear the same action label. Each policy has a value determined by its situation graph and assignments of probabilities and rewards to its arcs as described later. In the case that the ultimate goal of the policy is to clear away a collection of planks, it is obvious that one agent cannot achieve this. Section 4 extends the example to encompass multiple and co-operating agents.

In the above account the situations in the formulation were taken to be *concrete*, that is, corresponding one-to-one with the real situations arising in the problem domain. In the general case there may be large numbers of situations and/or agents which, if taken without simplification, would present serious scalability problems in policy evaluation and ranking. Therefore, the purpose in what follows is to show two aspects of such simplification, which we refer to as *Situation Abstraction* and *Group Abstraction*. Situation abstraction occurs when each generic situation in the formulation is chosen to represent many concrete situations and is discussed in Section 3. The natural extension of SGF to deal with multiple agents is to construct *multi-situations*, which represent for any concrete state the combined individual perceptions. Such a structure will in general be very complex and lends itself to simplification by considering transitions from the point of view of one agent which we refer to as *self*. We introduce *Group Abstraction* for this purpose and discuss it in Section 4.

3 Situation Abstraction in SGF

Abstraction in SGF partitions the set of concrete states into subsets called generic states and partitions the set of concrete perceptions into subsets called generic perceptions. A generic situation (O, P) is a pairing of a generic state O with a generic perception P. This abstraction process is required to satisfy the following constraints:

1. if (o, p) is a concrete situation then there must exist exactly one generic situation (O, P) such that $o \in O$ and $p \in P$;

[2] It is assumed that if an agent is holding an end then it must see a held end.

2. if (O, P) is a generic situation then it must contain at least one concrete situation (o, p) where $o \in O$ and $p \in P$;
3. if P is a generic perception then $\bigcap\{A(p) | p \in P\} \supseteq A(P)$.

Here, 1) and 2) ensure that the sets of concrete states and perceptions are partitioned such as to result in a partitioning of the complete set of concrete situations, whilst 3) ensures that every generic perception offers at least one action among those offered by each of its concrete members.

With fewer situations to deal with at the abstract level than at the concrete level, the equations relating situation values for any given policy are correspondingly fewer. Perhaps more importantly, having fewer perceptions at the abstract level than at the concrete level reduces the number of policies to be evaluated. Intuitively, a good abstraction is one whose discounting of differences at the concrete level yields a ranking of abstract policies that is approximately commensurate with the ranking of the concrete policies that would be obtained from concrete analysis. Given any situation abstraction in SGF, we will show how the standard Bellman formula (as detailed presently) incurs inaccuracy when applied to it, and how to obtain improved accuracy by modifying that formula.

3.1 Illustration

We illustrate these issues with an example, again assuming a single agent. The environment in which this agent operates is called *Token World* and contains some fixed number N of tokens. It is organized as one or more heaps of tokens together with a single region named *void* in which there are no tokens. As its perception, the agent always sees either a heap or *void* and always knows whether or not it is holding a token. Its possible actions are just the following: `gr`: grab a token from a heap; `dr`: drop a token onto a heap or onto *void*; `w`: wander. An agent dropping a token onto *void* thereby creates a new 1-token heap leaving *void* preserved, whilst an agent grabbing the token from a 1-token heap merely eliminates that heap. Using our single *void* representation, an agent dropping a token onto *void* thereby creates a new 1-token heap leaving *void* preserved, whilst an agent grabbing the token from a 1-token heap merely eliminates that heap. Prior to performing `gr` the agent must be not holding and seeing a heap, and is afterwards holding a token and seeing the reduced heap. Prior to performing `dr` it must be holding a token and seeing a heap or *void*, and is afterwards not holding and seeing the heap containing the token just dropped. Prior to performing `w` it can be holding or not and seeing anything, and is afterwards seeing a heap or *void* with its (not)holding status preserved.

Figure 3 shows some legal transitions in the case that $N = 10$. Taking the above notions to define the concrete representation, a concrete formulation of the complete situation graph would entail 72 states, 21 perceptions, 236 situations and 2^{20} policies. The goal will comprise some specified configuration of heaps and some perception, not necessarily perceivable in its entirety by the agent.

In order to compare meaningfully the possible policies of agents we need to discriminate between situations. In practice this is achieved by assigning numeric rewards to the graphs arcs. A situation regarded as a goal is one whose incident arcs are assigned the highest reward. We now consider one possible abstraction for the case when the goal is to achieve a configuration having exactly 3 identical heaps and not exactly 2 identical heaps, with the agent seeing *void* and not holding. The concrete states are partitioned into abstract states 1-4 and the concrete perceptions into abstract perceptions $a - h$:

1. exactly 3 identical heaps and not exactly 2 identical heaps

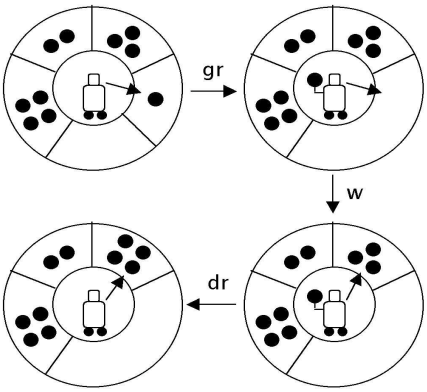


Figure 3. Transitions in *Token World*

2. exactly 2 identical heaps and not exactly 3 identical heaps
3. the heaps are all different
4. all other cases

Examples of states 1 and 2 are $\{2, 2, 2, 1, 1, 1, 1, void\}$ and $\{2, 2, 1, 1, 1, 1, 1, 1, void\}$ respectively.

$a(b)$. sees a heap of size < 4 and holding(not holding)
$c(d)$. sees a heap of size > 4 and holding(not holding)
$e(f)$. sees a heap of size $= 4$ and holding(not holding)
$g(h)$. sees *void* and holding(not holding)

The goal is therefore $(1, h)$. There are then 32 abstract situations in the graph and just 128 policies to evaluate. Each of the latter maps abstract perceptions to actions and is therefore an abstract policy spanning some set of concrete policies. In effect, the abstraction process partitions the concrete policy space as well as the concrete situation space. Concrete simulations indicate that the optimal abstract policy is $\{a \rightarrow \texttt{dr}, b \rightarrow \texttt{w}, c \rightarrow \texttt{dr}, d \rightarrow \texttt{w}, e \rightarrow \texttt{dr}, f \rightarrow \texttt{w}, g \rightarrow \texttt{dr}, h \rightarrow \texttt{w}\}$. The optimal policy value of 6.74 was computed using the Bellman formula below applied to the simulation runs.

The value of a policy partly depends upon the chosen goal situation. The status of the chosen goal is reflected in the assignment of numerical rewards to the arcs in the complete graph, each being a measure of the supposed benefit/disbenefit of effecting the associated transition. We might, for instance, assign a large positive reward $R = 100$ to each of the two arcs leading to $6b$ and a small but negative reward $r = -1$ to every other arc. A policy's value partly depends also upon the probabilities assigned to the arcs.

Once the assignments of rewards and probabilities are in place, a value for each situation s under a given policy can be calculated from $V(s) = \Sigma_{u \in SS}(P_{su} \times (\Upsilon_{su} + \gamma \times V(u)))$, the standard Bellman formula, in which SS is the successor set of s, Υ_{su} is the immediate reward for the action that takes s to u, P_{su} is the probability that from s the agent proceeds next to u and the parameter $0 < \gamma < 1$ is a discount factor that ensures the resulting set of linear equations has a unique solution. If it is assumed that the agent may begin its activity at any situation then the overall policy value is just the mean of the situation values.

The standard Bellman formula takes the probabilities on the arcs to be Markovian: the value of any situation is calculated using the expectation over its emergent arcs without regard to how that situation was reached. Consider situation $1a$ in the graph for some policy in which $a \rightarrow \texttt{w}$ and $c \rightarrow \texttt{w}$. The successors of $1a$ are then

$[1a, 1c, 1e, 1g]$. If the probabilities are estimated simply as the mean, over all concrete instances i of $1a$, of the probability that i can transit by w to these successors, their values are about [0.75, 0.025, 0.042, 0.183]. This takes the view that if the agent arrives by any means at $1a$ then the probability that it will next wander to $1e$ is 0.042. But this is not the case. Had the agent arrived at $1a$ from $1c$, for example, the probability of it next wandering to $1e$ would be zero: the only concrete state in which the agent can transit by w from $1c$ to $1a$ is $\{6, 1, 1, 1, \textit{void}\}$ in which it is impossible for the agent to wander to see a heap of size 4 (perception e).

Therefore, if the formula is applied to an abstract policy graph with probabilities estimated as just indicated above, it will perceive paths through the graph that might not be concretely traversable at all or traversable with quite different probabilities, yielding misleading policy values. We refer to this property of the paths as *piecewise incoherence*. Next we discuss our modification of the formula with the aim of ameliorating this deficiency.

3.2 Policy Evaluation Methods

The inaccuracies from piecewise incoherence in the abstract context can be reduced by modifying the standard Bellman formula. As it stands, the latter yields a set of linear equations expressing each value $V(i)$ of abstract situation i in terms of the values of the successor set $SS(i)$ of i. The first stage of our modification reformulates $V(i)$ as

$$V(i) = \Sigma_{j \in SS(i)} p_{ij}(r_{ij} + \gamma V(j|i))$$

where $V(j|i)$ is the contribution made to $V(j)$ by all those concrete transitions that transit to j from i, and p_{ij} is the average probability of those transitions. This stage therefore introduces a new set of conditional variables of the form $V(j|i)$. The second stage, which is somewhat more subtle, inter-relates these new variables as follows:

$$V(j|i) = \Sigma_{k \in SS(j)} p_{ijk}(r_{jk} + \gamma V(k|j))$$

where p_{ijk} is the probability that k is reached from i via j. Together these two formulations yield a new set of linear equations, involving more variables and probabilities than before, from which the various $V(i)$ values can be calculated. The abstract policy value is then again the mean of these. The modification thus gives recognition to the fact that, in the abstract context, the value of a situation has a non-Markov dependence upon its immediate predecessors.

Even significantly abstract formulations may, with this modification, offer a large number of equations, so it is important that an efficient procedure be used to extract the best policies. For this we use a branch-and-bound algorithm, adapted from Littman [5], which, for some number $n > 0$, develops a tree of partially-constructed policies whilst pruning those that could not - if fully extended to become complete policies - be among the n highest-value policies. To find only some optimal policy, n is chosen as 1.

3.3 Empirical Results

We have applied the modified treatment just described to a range of examples over different domains and have observed in all these cases an improvement in the correlation between predicted and simulated policy values. We illustrate this for two goals in *Token World*.

Goal-1: Achieving 3 Identical Heaps Here the example is that described in Section 3.1, where the world has 10 tokens and the goal is to transform any initial situation to one having exactly 3 identical heaps but not exactly 2, with the agent seeing *void* and not holding. We evaluated all the 128 policies using first the standard Bellman equations and then our modified equations. The probabilities were calculated by analysing all the concrete transitions. The predicted policy values were then compared with the results of simulating all the policies. Each one was run 500 times, with random initialization, for up to 50 transitions. (For economy, any run achieving the goal was terminated at that point, and to reflect this in the prediction we suppressed the goal's emergent arcs). The reward parameters were $R = 100$, $r = -1$ and $\gamma = 0.9$.

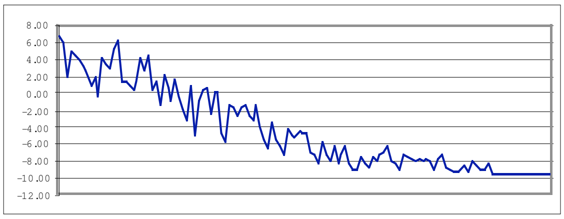


Figure 4. Using standard Bellman for Goal-1

Figure 4 charts the simulation values (vertical axis) against the increasing ranks of the values predicted by standard Bellman (horizontal axis), so that the predicted-best policies are on the left. If the prediction were perfect the chart would decrease monotonically from left to right. The correlation between predicted and simulated values is measured by the Kendal coefficient Q as a percentage ranging from 0 (worst) to 100 (best). For Figure 4, Q is 91.4% over all 128 policies and 69.5% across the first 20. Figure 5 shows the results using the modified equations. There, Q is 94.8% for all 128 and 78.4% for the first 20. In both cases the predicted optimal policy is the one that is optimal in simulation.

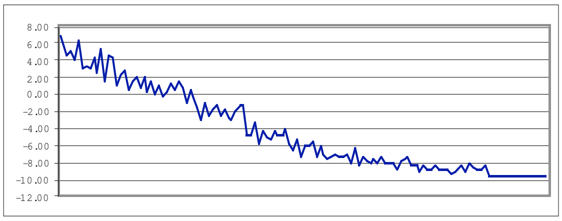


Figure 5. Using the modified equations for Goal-1

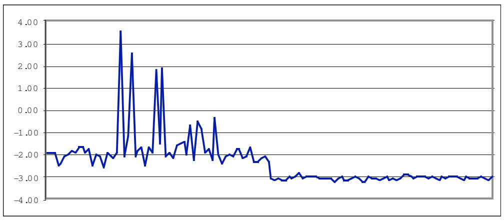


Figure 6. Using standard Bellman for Goal-2

Goal-2: Achieving Exactly 1 Heap Here there are again 10 tokens but the goal is the much harder one of arranging them into a single heap. For this problem we used a different abstraction, partitioning the states into 4 cases: one heap of 10; two heaps of 5; exactly one heap of 5 plus anything else; no heap of 5 or of 10. The perceptions were partitioned according to whether the agent was seeing *void*; seeing a heap less than 5; equal to 5; greater than 5. This yields 16 abstract situations and 128 policies.

Proceeding now as in the previous case, Figure 6 shows the chart using standard Bellman prediction, where Q is 66.4% for all 128 policies but only 31.1% for the first 10. Figure 7 shows the chart using the modified equations, where Q is 64.7% (not quite as good) for all 128 but 66.7% (radically improved) for the first 10. Moreover, the best policy from simulation is now predicted as best, whereas in the standard prediction it is ranked 21.

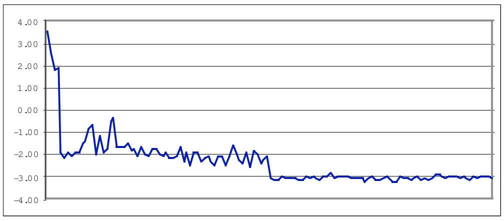


Figure 7. Using the modified equations for Goal-2

4 Group Abstraction in SGF

In a multi-agent context, an action performed by any one agent potentially updates the perceptions of other agents, so effecting a transition by the group as a whole. Without abstraction, the complete situation graph then depicts all such possible group transitions and is clearly a complex structure to deal with. Group abstraction projects this structure onto any one agent *self* to yield a *viewpoint graph*. As in the normal single-agent case, this viewpoint graph shows the transitions that *self* may effect by its own actions but also shows the transitions it may undergo exogenously by the actions of other agents. This kind of abstraction is simplest to achieve when the agents are all clones following the same policy, in which case the exogenous actions all bear the special label `x`, interpreted as "owing to an action by some other agent". This label is admitted into the action-set for *self*. Then, if a policy for *self* specifies `x` as the response to some perception, this is interpreted as making *self* wait. By this means, policies can control the manner in which agents variously act themselves or defer to the actions of others, a key aspect of co-operative behaviour. We next show how these ideas apply in a multi-agent Planks World and, in particular, how co-operation is mediated by the use of enriched perceptions representing communicable knowledge.

If non-communicating agents appear to co-operate in a task then this is a fortituitous manifestation of emergent behaviour, which we call "as-if" co-operation. This section shows how deliberate (planned) co-operation can be obtained by enabling agents to communicate. We avoid the need to devise special languages and protocols for this by restricting the communicable elements to be perceptions of the kind already employed. If there is an atomic perception p then we can allow another agent r_i to have the atomic perception kp (representing "knows" p), whose meaning is that one or more other agents are perceiving, and communicating to r_i, that p is true. We assume that the content of p is instantly transmissible from those other agents to r_i by some suitable broadcasting mechanism. With this provision in place, policy formation for r_i can then take account of what it receives from other agents in addition to its own direct perceptions.

Consider again the 2-plank problem of Section 2, now extended to more than 1 agent. Without co-operation, neither agent can perceive what the other is seeing, and so when each agent is holding one end of a plank they cannot tell whether they are holding the same plank or not. This means that the `dispose` action is liable to have null effect. Adding additional agents, but maintaining their limited perception, increases the chance of success when attempting a `dispose` action. Some of the null `dispose` attempts can be avoided if some communication between agents is allowed.

We can allow for agents to communicate to each other, in a limited way, by giving them certain useful percepts of the same form as those *self* could have. In this example, we choose to let an agent perceive whether no other agent is holding a plank end (knh), or whether at least one other agent is doing so (kh).

	$o = [r, t, f]$
0	[0, 0, 0]
1	[0, 0, 1]
2	[0, 1, 0]
3	[1, 0, 0]
4	[0, 0, 2]
5	[0, 1, 1]
6	[0, 2, 0]
7	[1, 0, 1]

	p	$A(p)$
a	s0, nh, knh	{`w`, `x`}
b	s0, nh, kh	{`w`, `x`}
c	su, nh, knh	{`li`, `w`, `x`}
d	su, nh, kh	{`li`, `w`, `x`}
e	sh, nh, kh	{`w`, `x`}
f	sh, h, knh	{`dr`, `x`}
g	sh, h, kh	{`di`, `dr`, `x`}

Figure 8. States, perceptions and actions

Thus the original example will now be extended in two ways: (i) by allowing communicated perceptions, and (ii) by allowing either two or three agents. The particular states and perceptions for this formulation for two agents are given in Figure 8. The *self*-restricted graph G_c for two agents and the policy $\{a \to$ `x`, $b \to$ `w`, $c \to$ `li`, $d \to$ `li`, $e \to$ `w`, $f \to$ `x`, $g \to$ `di`$\}$ is shown in Figure 9.

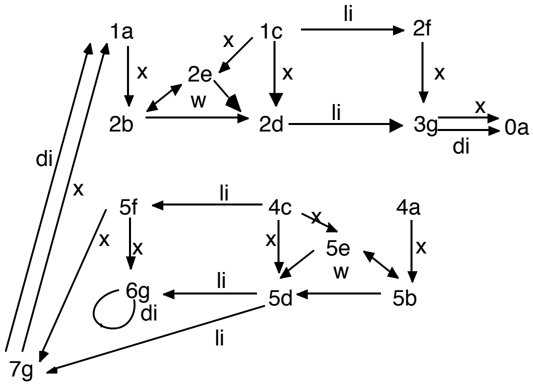


Figure 9. The policy graph G_c for 2 agents

Note that reflexive `x`-arcs are omitted to avoid clutter. In case there are three agents the unrestricted graph has an additional state, namely $8 = [1, 1, 0]$, i.e. the case in which one plank is raised and the second tilted, and 9 extra situations, namely $(3, b)$, $(3, e)$, $(6, b)$, $(6, d)$, $(6, e)$, $(7, b)$, $(7, d)$, $(7, e)$ and $(8, g)$ and numerous additional `x`-arcs. The simulator was run for the problems of disposing of two planks with either two or three agents. The particular policy illustrated in Figure 9 was ranked by the predictor as second for both the 2-agent case and the 3-agent case. The simulator ranked the policy, respectively, as fourth and third. More significantly, the observed (simulated) values of the policy were, respectively, 18.33 and 30.55.

Informally, the policy is to wait, rather than wander, except when there is a chance some agent may be holding the other end of a plank, in which case the policy is to wander. If the agent finds itself holding an end, it waits for another agent to lift the remaining end, ready for a dispose action. When there are more agents this policy has a better chance of success, hence the higher observed value (30.55) for 3 agents. The charts for the 2-agent and 3-agent case are shown in Figs. 10 and 11, respectively.

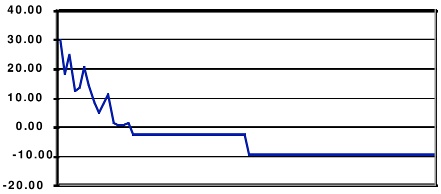


Figure 10. Policy chart for 2 agents

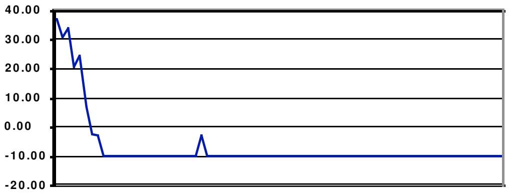


Figure 11. Policy chart for 3 agents

Using communication in this manner is semantically equivalent to increasing an agent's ability to perceive by its own means more of the state. Viewed in this way, we could have cast the new percepts $a : (s0, nh, knh)$ and $b : (s0, nh, kh)$ as $a : (s0, nh, nh1)$ and $b : (s0, nh, h1)$ where $h1$ and $nh1$ stand, respectively, for the other agent is, or is not, holding, on the assumption that an agent was physically equipped to know what the other agent was holding. In engineering terms, however, it is more practical to broadcast perceptions through one uniform technology than to equip agents with a diverse range of sensors. For more than one agent this approach introduces more complex perceptions, including disjunctions, so although the two views are equivalent, we prefer the one we interpret as communication.

5 Discussion and Conclusion

Policy design frameworks can be compared in terms of their tradeoffs between ease of problem formulation, complexity of policy optimization and predictive accuracy. Some are not directly comparable as they assume different agent architectures. MDP/POMDP methods assume agents capable of holding and consulting perception-action graphs whose paths represent the episodes an agent may experience only when undisturbed by exogenous events. By contrast, our agents presume a simpler policy structure and maintain optimal behaviour in all situations whether these have arisen by their own actions or not. The backward-planning design method of Nilsson [10] for teleoreactive agents also assumes a different agent architecture: there, the agent must have sufficient observability to take the best action in any situation and, crucially, in a goal situation. In many realistic contexts, however, the state component of a goal is too delocalized to be fully perceivable. All the above methods, including SGF, involve estimating probabilities, in contrast with those based upon learning. Q-learning [11, 6] can discover optimal policies having our structure but, like MDP methods, requires agents to have full observability. SGF has the distinctive feature that probability estimation is a once-only task for the given complete situation graph, independently of all policies and goals that might then be considered. In a POMDP framework each change of goal demands an evaluation of a new set of belief state probability distributions to find an optimal policy for achieving it.

The use of abstraction in SGF assumes that policy ranking is not overly sensitive to the small variations between the concrete situations spanned by an abstract one. The equations we use are expected to deliver for each abstract policy a value to which all its concrete policy instances closely approximate. Our simulation studies of abstraction using the standard Bellman equations showed that they cannot be relied upon to have this property. However, the modified equations in the cases we have tested, including those given here, have manifested this property. In future work on SGF we shall investigate how robust the property is in relation to the choice of abstraction. We will also investigate the use of both abstractions together to both homogenous and heterogenous communities of agents.

The more general conclusion of our work is that very simple agents can be mechanically designed and optimized so as to enable effective solution of communal goals, provided that suitable techniques such as abstraction are employed to address the issue of scalability. In those contexts that impose strong limitations upon physical architecture such agents may provide a realistic alternative to more elaborate rational agents, whose cognitive mechanisms present greater physical overheads besides compelling the agents to probe through large search spaces. An additional benefit of a framework like SGF is that it provides precise quantitative measurement of how well its agents perform.

REFERENCES

[1] K. Broda, C.J. Hogger and S. Watson, Constructing Teleo-reactive Robot Programs, *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI-2000)*, Berlin, pp. 653-657, 2000.

[2] K. Broda and C.J. Hogger, Policies for Cloned Teleo-Reactive Agents, 2nd Conference on Multi-Agent System Technologies, Ehrfurt, LNAI, 3187, Springer Verlag, pp. 328 - 340, 2004.

[3] K. Broda and C.J. Hogger, Abstract Policy Evaluation for Reactive Agents, *SARA-05, 6th Int. Symposium on Abstraction, Reformulation and Approximation*, Springer, LNAI 3607, pp. 44-59, 2005.

[4] L.P. Kaelbling, M.L. Littman and A.R. Cassandra, Planning and acting in partially observable stochastic domains, *Artificial Intelligence*, 101, pp. 99-134, 1998.

[5] M. L. Littman, Memoryless policies: theoretical limitations and practical results, *Proceedings of the 3rd International Conference on Simulation of Adaptive Behaviour* , MIT Press, pp. 297-305, 1994.

[6] J. Loch and S. Singh, Using Eligibility Traces to find the Best Memoryless Policy in Partially Observable Markov Decision Processes, *Proceedings of the 15th International Conference on Machine Learning*, pp. 323-331, 1998.

[7] T. Mitchell, *Machine Learning*, McGraw Hill, 1997.

[8] R. Nair, M. Tambe, M. Yokoo, D. Pynadath and M. Marsella, Taming Decentralised POMDPs: Towards Efficient Policy Computation for Multiagent Settings, *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*, pp. 705-711, 2003.

[9] K. L. Clark and F. G. McCabe, Go! for Multi-Threaded Deliberative Agents, *Annals of Mathematics and Artificial Intelligence*, Vol. 41, pp.171-206, 2004.

[10] N.J. Nilsson, Teleo-Reactive Programs and the Triple-Tower Architecture, *Electronic Trans. on Artificial Intelligence*, 5, pp. 99-110, 2001.

[11] R.S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 1998.

CyberCare: Reasoning about Patient's Profile in Home Healthcare

Alessandra Mileo[1] and **Davide Merico** and **Roberto Bisiani**[2]

Abstract. This paper proposes a framework based on modern tools and technologies to enable home healthcare through the observation of i) patient's clinical details by means of a Wearable Acquisition Device, ii) movements detected by sensors networks and iii) habits/actions inferred by an ASP logic program.

1 Introduction: the social and clinical context

In recent years, Communication and Information Technologies have been introduced in specific fields of medical sciences in order to allow the *delivery of clinical care*. A centralized view of medicine at a distance led to the integration of Communication Technologies and Clinical Decision Support Systems (CDSS) performing medical reasoning tasks. The resulting framework is known as Telemedicine.

Contrary to all expectations, the introduction of these techniques in real contexts showed they were not easily applicable to any medical task, and they have thus been restricted to some aspects of healthcare, such as medical prescriptions [17].

The limitations of Telemedicine for more general medical tasks, e.g. the formulation of a diagnosis, are related to the absence of standard protocols for data exchange and memorization between the Health Institutions.

Recent studies about the acceptance of technologies for the elderly [9, 6] showed that the best technological solution has to be chosen depending on the specific problem at hand, thus confirming previous research on this topic [15]. Moreover, while people tends to look for social relationships in activities such as cleaning or playing cards, in situations related to safety or health and personal care they are more likely to change their home environment in order to cope with their hierarchy of needs.

These considerations strengthen the argument that the success of intelligent technologies in healthcare depends on the need of each patient, and even on a given set of conditions under which he seems to need help.

In our proposal, we do not focus on the introduction of robots to domicile healthcare: beyond their high costs of set up and maintenance, their presence is rather intrusive and their acceptability is strongly related more to emotional components of people's image of them than to the effective help they can provide. Our perspective is shifted on *tools* and *technologies* that can unobtrusively help users and interact with them to increase home safety and personal healthcare. In order to do that, the interaction and communication between the patient and the system should be as much intuitive and transparent as possible through suitable interfaces and voice as well as other multimedia content processing.

We use ZigBee-sensors networks instead of classic monitoring tools (e.g. cameras), because they are simpler, faster and cheaper to install and use in existing domotic environments.

Our idea of CyberCare is based on an Intelligent Component aimed at collecting (temporally) local information about the patient's profile (medical parameter, clinical setting) and context (habits, localization), reasoning about it to determine the intervention required and provide the specialist with the context of the emergency, thus saving medical experts' time in determining how to operate.

Similar solutions to home healthcare have been proposed so far, such as the RoboCare project [4] and the KGP agent model [16], but they were mainly based on efficient and adaptive planning to monitor patient's daily activities. CyberCare is rather based on subsequent inferences to detect emergencies related to any single activity as they arise. In our reasoning model, a transition is referred to the state of an action (interrupted, abnormal, changed) and each inference allows to change this state. Patient's profile is updated through off-line reactivity capabilities, on the basis of the state transitions of each daily action. Nonetheless, we focus on single activities to detect emergencies when necessary, rather than considering a global daily plan that has to be monitored as a whole. In this way, thanks also to real time localization, any emergency related to a specific action execution can be treated faster as a single entity, while still keeping track of it to eventually update patient's habits off-line.

For these reasons, we argue that the introduction of modern sensors networks technologies as well as an Intelligent Component in our framework would enable patients to be, to some extent, self-sufficient in their own houses by increasing their safety.

Section 2 of this paper illustrates the features of a Location System based on sensors networks under the ZigBee protocol. Section 3 describes the CyberCare Intelligent Component system and Section 4 presents conclusions and further hints for development.

2 The Location System

The Location system used in CyberCare is based on concurrent use of ZigBee networks and Data from Inertial Measurement Units.

ZigBee is a wireless technology developed as an open global standard to address the unique needs of low-cost, low-power, wireless sensors networks. The standard takes full advantage of the 802.15.4 physical radio specification developed at the Institute of Electrical and Electronics Engineers (IEEE). The specification is a packet-based radio protocol that meets the needs of low-cost, battery-operated devices. The protocol allows devices to intercommunicate and be powered by batteries that last years instead of hours.

[1] Dipartimento di Informatica e Comunicazione, Università degli studi di Milano, email: mileo@dico.unimi.it

[2] Dipartimento di Informatica, Sistemistica e Comunicazione, Università Statale di Milano-Bicocca, email: roberto.bisiani@disco.unimib.it, davide.merico@disco.unimib.it

Among the ZigBee protocol features, we mention:

- Low duty cycle - Provides long battery life
- Low latency
- Support for multiple topologies: Static, dynamic, star and mesh
- Direct Sequence Spread Spectrum (DSSS)
- Up to 65,000 nodes on a network
- 128-bit AES encryption Provides secure connections between devices
- Collision avoidance
- Link quality indication
- Clear channel assessment
- Retries and acknowledgments
- Support for guaranteed time slots and packet freshness

ZigBee technology was not originally conceived to be used as location system. However, some commercial products have been developed to enable ZigBee nodes localization [13].

Inertial Measurement Units and components, which sense either acceleration or angular rate, are being embedded into common user interface devices more frequently as their cost continues to drop dramatically. These devices hold a number of advantages over other sensing technologies in that they measure relevant parameters for human interfaces and can easily be embedded into wireless, mobile platforms.

The Wearable Acquisition Device (WAD), developed by Microsystems [10], includes a three-dimensional accelerometer and a three-dimensional inclinometer that are used by the system to determine the position of the patient and his behavior to perceive when an emergency arises and how do patients react to the situation.

3 The Intelligent Component at a glance

The Intelligent Component of CyberCare is situated on a personal computer we call the Home Processor (Figure 1).

The reasoning process is based on (temporally) local details about the patient, collected real-time whenever an emergency is detected.

In Artificial Intelligence, the possibility of making assumptions rather than just doing deductions from a given knowledge base, has been considered very attractive and widely used for declarative representations of problems in a variety of areas. One way to use the assumption-based framework is that of applying Default Reasoning, especially in areas where you don't want to enumerate all of the exceptions of a situation, even if you could think of them all.

The intuition is that of using automated commonsense (nonmonotonic) reasoning to analyze patient's clinical[3] and environmental settings in order to detect the origins of an emergency and its resolution by reasoning on exceptions.

Our knowledge base does not contain the whole medical knowledge that may be related to the specific case. This would be too huge to manage when the inference process is running, while emergencies require rapid answers.

According to results provided by the inference engine the system can then update patient's clinical profile and habits definition by modifying and adapting numerical thresholds in logic predicates.

This means that instead of reasoning on similar medical cases only, the Intelligent Component is in charge of extracting and reasoning about the current situation of the patient at the time the emergency arose. The main tasks performed by the Intelligent Component are:

- monitoring patients through a sensors network and a WAD device;
- collecting both static and dynamic data about the patient (habits, biomedical data, location) and the environment the patient moves in (rooms, areas of interest);
- extracting such data as soon as the need of assistance is detected, an convert them into logic predicates;
- determining solutions to emergencies through an inference engine (ASP solver) evaluating the ASP program composed by the extracted predicates and logic rules[4];
- convert ASP solutions into actions through external modules;
- periodically (at the end of each day) adapting patient's profile automatically (off-line learning);

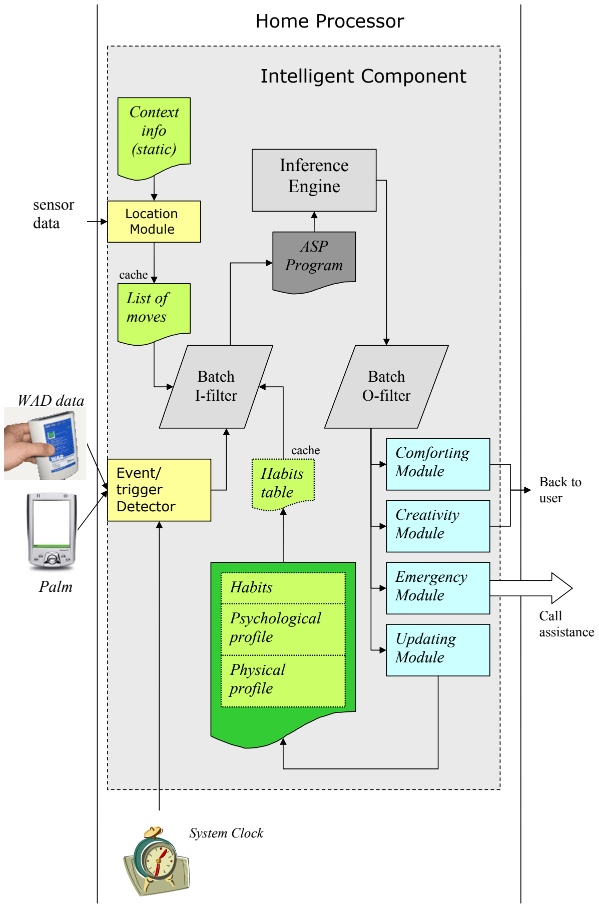


Figure 1. CyberCare: general Architecture

3.1 Event Detector

The Event Detector component is in charge of capturing external events and triggering the inference process in case of emergency. We consider three possible triggering events:

1. **significant changes of biomedical parameters** including blood pressure, ECG, temperature and others;

[3] Clinical details are formalized by the medical assistant during setup, while contextual information is extracted automatically by the Location Module and the Habits table.

[4] Logic rules enabling the system to reason about the emergencies and detect eventual anomalies in patient's behavior, are related to generic social settings and need to be defined by the knowledge engineer together with a medical expert.

2. **unexpected changes of patient's habits**, i.e., an action has not been performed within the time bound scheduled for it;
3. **explicit requests of help**.

It is worth mentioning that, in our view, clinical exceptions have higher priority than the behavioral ones.

This means that whenever a physical problem is detected together with a behavioral anomaly, only the former one is considered in the inference process. This allows the reasoning process to be selective, thus faster and much more effective in treating emergencies.

3.2 Location Module

The system has to collect information about patient's moves in the house in order to know what the context of an emergency is like. The Location Module is in charge of this task through Location Algorithms based on RSSI values (Received Signal Strength Indicator).

Location Engines use RSSI values combined with physical locations of reference nodes (with static location) to calculate positions of "blind nodes".

Reference nodes must be configured with an X and a Y value that correspond to the physical location. The main task for a reference node is to provide a "reference" packet that contains X and Y coordinates to the blind node, also referred to as *anchor node*.

A blind node communicates with its nearest reference nodes, collecting X, Y and RSSI for each of these nodes and calculates its position based on this parameter input using the location engine hardware.

The location estimation is performed at each node, hence algorithm is decentralized. This feature will reduce the amount of data transferred to radio, since only the calculated position is transferred.

All location information provided by ZigBee Location Engine are integrated with motion data computed by WAD inertial sensors to enhance position calculation precision.

3.3 Static and Dynamic Profile Extraction

General information about the patient is collected into a static profile to be later enriched run-time by dynamic data.

The static part of patient's profile includes:

- **clinical data**, represented as a set of logic predicates of the form
  $normal\_value(Parameter, Min\_value, Max\_value)$.
  and included in the ASP logic program when the system inference is triggered;
- **patient's habits**, expressed as
  $time(Action, Init, S1, Duration, S2)$.
  $place(Action, Room)$.
  indicating when and where an action is supposed to be performed; $S1$ and $S2$ represents the leeways in beginning and duration of the action, respectively;
- **patient's psychopathologies**, interaction-oriented details that have to be previously formalized by an expert and included in the knowledge base by the knowledge engineer[5].

Dynamic information is responsible for the system inference to be triggered. The dynamic component of patient's profile includes:

- **biomedical parameters** registered by the WAD and transmitted to the Intelligent Component by the Batch I-filter: $actual(Parameter, Min, Max, Time)$.
- **list of moves**, extracted by the Batch I-filter:
  $enter(Room, Time, N)$.
  $exit(Room, Time, N)$.
- **patient's requests**, transmitted through the palm device.

3.4 The Inference Process

To make the run-time reasoning task more efficient, we do not apply Machine Learning Techniques and case-based reasoning in the Inference Process, but rather Default Reasoning under the Answer Set Programming (ASP) paradigm.

ASP is based on the *stable model* semantics for Logic Programs proposed by Gelfond and Lifschitz [8] and it can be seen as bringing together concepts and results from Logic Programming, Default Reasoning and Deductive Databases.

In Default Reasoning you specify general knowledge for standard cases (the defaults) and modularly add exceptions. When you add an exception to default, you can't conclude what you could before. In that Default Reasoning is told to be *nonmonotonic*.

In the first prototype of CyberCare we evaluate the ASP program by using the Smodels solver, an implementation of the stable model semantics for logic programs implemented by Patrik Simons [12].

Smodels treats variable-free programs and it has to be used together with Lparse, a front-end in charge of performing the grounding procedure to produce a variable-free logic program for Smodels.

One may argue that Answer Set Programs grounding could be very costly; in this setting, the interaction-oriented approach allows us to restrict grounding to a finite and quite reduced domain including few biomedical data, daily activities and rooms of the house. As for the time unit, at the end of any inference we mark successfully completed actions and in the subsequent inference we limit time units from the current (discrete) instant of time back to the bound of the last unchecked activity.

The Smodels solver supports constraints, choice rules and weight rules [11, 14] and can be thus considered powerful enough to give interesting solutions to complex reasoning tasks.

To represent Default Reasoning in Smodels we have to express that an atom or predicate is an *assumable*[6] by telling that it is to be considered *true* unless some other rules indicates its negated[7] holds:

$assumable$:$-$*not* $exception$.

An activity A is considered $normal$ by default at time T, unless any anomaly is detected:

$normal(A, T)$:$-$*not* $anomaly(A, T)$.

The following *exception rule* indicates that an anomaly on action A holds by default at time T when A has not been performed within the given time bounds[8]:

$$\begin{array}{lll}
anomaly(A,T) & :- & time(A, Init, S1, L, S2), \\
 & & not\ done(A, R, T), place(A, R), \\
 & & Init + S1 + L + S2 < T. \\
done(A, R, T) & :- & was(R, T1, T2), place(A, R), \\
 & & time(A, Init, S1, L, S2), \\
 & & T1 < T2 < T, \\
 & & Init - S1 < T1 < Init + S1, \\
 & & L - S2 < T2 - T1 < L + S2. \\
was(R, T1, T2) & :- & enter(R, T1, N1), T1 < T2, \\
 & & exit(R, T2, N2), N1 = N2 - 1.
\end{array}$$

[5] Not included in the actual prototype.

[6] An *assumable* is a ground instance of a possible hypotheses that can be considered true when consistent.

[7] We consider Negation as Failure [7].

[8] This is only the simplest case. Further possible explanations for a behavioral anomaly can be treated by introducing corresponding exception rules.

The above rules refer to actions that are supposed to be completed at the time the inference is running. We also want the system to monitor actions that are being executed. These actions are checked to keep track of eventual delays in the initial time scheduled for them. Interruptions due to physical problems are also considered[9]. In exceptional cases, patients can be asked for indications related to their change of habits, but we don't consider this case in the first prototype of the system, cause we want it to be as less intrusive as possible.

3.5 The Output Modules

Inference results are interpreted by the batch output filter (Batch O-filter) that captures logic predicates included in the solutions (stable models) provided by Smodels and call the appropriate module, external to the inference engine.

The initial prototype of the system will include:

- the Updating Module analyzing what happened in the last twenty-four hours and updating[10] patient's profile accordingly;
- the Emergency Module redirecting the treatment of physical emergencies to the appropriated service.

Thanks to the modularity of the system, this list can be extended. As an example, we could provide a Creativity Module interacting with the patient when he asks for company, or a Support Module to provide psychological support through interactive screens.

4 Conclusions and Future Work

We propose a framework for home healthcare based on sensors network technologies, patient's profile observation and environment analysis through logic programming.

Patient's profile management and clinical assistance are oriented to the context of a specific emergency rather than to the general case-based analysis. This shift of perspective limits the amount of knowledge to be considered at a time, and that's the reason why we proposed not to use canonical Machine Learning Techniques, that need to treat a large amount of data in order to learn how to get to a solution/treatment.

The patient- and interaction-oriented approach based on Default Reasoning results in being unobtrusive, modular, declarative [1], efficient and self-adapting.

Non intrusiveness is granted by the fact that information about the user is extracted automatically by the location system and the WAD device: no complex statistical information or general medical knowledge are needed to determine the nature of the emergency.

Modularity is given by the Default Reasoning and a declarative specification of the problem, while efficiency and self-adaptation are granted by the fact that we use ASP inference engines and off-line profile update instead of case-based analysis and Machine Learning.

The interaction is fully intuitive, as we deal with multimedia contents and the patient is provided with a palm device that works like a remote control TV switch to interact with the system.

Preliminary tests on a few profile instances showed that CyberCare could be profitably employed in home healthcare services supporting the delivery of care. Our initial studies have been mainly addressed to the elderly, but the patient's profile specification and analysis we propose allows us to easily extend this framework to deal with other categories of subjects having social disadaptations (e.g. hyperactive children), and it could thus represent a desirable tool to support the National Social Service.

Nonetheless, we are aware of the fact that more detailed and huge experimental results are needed to evaluate effectiveness of this approach in different social contexts, and provide significant empirical data. This aspects will be detailed in a future extended paper, where an advanced prototype will also be presented. Such a complete prototype is supposed to be employed in a restricted area of the city of Milan (Italy).

A further issue is related to the solutions obtained by the system. One of the interesting aspects of using ASP semantics in this context is that all possible solutions to an emergency (interventions, diagnosis, modules' activation, etc.) are considered equally valid. There are efficient techniques to enforce priorities and ordering relations among solutions of an ASP program [2, 3], and it would be interesting to investigate how to apply these techniques in the Healthcare scenario.

In a future paper we want to investigate further extensions of the framework, such as i) ordering rules and predicates to select the preferred solution among the possible ones and ii) automatically updating rules rather than just thresholds, to efficiently reason about new unexpected situations [5].

References

[1] C. Baral, *Knowledge Representation, Reasoning and Declarative Problem Solving*, Cambridge University Press, 2003.

[2] G. Brewka, 'Logic programming with ordered disjunction', *Proc. of AAAI02. Extended version presented at NMR02*, (2002).

[3] G. Brewka, I. Niemelä, and T. Syrjänen, 'Implementing ordered disjunction using answer set solvers for normal programs', *Proc. of JELIA02*, 444–455, (2002).

[4] A. Cesta and F. Pecora, 'The robocare project: Multi-agent systems for the care of the elderly', *ERCIM News No. 53*, (2003).

[5] J. Del Grande, T. Schaub, and H. Tompits, 'A preference-based framework for updating logic programs: Preliminary report', *Workshop on Preferences and Their Application in Logic Programming Systems*, (2006).

[6] E. Dishman, 'Inventing wellness systems for aging in place', *Computer, IEEE*, 34–41, (2004).

[7] P.M. Dung and P. Mancarella, 'Production systems with negation as failure', *IEEE Transactions on Knowledge and Data Engineering*, **14**(2), 336–352, (2002).

[8] M. Gelfond and V. Lifschitz, 'The stable model semantics for logic programming', *ICLP/SLP*, 1070–1080, (1988).

[9] M.V. Giuliani, M. Scopelliti, and F. Fornara, 'Elderly people at home: Technological help in everyday activities', *Int's Workshop on Robots and Human Interactive Communication*, (2005).

[10] MsWebCare, 'Web based solutions for healthcare', *http://www.mswebcare.it*, (2001).

[11] I. Niemelä and P. Simons, 'Extending the smodels system with cardinality and weight constraints', *Jack Minker, editor, Logic-Based Artifcial Intelligence, Kluwer Academic Publishers*, 491–521, (2000).

[12] Helsinky University of Technology, 'Smodels solver', *http://www.tcs.hut.fi/Software/smodels/*, (2002).

[13] N. Patwari, A. O. Hero III, M. Perkins, N. S. Correal, and R. J. ODea, 'Relative location estimation in wireless sensor networks', *IEEE Trans. Signal Process. 51, 8 (August 2003)*, 21372148, (2003).

[14] P. Simons, I. Niemelä, and T. Soininen, 'Extending and implementing the stable model semantics', *aij*, **138**(1-2), 181–234, (2002).

[15] Y.A.W. Slangen-de Kort, C.J.H. Midden, and A.F. van Wagenberg, 'Predictors of the adaptive problem-solving of older persons in their homes', *Env. Psych*, **18**, 187–197, (1998).

[16] K. Stathis and F. Toni, 'Ambient intelligence using kgp agents', *2nd European Symposium for Ambient Intelligence*, 351, (2004).

[17] J. Teich, J. Schmiz, G. Kuperman, and D. Bates, 'Effects of computerized physician order entry on prescribing practices', *Archives of Internal Medicine*, **160**, 2741–47, (2000).

[9] For lack of space we omit the related code.

[10] Updates are made in terms of changing S1, L, S2 values according to some heuristics.

User Profile Agents for Cultural Heritage fruition

Stefania Costantini [1] and **Paola Inverardi** [2] and **Leonardo Mostarda** [3] and **Arianna Tocchio** [4] and **Panagiota Tsintza** [5]

Abstract. In this paper we present an application of a MAS (Multi-Agent System) composed of logical agents in an Ambient Intelligence scenario, related to the fruition of cultural assets. The users are located in an area which is known to the agents: in the application, the users are the visitors of Villa Adriana, an archaeological site in Tivoli, near Rome (Italy). Agents are aware of user moves by means of Galileo satellite signal, i.e., the proposed application is based on a blend of different technologies. The agents, developed in the DALI logic programming language, proactively learn and/or enhance users profiles and are thus capable to competently assist the users during their visit, to elicit habits and preferences and to propose cultural assets to the users according to the learned profile.

1 INTRODUCTION

The paradigm of Ambient Intelligence implies the objective of building a friendly environment where all of us will be surrounded by “intelligent” electronic devices, and this ambient should be sensitive and responsive to our needs. A multitude of sensors and actuators are already embedded in very-small or very large information and communication technologies, and a challenging task nowadays is to identify which advantages can be gained from these technology systems. Tourism for instance is a context where old and new aspects can be melted for reaching interesting results. In fact, tourism is a growing industry and it needs to evolve according to the tourists changing features. In the past, tourists were satisfied with standardized package tours. Today, with the popularization of traveling, tourists are expecting new tour experiences that are different and authentic [11].

Most cultural tour sites today still maintain a conventional form of tour that is static and provides a visitor with plenty of information, which is however lacking any form of customization. Several interesting works have proposed a new manner of enjoying cultural places, as technology may support more dynamic and personalized methods to conceive the fruition of cultural assets. Park et al. in [9] propose a system named “Immersive tour post” that uses audio and video technology to provide improved tour experiences at cultural tour sites. This system take the form of a post that stays fixed in one location and reproduces the vision and sounds of the historical event that occurred at the particular space. Mobile applications in a mobile-environment have been experimented by Pilato et al. in [10].

Visitors are assisted in their route within the “Parco Archeologico della Valle dei Templi” (archaeological area with ancient Greek temples) in Agrigento (Sicily, Italy) by an user-friendly virtual-guide system called MAGA, adaptable to the users needs of mobility. MAGA exploits speech recognition technologies and location detection, thus allowing a natural interaction with the user. Each visitor can access the information on cultural assets via a portable device (PDA, or “Personal Digital Assistant”) that, through RFIDs (which are a well-known standard for an automatic identification method), is able to capture where the person is in the Parco.

Several other proposals can be found in the literature, exploring the integration between human-computer interaction and information presentation. The system Minerva, proposed by Amigoni et al. in [1] organizes virtual museums, starting from the collections of objects and the environments in which they must be displayed while the DramaTour methodology presented by Damiano et al. in [8] explores a visit scenario in an historical location of Turin. Visitors are assisted by a virtual spider that monitors their behavior and reactively proposes the history of the palace in detail and a lot of funny anecdotes about the people. Studying the human behavior during the visit in a cultural heritage scenario is an exciting aim.

The systems presented above have a common characteristic: they try to improve the traditional methods to inform the visitor by means of new catchy techniques for making the human-machine interface more friendly and intuitive. But, is it possible to go beyond, towards capturing the visitors desires and expectations? A particular mechanism for capturing the visitor interest for one or more cultural assets has been presented by Bhusate at al. in [2]. Each visitor receives a PDA associated to non-invasive sensors that measure “affective” context data such as the user’s skin conductance and temperature. The sensor readings are reported to a control module that determines, according to other data, the visitor’s mood. Preferences can be also catched by asking questions directly to the user before starting the visit.

This method has been adopted in the system KORE [3] where parameters such as age, cultural level, preferences in arts, preferred historical period, etc., are taken into account for “tuning” the pieces of information provided, by omitting those useless for the user (either too difficult or too easy to understand) and delivering only data which match the user profile. The architecture of KORE is based on a distributed system composed of some servers, installed in the various areas of museums, which host specialized agents. The KORE system practically demonstrates that intelligent agents can have a relevant role in capturing the user profile by observing the visitor behavior. They possess the capability to be autonomous and to remain active while the visitor completes her/his visit; they can perceive through the sensors all choices performed by the user and, consequently, activate a reasoning process.

In this paper, we present the architecture of the MAS DALICA applied to the Villa Adriana scenario for capturing the visitors interests and enhancing their profiles. Similarly to what happens in the KORE

[1] University of L’Aquila, Italy, email: stefcost@di.univaq.it
[2] University of L’Aquila, Italy, email: inverardi@di.univaq.it
[3] University of L’Aquila, Italy, email: mostarda@di.univaq.it
[4] University of L’Aquila, Italy, email: tocchio@di.univaq.it
[5] University of L’Aquila, Italy, email: panagiota.tsintza@di.univaq.it

system, each DALICA intelligent agent starts its activity with the caching of data such as the visitors' age, preferences, cultural level and so on. Then, it captures additional data about the visitor's movements and choices, elaborates them and updates the user profile. The visitor's movements are traced by means of the Galileo satellite. The learned profile allows DALICA to offer information on the cultural assets adapted to the visitor, and to proactively propose to see those assets closer on the one hand to the visitor's physical position and one the other hand to the the visitor's preferences. The related items of information are provided in an appropriate customized form. As acknowledged in Section 4, the DALICA system has been developed within the CUSPIS European project.

In Section 2 we present the scenario where DALICA has been put at work and the features of the system. We also discuss the methods through which the intelligent agents are capable to capture the visitors' interest and the monitoring capabilities of the agents. Finally, we conclude in Section 3.

2 THE DALICA SYSTEM

The case-study on which the development of DALICA has been based is that of constructing and updating the user profile of visitors of Villa Adriana in Tivoli near Rome (Italy). Villa Adriana is an exceptional complex of classical buildings created in the 2nd century A.D. by the Roman emperor Hadrian. It combines the best elements of the architectural heritage of Egypt, Greece and Rome in the form of an 'ideal city' [12].

For a visitor, Villa Adriana is a unique wonderful place. For DALICA, Villa Adriana as a set of Points of Interest (POI's). For "POI" we intend either a specific cultural asset like the "Pecile" or public places like restaurants located nearby. The first part of our work has been concerned with the study of the scenario of Villa Adriana for individuating the characteristics of the POIs useful for the agent reasoning process. For this purpose we have defined a POI as a set of the following fields:

- *Identifier*: a string identifying uniquelly the POI;
- *Latitude*: the latitude of the POI defined through the Galileo satellite.
- *Longitude*: the longitude of the POI defined through the Galileo satellite.
- *Radius*: the radius of the circle that contains the POI area.
- *Keywords*: a list of the POI characteristics like, for example, 'mosaic' if the POI contains a mosaic, or 'water' if in the POI there is a fountain or a water basin. Considering that each POI can have one or more keywords, we combined each one with a number indicating its weight (relative importance) in the POI description. For example, assuming that the "Pecile" usually captures the attention of a visitor prevalently for the water basin while the mosaic maintains a very marginal role, the list of the keywords will be $[(water, 60), (garden, 30), (mosaic, 10)]$. Clearly, this information has been provided by experts.
- *Time for visit*: is an average of the time that we suppose an user will employ for visiting the specific POI.

The POIs descriptions have been collected into an appropriate dynamic ontology (developed by the group of Artificial Intelligence and Natural Language Processing at the Dept. for Computer Science, Systems and Management of the University of Rome Tor Vergata, in the context of the CUSPIS project).

For example, for defining the "Pecile", we use the following string: $poi('VA\_PecileV1', 41.94201257700091, 12.77403535070269, 80, [('mosaic', 10), ('water', 40), ('statue', 20), ('garden', 10), ('column', 20)], 10)$. Keywords are important because they allow to establish the possible similarities between POIs and, consequently, to discover if the visitor is interested in some particular feature which is common to them. E.g., if in Villa Adriana a visitor decides to visit the "Pecile", the "Teatro Marittimo", the "Canopo", the "Piccole" and the "Grandi Terme", it is plausible to assume that she/he could be interested in those POIs where the water has a relevant role.

In this scenario, we have developed and experimented DALICA MAS. The main goal of the system has been that of supporting users during their visits. This implies capturing their profiles and offering them a customized information on the cultural assets, including proposals for extending the visit, for new visits or for other visits to related places, better if located nearby. Each visitor, at the beginning of the visit, has to book the route on an Internet site where she/he can express some preferences and choices about the service fruition. Then, each visitor is provided with a PDA by which the movements and the choices of the visitor can be observed, so that she/he can receive suitable information on the cultural assets.

When the visitor starts her/his route, an intelligent agent, called User Profile One, is generated. At the staring phase, it elaborates the data coming from the user-profile stored on Internet and determines an initial fruition profile. Then, it re-elaborates the fruition profile according to the new data derived from the user behavior. New enhanced fruition profiles will possibly substitute the former one while the visitor proceeds in the route.

At this point, it is necessary to explain through which strategies it is possible to capture the visitors interests in a scenario such as Villa Adriana, where the cultural assets are arranged in an area of 300 hectares.

2.1 Deducing the Visitor's Interests

Intelligent agents in DALICA are reactive, pro-active and communicative. They are capable of perceiving the data coming from the environment such as the satellite coordinates or the POIs chosen by the visitor and to react appropriately. While reactivity allows the agents to adopt a specific behavior in response to the external perception, pro-activity has a main role, because the reasoning process that leads to the interests deduction is based on the correlation of several data coming from the environment, from the ontology and from some basic inferential processes.

Communication capabilities intervene whenever it is necessary to send data to the visitor's PDA: e.g., the explanations of what is being seen or the list of the deduced interests or the proposed other POIs to see or the warning that the visitor is entering in a restricted area. In the rest of this section we concentrate the attention on the methods used for deducing the user interests, while in next section we present the strategies for assisting her/him during the visit and for checking her/his behavior.

We divide the agent deduction process into three phases: the first one represents a basic deduction level while the second and third ones elaborate the results by concatenating the previous deductions. We starts the explanation by illustrating the algorithms concerning the first phase:
Deducing the interests based on time: This algorithm is founded on the consideration that a visitor is interested in a POI if she/he observes it for a time interval "longer" than the average time of the visit for the specific cultural asset. The meaning of "longer" can be modulated according to the current visitor's profile. So, if a visitor has booked a visit that lasts up to six hours the time interval for the

observation will be longer than that of a visitor that booked a visit lasting for two hours.

How is it possible to determine which POI the visitor is looking at? The method is based on the Galileo Satellite. Each POI, as explained in the previous section, is identified by a circle (whose center is defined by a latitude and a longitude) and by a radius. If the visitor position (expressed in latitude and longitude and coming from the PDA) belongs to the circle related to a specific POI, we can suppose that she/he is visiting that POI. If two or more POIs are close enough to determine an intersection between their circles and the visitor is located in the intersection, then the algorithm, not being able to capture the real intention of the visitor, presumes that the visitor is interested in all those POIs. Each POI which is selected according to the visitor movements is identified by a list of keywords. The algorithm elaborates the keywords of all selected POIs and then extrapolates the most frequent ones. These keywords represent the hypothetical user interests that, once deduced, will have to be confirmed both by subsequent user behavior and by other deduction mechanisms.

Deducing the interests based on the visited POIs: This algorithm considers the POIs chosen by the user and its outcome improves when several POIs have already been visited. In fact, for each POI the algorithm extracts the keywords and the most frequent ones are asserted as "deduced interest".

Deducing the interests based on the chosen route: If a visitor decides to follow a predefined route chosen among those proposed by the system, the agent tries to capture the visitor's interests by studying the POIs included in the route. Each POI will have a list of keywords and those most relevant for describing the route will be selected for the next step of the deduction process.

Deducing the interests by similarity: This algorithm employs a similarity measure. In particular, the interests expressed by the visitor in the web site are matched with those in the ontology. Those in the ontology which look to be similar enough are selected as deduced interests.

Deducing the interests according to some questions: Another strategy for capturing the visitor's interests is centered on some occasional questions about the POIs located near the visitor. The agent observes the POIs around the PDA, chooses one of them and asks the visitor's opinion on it. A positive response such as ("Yes, I like the Odeon") will trigger the interests deduction process.

Deducing the interests according to cultural questions: The last strategy for deducing the visitor's interests takes into consideration the cultural level of the visitor. Some questions such as "Do you like the ancient art? Do you know what is a cavea?" are useful to determine the information level to submit to the visitor. Moreover, some parameters such as the visitor's job and age are involved in the process. The agent compares the data acquired via the questions and via the other parameters and elaborates them in order to determine the appropriate degree of the information. We have identified for now three degrees.

-**Basic**: It is related to a basic information level where the user prefers a superficial information on the POIs combined with details on the ancient people's life. This level usually fits primary and secondary students and occasional visitors.

-**Medium**: Provides more technical data on POIs and particular attention is reserved to their structure. This level fits people fond of art.

-**Specialized**: Provides the visitor with a detailed information on POIs combined with information about the materials and techniques used to manage the cultural assets. This level is tailored to specialized students, technical people, researchers and so on.

The second deduction phase captures the results of the previous deduction algorithms and tries to compare them, with the aim of reaching a more precise user profile definition. In particular, those interests coming from the previous phase and confirmed by this second one are involved in a process that selects only the most frequent ones. These interests are sent to the visitor's PDA in order to be confirmed by her/him. Precisely, this second phase is based on the following algorithms:

Filtering the deduced interests according to the time: This filter combines the deduction of the interests based on the permanence near a certain POI and the moment when the deduction itself has been reached. In particular, this step has the objective of understanding whether a visitor remained in a specific area because interested in a POI or for some other reason (e.g., she/he was sitting on a lawn eating a sandwich). The reasoning process is presently pretty simple, and will be improved in the future. We suppose that a visitor could be interested in eating especially at a certain time (say from 12:30 to 14:30). If the visitor has not spent some time in a restaurant area and the deduction has been reached after 12:30 and before the 14:30, then the hypothesis of eating a sandwich has to be taken into account with a higher priority than at other moments of the day.

Each deduced interest is involved in a *interests updating process*. More precisely, each interest/keyword is associated to a weight (priority) N. For a specific deduced interest K, we have defined a global evaluation function computed on the weights. In this manner, the system takes in account not only the interests more frequently deduced but also their 'relevance' in the deduction process.

Combining the deduced interests: The interests deduced by the previous algorithms based on time, on visited POIs, on the chosen route and according to some questions are crossed in order to obtain a more reliable user profile definition. The interests which are confirmed will be involved in the *interests updating process*.

Using similarity for confirming the deduced interests: Reliability of the interests deduced in the previous phase is checked according to the similarity degree with those inserted in the visitor's profile in the web site. If the similarity is greater than a prefixed threshold, the interest will be involved in the *interests updating process*.

The third phase delivers data related to the elicited interests to the visitor's PDA. When the visitor receives the interests list, she/he can confirm either all interests or a subset of them. The selected interests are managed by the agent for updating the user profile. Moreover, the agent communicates them to a central system that manages the information for the visitor in order to propose (through the agent) data and POIs closer to his desires and expectations.

2.2 Monitoring Visitor's Behavior

Intelligent agents in DALICA are also used for monitoring the users behavior with a fixed frequency. The situations where the reactive and proactive capabilities of the agents are put at work for this kid of monitoring are at least the following.

Checkinging the forbidden areas: In Villa Adriana there are areas where visitors cannot enter. These areas are defined in the ontology and an agent monitors from time to time the visitors' movements in order to guarantee that no one violates the rules.

Monitoring the visitors route: The agent has the ability to follow the visitor that has chosen a predefined route along his visit. For instance, the agent is able to make the itinerary shorter or longer (by either removing or adding POIs) according to the user pace, so that the user can complete the itinerary in time.

Creating a list of POIs: When the visitor has finished the visit, the agent collects all POIs that he has visited and puts them in a file with texts and images. This allows the visitor to keep a reminder of his visit to Villa Adriana.

2.3 The DALICA Architecture

The DALICA architecture involves a MAS and a central external system. This system on the one hand acts as a "router" between the MAS and the PDA's: in fact, the MAS is presently too heavy to be directly installed on the PDA's. Thus, the MAS resides on a more powerful machine and uses the central system to exchange data with the PDA's. It receives messages from/to the agents and delivers them from/to to the PDAs of the visitors. On the other hand, the central system collects and stores data about visitors and visits for future use.

In the DALICA MAS, several intelligent agents cooperate in order to support the users during their visit. The three most important agents composing the MAS are the following.

Generator Agent: The role of this agent is to automatically generate the User Profile agents when a user starts a visit. The generation process happens when PDA sends a positioning message related to a new visitor. This reactive capability is combined with a set of pro-active functions that check from time to time if the User Profile agents are active and, if not, generate them again.

User Profile Agent: Acts as described before in this section. They deduce the visitors interests and monitor their behaviors.

Output Agent: Manages communications between the DALICA MAS and an external central system.

The MAS receives data about the user movements and actions coming from the visitors PDAs via the Input Interface. This interface is not an agent. It has the role to deliver messages to and from the external system into the Linda Tuple Space through which the intelligent agents in the DALICA MAS communicate.

DALICA agents have been implemented in the DALI language [4] [5] [13] [6] [7], an Active Logic Programming language designed for executable specification of logical agents. Reactivity allows DALI agents to perform a number of tasks, including the interaction with the Galileo satellites and the perception of external stimuli coming from the user, the external system or the other agents. Most of the reasoning processes outlined above take place proactively, by exploiting the DALI mechanism of *internal events*. Conditions that may trigger new reasoning are checked from time to time (at a frequency and with priorities stated by separate directives). If such a condition is true, the reasoning starts, and appropriate actions are undertaken. Reasoning processes take also profit from the ability of DALI agents to store past events and actions together with the time when they have occurred. They are thus able to reason on the past and thus learn from experience.

3 CONCLUSIONS

We conclude this paper by making some considerations about our work. It is not so easy to find an application where intelligent agents are put at work in a real scenario but it is even less frequent to find intelligent logical agents at work. In the light of these considerations, the DALICA MAS is a novelty. This also because DALICA exploits the signal of Galileo Satellites to deduce the Users Profiles. DALICA at work in the area of Villa Adriana practically demonstrated that logical agents can be applied successfully for capturing the visitors habits and preferences.

Our system cannot be compared with platforms such as MAGA and DramaTour where the main goal is to offer information to the visitors via specialized interfaces. DALICA mainly deduces the visitors interests and leaves the job of presenting the information to an external component. KORE is the system closer to DALICA because it uses agents for managing the information through the study of the User Profile. KORE does not use the Galileo signal and its agents are not logical. Moreover, DALICA is more centered on the deduction profile process while KORE mainly filters the information according to the User Profile characteristics.

As future developments, the system reasoning capabilities that are presently quite basic can be improved. Also, previous experience can be better exploited. Different agents managing different visitors might communicate so as to cooperate in improving their performance and enhancing the services they offer.

4 ACKNOWLEDGEMENTS

This work has been partially supported by the project CUSPIS (GJU/05/2412/CTR/CUSPIS) "A Cultural Heritage Space Identification System".

REFERENCES

[1] F. Amigoni, S. Della Torre, and V. Schiaffonati, 'Yet another version of minerva: The isola comacina virtual museum', in *Proc. of the First European Workshop on Intelligent Technologies for Cultural Heritage Exploitation, at The 17th European Conference on Artificial Intelligence*, pp. 1–5, (2006).

[2] A. Bhusate, L. Kamara, and J.V. Pitt, 'Enhancing the quality of experience in cultural heritage settings', in *Proc. of the First European Workshop on Intelligent Technologies for Cultural Heritage Exploitation, at The 17th European Conference on Artificial Intelligence*, pp. 1–13, (2006).

[3] M. Bombara, D. Cal, and C. Santoro, 'Kore: A multi-agent system to assist museum visitors', in *Proc. of the Workshop on Objects and Agents (WOA2003), http://citeseer.ist.psu.edu/708002.html*, (2003).

[4] S. Costantini and A. Tocchio, 'A logic programming language for multi-agent systems', in *Logics in Artificial Intelligence, Proc. of the 8th Europ. Conf.,JELIA 2002*, LNAI 2424. Springer-Verlag, Berlin, (2002).

[5] S. Costantini and A. Tocchio, 'The dali logic programming agent-oriented language', in *Logics in Artificial Intelligence, Proc. of the 9th European Conference, Jelia 2004*, LNAI 3229. Springer-Verlag, Berlin, (2004).

[6] S. Costantini and A. Tocchio, 'About declarative semantics of logic-based agent languages', in *Declarative Agent Languages and Technologies*, eds., M. Baldoni and P. Torroni, LNAI 3229, Springer-Verlag, Berlin, (2006). Post-Proc. of DALT 2005.

[7] S. Costantini, A. Tocchio, and A. Verticchio, 'A game-theoretic operational semantics for the dali communication architecture', in *Proc. of WOA04, Turin, Italy, December 2004, ISBN 88-371-1533-4*, (2004).

[8] R. Damiano, C. Galia, and V. Lombardo, 'Virtual tours across different media in dramatour project', in *Proc. of the First European Workshop on Intelligent Technologies for Cultural Heritage Exploitation, at The 17th European Conference on Artificial Intelligence*, (2006).

[9] D. Park, T. Nam, C. Shi, G.H. Golub, and C.F. Van Loan, *Designing an immersive tour experience system for cultural tour sites*, ACM Press, New York, NY, 1193-1198, Montral, Qubec, Canada, April 22 - 27, in chi '06 extended abstracts on human factors in computing systems edn., 2006.

[10] G. Pilato, A. Augello, A. Santangelo, A. Gentile, and S. Gaglio, 'An intelligent multimodal site-guide for the parco archeologico della valle dei templi in agrigento', in *Proc. of First European Workshop on Intelligent Technologies for Cultural Heritage Exploitation, at The 17th European Conference on Artificial Intelligence*, (2006).

[11] A. Poon, 'The new tourism revolution', *Tourism Management,vol.15, no.2*, (1994).

[12] UNESCO site. Villa adriana. http://www.villa-adriana.net.

[13] A. Tocchio. Multi-agent sistems in computational logic. Ph.D. Thesis, Dipartimento di Informatica, Universitá degli Studi di L'Aquila, 2005.

Curious Places: Curious, Proactive, Adaptive Built Environments

Kathryn Merrick[1] and **Rob Saunders**[2] and **Mary Lou Maher**[2]

Abstract. Advances in intelligent agent research, such as curious agents and motivated learning agents, make possible a new kind of intelligent environment: a curious place. Previously, intelligent environment research has focused on developing reactive and interactive systems that control sensor and effector architectures, achieve context awareness and support human activities. This paper identifies the key attributes of curious places that differentiate them from existing intelligent environments and proposes new focus areas for intelligent environment research: proactive problem finding, life-long adaptability and enhancement of human activities. An example of a curious place application is discussed with emphasis on its adaptability and its potential to enhance human experiences.

1 INTRODUCTION

Recent intelligent agent research developing intrinsically motivated learning agents presents opportunities for the design of places able to respond with motives such as interest and curiosity to support and enhance human activities. Maher et al [7] introduced three motivated learning agent models for intrinsically motivated intelligent sensed environments that incorporate computational models of motivation with reinforcement learning, supervised learning and unsupervised learning. These models aim to achieve adaptive responses using motivation to direct learning towards useful or interesting behaviour.

When computational models of curiosity are used as the model of motivation in intelligent environments, a new kind of space emerges: a curious place. In addition to supporting human activities, the environment is able to proactively anticipate and identify courses of action to enhance the human experience. These abilities suggest new focus areas for intelligent environment research: curiosity and proactive problem finding, life-long adaptability in dynamic environments and enhancement of human activities.

This paper discusses these focus areas and motivates the need for further research in these directions. An example of a curious place application is discussed with an emphasis on how it extends the capabilities of traditional agent-based approaches to similar systems.

2 INTELLIGENT ENVIRONMENTS

A practical application of intelligent environments is the C-Bus home automation package [2] where computational processes monitor activities within the home and respond by turning lights on and off, locking and opening doors, triggering zoned heating or cooling and activating automatic watering systems. Such home automation systems are possible with sensors and effectors that are programmed to respond deterministically to predefined triggers.

Another approach to intelligent environments is to use agents. Agents reason about the use of the room in order to facilitate human activity. This research started with the Intelligent Room Project [1, 3] and has progressed in several directions, from sensor technology and information architectures, to possible agent models for intelligent reasoning [5]. Agent societies in intelligent environments have the potential to exhibit complex emergent behaviour as a result of collaboration between agents performing different roles [10]. However, while existing agent-based systems go beyond the home automation systems to proactively support human activities, the agents still respond with programmed reflexes to predefined triggers. Curious places introduce the use of intrinsically motivated agent models to the design of intelligent environments.

3 ATTRIBUTES OF A CURIOUS PLACE

Brooks [1] and Coen [3] argue that intelligent environments should:

- adapt to and be useful for everyday activities;
- assist the user without requiring the user to attend to them;
- have a high degree of interactivity; and
- be able to understand the context in which people are trying to use them and behave appropriately.

Projects such as Active Spaces [8] and the Interactive Workspace Project [6] have produced environments that support everyday activities without user attention and can behave appropriately within a context. Motivated learning agents [7] are a type of agent that provides a way to extend the usefulness of intelligent spaces by giving them the ability to better adapt to changing patterns of human activity and potentially allowing them to anticipate user demands.

Motivated learning agents use a model of intrinsic motivation to reward activities that may be beneficial to the long-term development of the agent but may not have an immediate extrinsic reward attached. Figure 1 illustrates how the motivation process $\mathcal{M}$ takes inputs from the sensors and memory of the agent and produces output events and rewards that can be used by other processes of the agent to guide action selection and learning.

Curious agents [9] are a type of motivated learning agent that produce an intrinsic motivation reward based on the perceived novelty of a sensed experience. Computational models of curiosity incorporate adaptive components that monitor and learn from experience by paying attention to unexpected, or novel, changes in the environment. Curious agents model interest in new experiences based on their similarity with remembered experiences. Curious agents can also model

[1] School of Information Technologies, University of Sydney and National ICT Australia, email: kkas0686@it.usyd.edu.au

[2] Faculty of Architecture, Design and Planning, University of Sydney, email: [rob,mary]@arch.usyd.edu.au

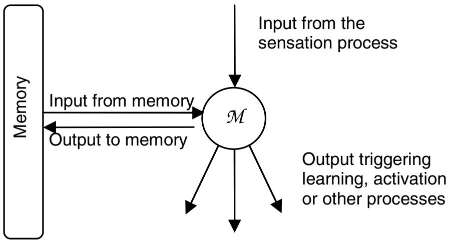

Memory
Input from the sensation process
Input from memory
Output to memory
$\mathcal{M}$
Output triggering learning, activation or other processes

Figure 1. Interaction of motivation with other agent processes.

boredom, for example, when the agent's level of interest over multiple experiences falls below a threshold.

The research presented here focuses on the development of intelligent environments using curious agents that adapt to changing user behaviour and anticipate user demands. The following sections outline three focus areas for curious place research that have the potential to extend the ability of intelligent environments to:

- proactive problem finding;
- life-long adaptability; and
- enhancing human activity.

3.1 Proactive Problem Finding

Research in intelligent environments and intelligent agents has typically focused on the development of systems that solve known problems by learning, planning or rule-based responses. The significant problem of identifying interesting problems, unknown at design time, has received less attention.

Curious places can generate their own problems to solve. The generation of a problem is triggered by a curious agent becoming bored with a predictable routine of experiences. The level of interest an agent has in a generated problem can be determined from how similar the new problem is to one that the agent has solved before [9].

3.2 Life-Long Adaptability

Human activities are not static: the daily, monthly and yearly behavioural cycles of individuals and groups shift and change over time as a result of changing biological, cognitive and social needs. Human activity is often characterised by creativity that leads to unpredictable changes in behavioural patterns. Consequently, it is difficult for system designers to predict in advance all the human behaviours that an intelligent environment may need to adapt to.

Although Brooks [1] and Coen [3] identified adaptability as a key requirement of intelligent environments relatively little research has focussed on building systems that can monitor and respond to unexpected changes in human behaviour. Machine learning has been used in intelligent environments but the focus has been on learning responses to human behaviours the system's designers have identified in advance as being important.

Curious places can monitor human activities and can identify unexpected, or interesting, behaviours. Identification and adaptation to interesting behaviours is strongly rewarded by the model of curiosity, providing the necessary feedback for a curious place to respond and adapt to novel human behaviours as they emerge.

3.3 Enhancement of Human Activities

The ability to support and be useful for human activities is a key requirement of any intelligent environment. Curious places have the potential to not only support human activities but also provide new services that enhance the human experience and can, in turn, modify the way humans interact with their environment.

Curious places can autonomously explore the potentials of their sensors and effectors allowing them to develop new behaviours or, when connected to appropriate sources of information, discover new information that was not provided by the system architects. Research in this area is thus a step in the direction of building intelligent environments that can not only assist with routine tasks but anticipate and actively contribute to creative activities within the space. The curious research space described in Section 4 is an example of such a system.

4 A Curious Research Space

We are currently implementing a curious place in our university environment as a 'curious research space'. This application is situated in a university meeting room that is equipped with sensor and effector hardware and a device control layer as shown in Figure 2.

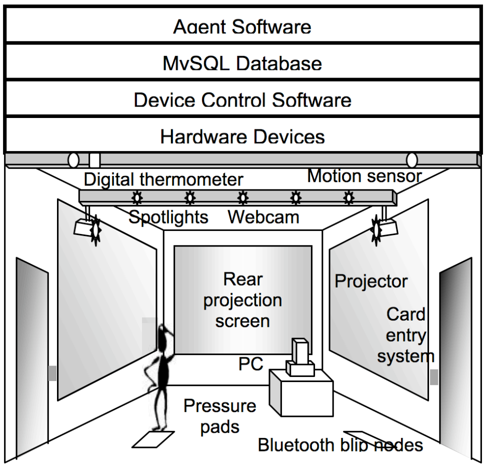


Figure 2. System architecture for a curious place.

Traditional research environments provide a physical space where human researchers can perform research activities, disseminate research findings, store equipment and collaborate. Curious research spaces extend the built environment with motivated agent technology to monitor and actively contribute to research by conducting their own research activity.

The curious research space is implemented as a society of motivated reflex agents (MRAs). MRAs incorporate models of motivation into reflex agent architectures such that actions are triggered not only by environmental stimuli but by the agents motivations. This allows MRAs to exhibit adaptive behaviour. MRAs use motivated reflexes to reason about motivation and the environment. Motivated reflexes trigger behaviour according to conditions about both the environment and the motivational state of the agent. Motivated reflexes can be implemented as rules with the following form:

```
if condition(environment stimuli) and
   condition(motivational state)
then behaviour
```

Environmental stimuli may be an observed state of the environment or an observed change in the state of the environment. Conditions define constraints on the observations or changes that trigger a particular response. A behaviour may be a single action or it may be a sequence of actions that achieve some goal. Motivation may be intrinsic as described in Section 3 or extrinsic, from the environment, e.g., rewards from other agents. In MRAs reasoning is characterised by three processes: sensation, motivation and activation as shown in Figure 3. The sensation process transforms raw sensor data into three structures: a set O of observations of the current state of the environment; a set E of events representing changes between successive states of the environment; and an environmental motivation M_e.

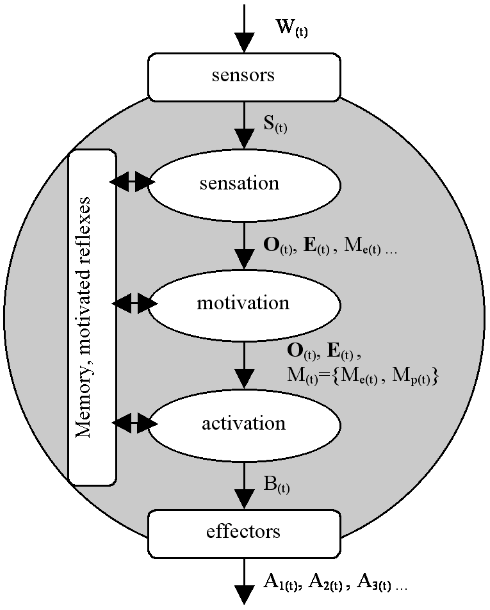


Figure 3. The motivated reflex agent architecture.

The motivation process computes intrinsic motivation, M_p, and combines it with extrinsic motivation to produce a motivational state M. The activation process uses rules representing motivated reflexes to select a behaviour B comprising actions A_1, A_2, A_3 ... that trigger effectors to make changes in the environment.

In societies of MRAs, agents playing different roles are defined by different rule sets. We define a curious research space using a society of MRAs that play the roles of keyword agents, search agents, content agents, and narrative agents. Keyword agents analyse presentations given in the room and extract interesting keywords. Keywords are communicated to search agents using the FIPA [4] communication protocol. Search agents use interesting keywords from one or more keyword agents to search the internet for related documents. Content agents analyse documents found by search agents to identify interesting documents. Structure agents identify interesting phrases, sentences or illustrations and build slides. Narrative agents construct slide shows and presentation agents perform those slide shows while monitoring the human audience.

5 DISCUSSION

We envisage that future curious places might be developed as intelligent rooms, entertainment arcades or data centres. As an intelligent room a curious place observes the actions of its inhabitants, identifies novel or interesting actions, learns about them using unobtrusive techniques, then modifies the physical environment to meet the changing needs of its users.

A curious place as an entertainment arcade might include characters or augmented reality displays that directly interact with occupants via active learning methods such as reinforcement learning. Characters and displays would be capable of actively seeking novel stimuli to provoke interaction and entertain users.

Finally, a curious place as a data centre would observe the actions of its inhabitants, or even a wider space such as an entire building or the internet, identify novel or interesting phenomena to learn about using techniques such as data mining, then modify a digital environment to reveal these finding to users.

The idea of a curious place promises a kind of sensed environment that is interested in the people that inhabit it and that may in turn be interesting to its inhabitants. Curious places extend intelligent environments with proactive problem finding ability, life-long adaptability and the ability to enhance human experience in the environment. In addition, curious places have the potential for long term support of human activity by adapting to the changing behavioural cycles of their human inhabitants.

ACKNOWLEDGEMENTS

This research was supported by a National ICT Australia PhD scholarship. National ICT Australia is funded by the Australian Governments Backing Australias Ability initiative, in part through the Australian Research Council.

This paper was written while Mary Lou Maher was working at the National Science Foundation in the United States. Any opinion, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Rodney A. Brooks, ‘The intelligent room project’, in *The Second International Cognitive Technology Conference*, pp. 271–279, (1997).

[2] Clever Home Automation, *Clipsal C-Bus*, 2007.

[3] Michael H. Coen, ‘Design principles for intelligent environments’, in *The Fifteenth National / Tenth Conference on Artificial Intelligence / Innovative Applications of Artificial Intelligence*, pp. 547–554, Madison, Wisconsin, USA, (1998).

[4] FIPA. Agent communication language specifications, http://www.fipa.org/repository/aclspecs.html, January 2007.

[5] Tracy Hammond, Krzysztof Gajos, Randall Davis, and Howard E. Shrobe, ‘An agent-based system for capturing and indexing software design meetings’, in *Proceedings of International Workshop on Agents In Design, WAID’02*, eds., John S. Gero and Frances M. T. Brazier, pp. 203–218, Cambridge, MA, (August 2002).

[6] Brad Johanson, Armando Fox, and Terry Winograd, ‘The interactive workspaces project: Experiences with ubiquitous computing rooms’, *IEEE Pervasive Computing*, **1**(2), 67–74, (2002).

[7] Mary-Lou Maher, Kathryn Merrick, and Owen Macindoe, ‘Intrinsically motivated intelligent sensed environments’, in *The Thirteenth International Workshop of the European Group for Intelligent Computing in Engineering*, pp. 455–475, Ascona, Switzerland, (2006).

[8] Manuel Román, Christopher K. Hess, Renato Cerqueira, Anand Ranganathan, Roy H. Campbell, and Klara Nahrstedt, ‘Gaia: A middleware infrastructure to enable active spaces’, *IEEE Pervasive Computing*, **1**(4), 74–83, (2002).

[9] Rob Saunders, *Curious Design Agents and Artificial Creativity*, Ph.D. dissertation, University of Sydney, 2002.

[10] Michael Wooldridge and Nicholas R. Jennings, ‘Intelligent agents: Theory and practice’, *Knowledge Engineering Review*, **10**(2), 115–152, (1995).

Modelling Group Decision Making Processes with Artificial Societies considering Emotional Factors

Goreti Marreiros[1,2], Ricardo Santos[1,3], Carlos Freitas[1], Carlos Ramos[1,2], José Neves[4] e José Bulas-Cruz[5]

Abstract. The organisational complexity, the globalisation and the internationalisation of the markets and the individual limits of the group members stress the decision taken. Actually making decisions imply to consider many different points of view, so decisions are commonly taken by formal or informal groups of persons. Group meetings are important events where ideas are exposed, alternatives considered, argumentation and negotiation take place, and emotional aspects take sometimes the same importance of rational aspects. In this work it is proposed an agent-based architecture to support a ubiquitous group decision support system for ambient intelligence environments that considers emotional factors.

1 INTRODUCTION

Groups of individuals have access to more information and more resources what will (probably) allow to reach "better" and quicker decisions. However working in group has also some difficulties associated, e.g. time consuming; high costs; improper use of group dynamics and incomplete tasks analysis.

If the group members are dispersed in time and space, the need of coordination, informal and formal forms of communication and information sharing will increase significantly. And is a fact that in a Global World meeting participants may be in different places (some in a meeting room, others in their offices, others in different countries) with access to different devices (computers, PDA, mobile phones, embedded systems in the meeting room or in their clothes) and available at different times (asynchronous meetings).

Group Decision Support Systems (GDSS) aim at reducing the loss associated to group work and to maintain or improve the gain. During group decision making process different types of conflicts and disagreements arise, and it is necessary to overcome them. Argumentation can be an excellent choice to justify possible choices and to convince other elements of the group that one alternative is better or worst than another.

Traditional meeting rooms (Figure 1) are places where groups members present ideas, analyze alternatives, make proposals, exchange arguments, considering rational and emotional aspects, vote, and make decisions.

In this work it is proposed an architecture for a ubiquitous group decision support system that is able to help people in group decision making processes and considers the emotional characteristics of participants.

Figure 1. Traditional Meeting room

This system is intended to be used for intelligent decision making, a part of an ambient intelligence environment where networks of computers, information and services are shared [1].

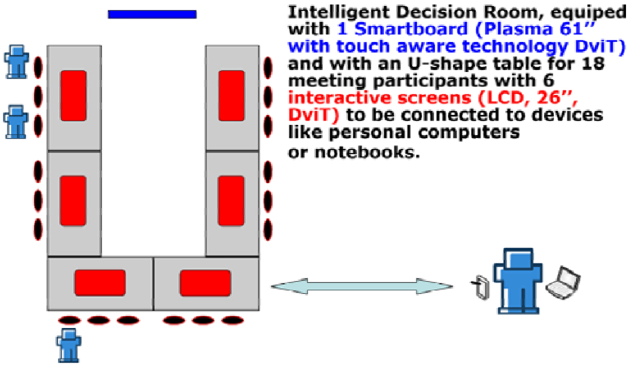

Intelligent Decision Room, equiped with 1 Smartboard (Plasma 61" with touch aware technology DviT) and with an U-shape table for 18 meeting participants with 6 interactive screens (LCD, 26", DviT) to be connected to devices like personal computers or notebooks.

Figure 2. Distributed decision meeting

As an example of a potential scenario, it is considered a distributed meeting involving people in different locations (some in a meeting room, others in their offices, possibly in different countries) with access to different devices (e.g. computers, PDAs, mobile phones, or even embedded systems as part of the meeting room or of their clothes).

Figure 2 shows an Intelligent Decision room with several interactive Smartboards. The meeting is distributed but it is also asynchronous, so participants do not need to be involved at any time (like the meeting participant using a PDA and/or a notebook in Figure 1). However, when interacting with the system, a meeting participant may wish to receive information as it appears. Meetings are important events where ideas are exposed, alternatives are considered, argumentation and negotiation take place, and where the emotional aspects of the participants are so important as the rational ones. This system will help participants, showing available information and knowledge, analyzing the meeting trends and suggesting arguments to be exchanged with others.

Group decision making processes represent very complex human activities. A better understanding of those processes implies the relation of several disciplines like for instance, psychology, sociology, political science, etc. Since a few years

[1]GECAD – Knowledge Engineering and Decision Support Group, Portugal

[2]Institute of Engineering – Polytechnic of Porto, Portugal, {goreti; csr}@dei.isep.ipp.pt

[3]College of Management and Technology– Polytechnic of Porto, Portugal, rjs@estgf.ipp.pt

[4]University of Minho, Portugal, jneves@di.uminho.pt

[5]University of Trás-os-Montes e Alto Douro, Portugal, jcruz@utad.pt

ago specialists in decision making area started to consider emotion as a factor of influence in the decision making process [2][3][4]. In psychological literature several examples could be found on how emotions and moods affects the individual decision making process. For instance, individuals are more predisposed to recall memories that are congruent with their present emotional state. There are also experiences that relate the influence of emotional state in information seeking strategies and decision procedures.

The rest of the paper is structured as follow. Section 2 presents the system architecture that we are proposing to address the representation of ubiquitous group decision making problems. Section 3 introduces one of the modules of the architecture, the Agent Based Simulation for Group Decision, with special attention to the participant agents. decision protocol used in simulation. In section 4 is presented the participant agents architecture and detailed its main components and interactions. Section 5 presents some implementation details, and finally section 6 presents some conclusions.

2 SYSTEM ARCHITECTURE

One's aim is to present a ubiquitous system able to exhibit an intelligent and emotional behaviour in the interaction with individual persons and groups. This system supports persons in group decision making processes considering the emotional factors of the intervenient participants, as well as the argumentation process.

Groups and social systems are modelled by intelligent agents that will be simulated considering emotional aspects, to have an idea of possible trends in social/group interactions.

The system consists of a suite of applications as depicted in Figure 3.

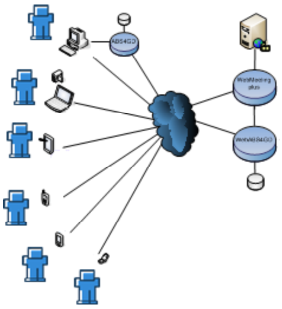


Figure 3. System architecture

The main blocks of the system are:

- WebMeeting Plus – this is an evolution of the WebMeeting project with extended features for audio and video streaming. In its initial version, based on WebMeeting [5], it was designed as a GDSS that supports distributed and asynchronous meetings through the Internet.
- ABS4GD – this is the simulation tool resulting from the ArgEmotionAgents project. ABS4GD (Agent Based Simulation for Group Decision) is a multi-agent simulator system whose aim is to simulate group decision making processes, considering emotional and argumentative factors of the participants.
- WebABS4GD – this is a web version of the ABS4GD tool to be used by users with limited computational power (e.g. mobile phones) or users accessing the system through the Internet.

3 GROUP DECISION SIMULATOR

Agent Based simulation is considered as important tool in a broad range of areas e.g. individual decision making (what if scenarios), e-commerce (to simulate the buyers and sellers behaviour), crisis situations (e.g. simulate fire combat), traffic simulation, military training, entertainment (e.g. movies).

According to the architecture that we are proposing we intend to give support to decision makers in both of the aspects identified by Zachary and Ryder [6], namely supporting them in a specific decision situation and giving them training facilities in order to acquire competencies and knowledge to be used in a real decision group meeting. We defend that agent based simulation can be used with success in both tasks. Multi-agent systems seem to be quite suitable to simulate the behaviour of groups of people working together [7][9], as well as to assist the participants presenting new arguments and feeding the simulation model of the group by observing the interaction and history of the meeting.

The Agent Based Simulator for Group Decision (ABS4GD) is a tool that can be used by one or more participants to simulate possible scenarios, to identify possible trends and to assist these participants (in this way it can be seen as a what-if tool of a decision support system). However, the criteria used by this decision support system are not just rational, since they will consider emotions [8]. In our approach the decision making simulation process considers emotional aspects and several rounds of possible argumentation between meeting participants. It is important to notice that this simulator was not developed in order to substitute a meeting or even to substitute some meeting participants.

The simulator is composed of several agents: Facilitator agent, Voting agent, Information agent and Participant agents.

The Facilitator agent will help the responsible for the simulation in it organization (e.g. decision problem and decision rules configuration). According to coordinator instructions, he will proceed to the formation of a group of agents to participate in a specific simulation. This agent will also be responsible for the inclusion of new participant agents in the community. During the simulation, the facilitator will coordinate all the process and, at the end, will summarize the results of the simulation.

Experience tells that almost all the group decision making meetings have one or more voting rounds. The Voting agent will be responsible for the tasks related with the voting simulation process, according to the decision rules settled by the Facilitator agent.

The Information agent holds information about the different proposals (alternatives) that will be evaluated by the group of agents during the group decision making simulation.

The Participant agents will simulate the role of persons in the group decision making process. All the set of participant agents form a community of participant agents. The agents are dotted of social and emotional characteristics that will personalize its behaviour. Each agent will have a model of himself, a model of the others agents, and a model of the community where he is. Through the analysis of the realized simulation the agent will constructs the others agent's profile, particularly in what is related to: reputation, credibility, preferred arguments and emotional state.

As we see the simulator is composed of several agents, but the more relevant are the participant agents because they simulate the human participants of a meeting, for that reason in section 4 in presented the participant agent architecture and particularly the way how emotions are handling.

4 PARTICIPANT AGENTS ARCHITECTURE

In figure 4 it is represented the architecture of participant agents. This architecture contains three main layers: the knowledge layer, the reasoning layer and the interaction layer.

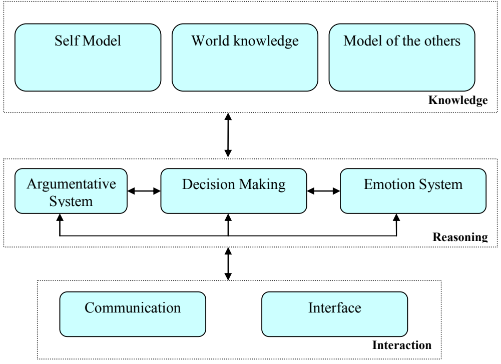


Figure 4. Participant Agents structure

4.1 Knowledge layer

In the knowledge layer the agent has information about the environment where he is situated, about the profile of the other participant's agents that compose the simulation group, and regarding its own preferences and goals (its own profile). The information in the knowledge layer is dotted of uncertainty [10] and will be accurate along the time through interactions done by the agent.

A database of profiles and history with the group's model is maintained and this model is built incrementally during the different interactions with the system.

The community should be persistent because it is necessary to have information about previous group decision making processes, focusing credibility, reputation and past behaviours of other participants [11].

4.2 Interaction layer

The interaction layer is responsible for the communication with other agents and by the interface with the user of the group decision making simulator.

4.3 Reasoning Layer

The agent must be able to reason based on complete or incomplete information. In this layer the reasoning mechanics is based on the information existent in the knowledge layer and on the messages receive from other agents through the interaction layer. The reasoning mechanism will determine the behaviour of the agent and allow the acquisition of new knowledge, based on, essentially previous experiences (simulations).

The reasoning layer contains three major modules:

- The argumentative system [9] – that is responsible for the arguments generation. This component will generate explanatory arguments and persuasive arguments, which are more related with the internal agent emotional state and about what he, think of the others agents profile (including the emotional state);
- The decision making module – will support agents in the choice of the preferred alternative and will classify all the set of alternatives in three classes: preferred, indifferent and inadmissible;
- The emotional system – will generate emotions and moods, affecting the choice of the arguments to send to the others participants, the evaluation of the received arguments and the final decision.

4.4 Emotional System

The emotions that will be simulated in our system are those identified in the reviewed version of the OCC (Ortony, Clore and Collins) model: joy, hope, relief, pride, gratitude, like, distress, fear, disappointment remorse, anger and dislike [12].

An emotion in our system is characterized by the following properties: if it is positive or negative, moment in time (simulation time) when it was initiated, identification of the agent or event that cause the emotion and emotion intensity.

The Facilitator agent will support the setup of a set of rules to configure the emotion generation. The system is prepared to allow the configuration of all the set considered in the reviewed OCC model, but the responsible may opt just to configure a subset of it. The emotional system is composed by three major blocks: appraisal, selection and decay.

4.4.1 *Appraisal*

The appraisal mechanism is based on the OCC model, where the simulator user defines the conditions for the emotion activation. An example may be:

$$Hope(AgPi,X){:}\text{-}Goal(AgP_i,X),$$
$$Request\ (AgP_j,X).$$

In the previous example the emotion Hope is appraised if Agent AgP_i has the goal (X) and asks to agent AgP_j to perform the goal X then the emotion Hope is generated.

For each condition in the emotion generation rule is settled a weight, in the interval [0,1]. The intensity of the emotion is calculated according to the conditions weight.

A particular emotion could be or not be expressed by the agent depending on the intensity of the others emotions.

4.4.2 *Selection*

All the emotions defined in the simulator have an threshold activation, that can be influenced by the agent mood. The activation threshold is a value between 0 and 1.This component selects the dominant emotion. $AgP_{i,Emo,t}$ is the set of all the emotions generated by the agent AgP_i and respective intensities and activations thresholds.

$AgP_{i,Emo,t}=\{(Emo_1,Int_1,Act_1),\ldots(Emo_n,Int_n,Act_{n)}\}$

The selected emotion in instant t, $AgPi_{ActEmo,t;}$ will be the one that have a higher differential between the intensity and the activation.

4.4.3 *Decay*

Emotions have a short duration, but they do not vanish instantaneously, they have a period of decay. There are several proposals for this calculation. In our model we consider three possibilities: Linear, Exponential and Constant.

The characterization of the decay function for each type of emotion, allows modelling the decay celerity of the different emotions.

4.4.4 *Mood*

The agent mood is calculated based on the emotions agents felt in the past and in what agent think about the moods of the remaining participants. In our approach only the process of mood contagion is being considered, we do handle the process of emotions contagion. We consider only three stages for mood: positive, negative and neutral. The mood of a specific participant is determined according the following:

$$M^+ = \sum_{i=t-n}^{t-1} I_i^+ , M^- = \sum_{i=t-n}^{t-1} I_i^-$$

Where *M*+ and *M*- represents the sum of the intensities of the emotions felt in the last *n* periods, and n can be parameterized by the simulator user. Only emotions that are above the threshold activation are considered, some of the considered emotions may not be selected as dominant, because at each moment only one emotion may prevail.

$$\begin{cases} If\ M^+ \geq M^- + l\, , then\ positive\ mood \\ If\ M^- \geq M^+ + l\, , then\ negative\ mood \\ If \left| M^+ - M^- \right| < l\, , then\ neutral\ mood \end{cases}$$

The variable *l* has an empirical value that varies according what a specific participant thinks about the mood of the group and his potential mood. We could have for instance the following values for *l*.

$$\begin{cases} l = 0.10\, , if\ group\ mood\ is\ positive\ and\ M^- \geq M^+ \\ l = 0.10\, , if\ group\ mood\ is\ negative\ and\ M^+ \geq M^- \\ l = 0.05\, , if\ group\ mood\ is\ neutral \\ l = 0.01, if\ group\ mood\ is\ negative\ and\ M^- \geq M^+ \\ l = 0.01, if\ group\ mood\ is\ positive\ and\ \ M^+ \geq M^- \end{cases}$$

5 ABS4GD implementation

Some details of the implementation of the simulator are described here. The system was developed in Open Agent Architecture (OAA), Java and Prolog. OAA is structured in order to: minimize the effort involved in the creation of new agents, that can be written in different languages and operating on diverse platforms; encourage the reuse of existing agents; and allow for dynamism and flexibility in the makeup of agent communities. More information about OAA can be found in www.ai.sri.com/~oaa/.

6 CONLUSIONS

This work proposes a simple architecture for a ubiquitous group decision making system able to support distributed and asynchronous computation. This system supports a group of people involved in group decision making, being available in any place (e.g. at a meeting room, when using a web based tool), in different devices (e.g. computers, note-books, PDAs) and at different time (e.g. on-line meeting, asynchronous meetings).

One of the key components of this architecture is a multi-agent simulator of group decision making processes, where the agents present themselves with different emotional states, being able to deal with incomplete information, either at the representation level, or at the reasoning one. The discussion process between group members is made through the exchange of persuasive arguments, built around the same premises stated to above. Future work includes the refinement of the architecture, as well as the improvement of the interaction between the simulator and the group members.

Most of these ideas covered by this work are not exclusive to Decision Making processes. There are other social interaction domains in which emotion, argumentation, ubiquitous computing and ambient intelligence are important. It is expected that in the future many new ways to perform collaborative work will appear. We expect that the experience with Group Decision Making support presented here will give some useful insights for this new way to interact in the future.

ACKNOWLEDGEMENTS

The work described in this paper is included in ArgEmotionAgents project (POSC / EIA / 56259 / 2004).

REFERENCES

1. G. Marreiros, R. Santos, C. Ramos, J. Neves, P. Novais, J. Machado, J. Bulas-Cruz, 'Ambient Intelligence in Emotion Based Ubiquitous Decision Making'. *Proc. Artificial Intelligence Techniques for Ambient Intelligence, IJCAI'07 ,* Hyderabad, India, 2007.
2. A. Damasio. *Descartes' Error: Emotion, Reason and the Human Brain*. Picador, 1994.
3. J. LeDoux, *The emotional brain*. Simon & Shuster: New York, 1996.
4. D. Goleman, *Emotional Intelligence*. New York: Bantam Books, 1995.
5. G. Marreiros, J.P. Sousa and C. Ramos, 'WebMeeting - A Group Decision Support System for Multi-criteria Decision Problems'. *Proc. ICKEDS*, Porto, Portugal pp. 63-70, 2004.
6. W. Zachary, J. Ryder, 'Decision Support Systems: Integrating Decision Aiding and Decision Training'. *In: Handbook Of Human-Computer Interaction*, pp. 1235-1258. The Netherlands: Elsevier Science, 1997.
7. G. Marreiros, C. Ramos, J. Neves, 'Dealing with Emotional Factors in Agent Based Ubiquitous Group Decision'. *Lecture Notes in Computer Science*. Vol. 3823 pp. 41-50, 2005.
8. R. Santos, G. Marreiros, C. Ramos, J. Neves, J. Bulas-Cruz, 'Multi-agent Approach for Ubiquitous Group Decision Support Involving Emotions'. *Lectures Notes for Computer Science*, 4159, pp. 1174 – 1185, 2006.
9. P. Davidsson, 'Multi agent based simulation: beyond social simulation', *Proc. of the Second international Workshop on Multi-Agent Based Simulation*. S. Moss and P. Davidsson, Eds. Springer-Verlag NJ, pp. 97-107, 2001.
10. J. Neves, 'A Logic Interpreter to Handle Time and Negation in Logic Data Bases', in *Proceedings of ACM'84*, The Fifth Generation Challenge, pp. 50-54, 1984.
11. F. Andrade, J. Neves, P. Novais, J. Machado, A. Abelha, 'Legal Security and Credibility in Agent Based Virtual Enterprises, in *Collaborative Networks and Their Breeding Environments',* Springer-Verlag, pp 501-512, 2005.
12. A. Ortony, A., 'On making believable emotional agents believable', I*n R. P. Trapple, P. (Ed.), Emotions in humans and artefacts. Cambridge*: MIT Press, 2003.
13. M. El-Nasr; J. Yen.; T.R. Ioerger. 'FLAME -Fuzzy Logic Adaptive Model of Emotions'. *Autonomous Agents and Multi-agent systems*, Vol. 3, pp. 217-257, 2000
14. R. Picard *.Affective Computing*. MIT Press, Cambridge, MA.

Towards a Model of Evolving Agents for Ambient Intelligence

Stefania Costantini[1] and **Pierangelo Dell'Acqua**[2] and **Luís Moniz Pereira**[3] and **Francesca Toni**[4]

Abstract. We propose a general vision for agents in Ambient Intelligent applications, whereby agents monitor and train unintrusively human users, and learn their patterns of behavior by observing and generalizing their observations, but also by "imitating" them. Agents can also learn by "imitating" other agents, after being told by them. Within this vision, agents need to evolve to take into account what they learn from or about users, and as a result of monitoring the users. In this paper we focus on modelling, by means of dynamic-logic-like rules, the monitoring behavior of agents, and by modelling the corresponding evolution of the agents.

1 Motivation

We envisage a setting (see Figure 1) where agents interact with users (i) with the objective of training them in some particular task, and (ii) with the aim of monitoring them for ensuring some degree of consistence and coherence in user behavior. We assume that agents are able (iii) to elicit (e.g. by inductive learning) the behavioral patterns that the user is adopting, and (iv) to learn rules and plans from other agents by imitation (or being told). In fact, learning may allow agents to survive and reach their goals in environments where a static knowledge is insufficient. Here, for some aspects related to learning we take inspiration from recent evolutionary cultural studies of human societal organization to collectively cope with their environment. We believe in fact that some principles emerging from these studies can equally apply to societies of agents. This especially when agents cooperate to help humans adapt to environments that are new to them and/or their ability to cope with the environment is too costly, non-existent or impaired.

The envisaged agents will try to either modify or reinforce the rules/plans/patterns they hold, based on appropriate evaluation performed by an internal meta-control component. The evaluation might also convince the agent to modify its own behavior by means of advanced evolution capabilities.

This overall agent model is in accordance with the vision of Ambient Intelligence as that of a digitally augmented environment centered around the needs of humans, where appliances and services proactively and unintrusively provide support and assistance.

We consider it necessary for an agent to acquire knowledge from other agents, i.e. learning "by being told" instead of learning only by experience. Indeed, this is a fairly practical and economical way of increasing abilities, widely used by human beings, as widely studied in evolutionary biology [12].

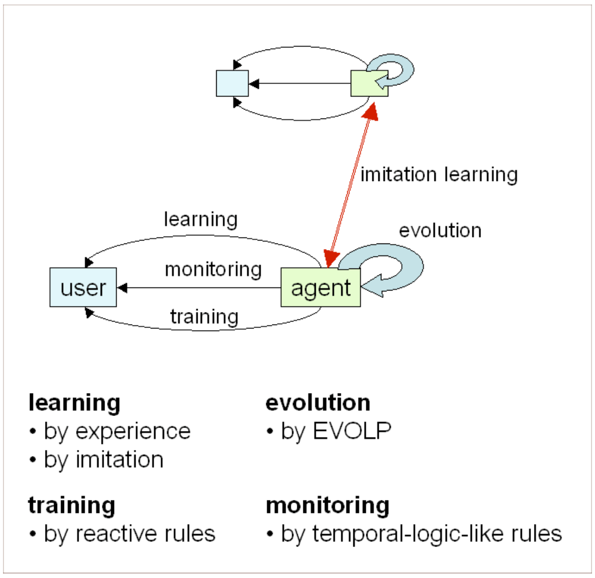


Figure 1. Agent interaction model

Note that avoiding the costs of learning is an important benefit of imitation, but nevertheless learning involves many issues and some potential risks. The issues are at least the following: (a) how to ask for what an agent needs; (b) how to evaluate the actual usefulness of the new knowledge; and, (c) how this kind of acquisition can be semantically justified in a logical agent. We will discuss issues (b) and (c) in Sections 3 and 4 while we shortly discuss (a) in Section 5. We make the simplifying assumption that agents speak the same language, and thus we overlook the problem of ontologies. We also assume that agents involved in the society are benevolent and trusted. Otherwise, incorporating and using learned knowledge would involve the management of related risks.

Note also that an agent that learns and re-elaborates the learned knowledge becomes in turn an information producer, from which others can in turn learn. Instead, an agent that just imitates blindly can be a burden for the society to which it belongs. Then, one my wonder about the effects and risks for the society to allow imitation. Evolutionary biology shows that the long-run of evolution of human

[1] Università degli Studi di L'Aquila, Dipartimento di Informatica, L'Aquila, Italy
[2] Department of Science and Technology - ITN, Linköping University, Norrköping, Sweden
[3] Centro de Inteligência Artificial (CENTRIA), Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal
[4] Department of Computing, Imperial College London, South-Kensington Campus, London, UK

societies is a mixture of learners and copiers, in which both types have the same fitness as purely individual learners in a population without copiers. To understand this result, think of imitators as information scroungers and of learners as information producers. Information producers bear a cost to learn. When scroungers are rare and producers common, almost all scroungers will imitate a producer. If the environment changes, any scroungers that imitate scroungers will get caught out with bad information, whereas producers will adapt.

Then, an agent will be able to increase its fitness in such a society in two ways: if it is capable of usefully exploiting learned knowledge thus deriving new knowledge and becoming an information producer; if it is capable to learn selectively, learning when learning is cheap and accurate, and imitating otherwise. A future direction of this work is that of equipping agents with a higher level responsible for coping with this kind of information exchange.

2 Application to Agent Societies for Ambient Intelligence

In the sequel we shall outline a model for the construction of logical agents that are able to learn and adapt agents in interaction with humans.

Let us emphasize that, to engage with humans, agents should have a description of how humans normally function. Clearly, the description will in general be initially limited to the "normal" user behavior in that ambient setting. We assume that the agents are deliberately designed and originally primed with some ambient setting in mind, and the humans are new to the setting and/or experience difficulties or impairments in coping with it. As deep learning is time consuming and costly, and thus needs not be repeated by one and all, an agent may apply a hybrid combination of both deep and imitation. The view is that all agents and the society as a whole will eventually take profit from the learning/imitation process, that can here be seen as a form of cooperation.

Each agent will thus initially contain abilities related to its supervision task. These may be enhanced by interaction with both the user and the environment, and with other similar agents. However, when some piece of knowledge is missing and a task cannot be properly carried out by an agent, that piece may eventually be acquired from the society, if extant there, for the agent may be unable or unwilling to deep learn it. Then it will exercise it in the context at hand, subsequently evaluate on the basis of such experience, and report back to the society. This evaluation of imparted knowledge builds up a network of agents' credibility and trustworthiness.

3 Agent model: sketch

In order to meet the vision outlined in Section 1, we consider an agent model composed of two layers:

- A *base layer* PA (for Personal Assistant) in charge of giving immediate answers to a user. We will assume that PA is a logic program, but will not commit to a particular semantics for it (for a survey of semantics for logic programs, see e.g. [1]). We will assume however a semantics possibly ascribing multiple models to PA, in order to deal with "uncertainty" (as we will see later). One such a semantics might be the stable model semantics [7].
- A *meta-layer* MPA in charge of updating PA when no model exists according to the chosen semantics for PA. This meta-layer relies on meta-knowledge, e.g. reporting long-term objectives about the user (e.g., safety and good health) and some domain-dependent meta-knowledge related to the PA. This domain-dependent knowledge may be updated by learning (by being told) from other agents.

To describe the dynamic changes of the user behavioral patterns as well as the environment, we assume that both PA and MPA are formalized via some kind of evolutionary programming paradigm. One possibility would be to exploit EVOLP, an extension of logic programming [8] that allows to model the dynamics of knowledge bases expressed by programs, as well as specifications that dynamically change.[5] EVOLP augments a given logic programming language by adding the new distinguished atom $assert(R)$, where R is a rule. The intended meaning is that whenever $assert(R)$ is a consequence of the given program, then the rule R is added to the program itself. Symmetrically, a rule can be removed by asserting its negation. The semantics of EVOLP is given in terms of *program sequences*, i.e., addition or removal of a rule transforms the given program into a new one which is its successor in the sequence. EVOLP is originally based on the "stable model" or (equivalently) "answer set" semantics [7]. However, EVOLP is a general framework that does not strictly depend upon the underlying semantics.

Note that the agent model we envisage here is an instance of a more general model, outlined in [6], whereby an agent results from activating some form of *control* in the context of an environment where the sensing, acting and communication capabilities can be put to work. An *initial agent* A_0 will in general *evolve* into $A_1, A_2, \ldots$ through a sequence of stages, that will be affected by the interaction with the environment, that will lead it to respond, to set and pursue goals, to either record or prune items of information, etc. This more general model admits also KGP [2, 10, 13] and DALI [4, 5, 14] as instances.

In [3] we have introduced the possibility for the agent to learn reactive rules and plans. Once acquired, the new knowledge is stored in two forms: as plain knowledge added to the set of beliefs, so that the agent is able to use it and as meta-information, that permits the agent to "trace" the new knowledge, in the sense of recording what has been acquired, when and with what expectations. The meta-information allows the meta-control to reason about these aspects. If the agent should conclude that the new rules must be removed because the expectations have not been met, the meta-information will be used to locate the rules in the set of beliefs and remove them.

In this paper, we focus on the monitoring aspects of our agent vision (see Figure 1). For this purpose, we will assume that the meta-control includes rules inspired by temporal logic, but adapted to the agent context: the basic difference is that in the case of agents there is no way of verifying these rules along the full time-line (like it is e.g. done by model-checking). In fact, the notion of truth of a temporal formula in agents that evolve is necessarily bound to be checked at certain times and, as the agents evolve, the truth value may change, thus introducing an element of non-monotonicity. These temporal logic-like rules are described in the next section.

3.1 Temporal logic-like rules

Assume given a logical formalism L in which we can express sentences (including the sentence $true$). L will possibly include quantification. Then, temporal logic-like rules are defined as follows.

Definition 3.1 *Let P and C be sentences in L. Let T be a time-stamp or a time-interval. A* safety formula *$\mathcal{F}$ is a formula of the*

[5] An implementation of EVOLP is available from http://centria.fct.unl.pt/˜jja/updates.

form $K\ F$ WHEN C *where either* $F = P$ *or* $F = P : T$, *and* $K \in \{$ALWAYS, SOMETIMES, NEVER, EVENTUALLY$\}$. *If* $K =$ EVENTUALLY *then the time-stamp is mandatory. If* C *is* $true$ *then the safety formula is abbreviated to* $K\ F$.

At a certain time t a safety formula can be either true or false (for simplicity we consider solely safety formulas $K\ P : T$ below). First, the inner sentence P will be either true or false according to the concrete logical formalism we are adopting. Then, $P : T$ is true at t if P is true at t and either T is a time-stamp and $t \leq T$ or T is a time-interval and $t \in T$.

Definition 3.2 *Let* T *be a time-stamp and* $\mathcal{F} = K\ P : T$ *a safety formula. Then:*

- $\mathcal{F}$ *is true at time t iff* $P : T$ *is true at t whenever* $K \in \{$ALWAYS, SOMETIMES, EVENTUALLY$\}$*;*
- $\mathcal{F}$ *is true at time t iff* $P : T$ *is false at t whenever* $K =$ NEVER.

Defining the truth of safety formulas in time-intervals requires the interval to be specified as a totally ordered set of discrete points.

Definition 3.3 *Let* T *be a time-interval and* $\mathcal{F} = K\ P : T$ *a safety formula. Then,* $\mathcal{F}$ *is true iff:*

- $K =$ ALWAYS *and* $\forall\ t \in T$ *it holds that* $K\ P$ *is true at t;*
- $K =$ NEVER *and* $\forall\ t \in T$ *it holds that* $K\ P$ *is false at t;*
- $K =$ SOMETIMES *and* $\exists\ t \in T$ *such that* $K\ P$ *is true at t;*
- $K =$ EVENTUALLY *and* $\exists\ t \in T$ *such that* $K\ P$ *is true at t and* $\forall\ t_2 \in T, t_2 > t$ *implies that* $K\ P$ *is true at* t_2.

Notice that the notion of truth/falsity is necessarily bound to be checked at certain times and the outcome in general will change as the agent evolves: what was *ALWAYS* (or *NEVER*, etc.) true at some point may not be so in a previous or later point.

4 User monitoring by learning-by-imitation and evolution: case study

The following scenario illustrates the dynamic aspects of the knowledge base of a PA/MPA whose knowledge evolves to reflect changes in the user behavior as well as in the environment.

Suppose we have a user who must undergo treatment for some illness and therefore needs to take medicine. He/she asks his/her personal assistant about what to do during treatment, e.g., “Can I drink a glass of wine if I have to take this medicine?” Or, more generally, the user may just ask “Can I drink a glass of wine now?” where the personal assistant should give advice based on whether there is medicine to be taken (or other related matters). Referring to the first question, PA may initially contain:

$\perp \leftarrow drink, take\_medicine$

plus default usage rules:

$drink \leftarrow not\, abnormal(drink)$

$take\_medicine \leftarrow not\, abnormal(take\_medicine)$

When asserting (triggered by the user’s question):

$drink$

$take\_medicine$

an integrity violation is detected because the symbol $\perp$ is in some models (in fact all, in the stable model semantics). The PA can ask the MPA for help, and it might provide in the first place general rules such as:

$abnormal(drink) \leftarrow not\, abnormal(take\_medicine)$

$abnormal(take\_medicine) \leftarrow not\, abnormal(drink)$

together with rules stating that facts about abnormality should be rejected. The MPA can however have meta-axioms stating that a user action which is necessary to reach a basic objective should be undertaken, e.g.

ALWAYS $do(user, A)$
WHEN $goal(G), necessary(G, A)$

$goal(healthy)$

$necessary(healthy, take\_medicine)$

then, the provided rules might be, accordingly:

$abnormal(drink) \leftarrow not\, abnormal(take\_medicine)$

$abnormal(take\_medicine) \leftarrow not\, abnormal(drink)$

$\perp \leftarrow not\, take\_medicine, mandatory(take\_medicine)$

$mandatory(take\_medicine)$

where the latter fact signifies taking the medicine cannot be avoided.

Let us assume something more, to complicate the matter a little so as to show how the MPA can evaluate rules acquired from its siblings. Assume that the MPA knows:

$illness(user, cold)$

$goal(healthy) \leftarrow illness(user, X), recover(X)$

and has learnt:

$recover(cold) \leftarrow do(user, take\_aspirin)$

Now, the MPA will check the usefulness of the learnt rule, e.g. by means of the meta-axiom:

EVENTUALLY $goal(G) \leftarrow$
$known\_conds(C),\ learnt(Cond)\ :\ t$

The intended meaning is that goal G is expected to be reached by time t, by means of: (i) what the agent knew before, here indicated as $known\_conds(C)$ and corresponding to $illness(user, cold)$ in the example; (ii) the learnt condition $Cond$, i.e., $recover(cold)$ in the example. If this does not happen, in that the PA, by virtue of its interaction with the user, does not confirm recovered health, the learnt rule can be either de-activated or removed.

In accordance with the vision of Ambient Intelligence as a digitally augmented environment which is omnipresent and can observe and supervise the situation, our assistant agent will be able to perceive and record data about user behavior. In fact, the description of the user begins with an initial form and will then be subject to evolution according to what the agent observes along the interaction. These data can be exploited by means of either induction, abduction, or some other classification method so as to predict plausible future user behavior (for induction and abduction in logic programming, see e.g. [11] and [9] respectively) . For instance, assume that our agent is able to learn that the user normally takes a drink when coming back home. This can be represented by a rule such as:

$drink \leftarrow arrive\_home$

This learnt rule can be associated with a certainty factor. When the rule becomes later confronted with subsequent experience, its certainty factor will be updated, accordingly. Whenever this factor exceeds a threshold, this may lead to assert new meta-knowledge, such as:

USUALLY $drink$ *WHEN* $arrive\_home$

Formulas including *USUALLY* are an extension to the language for meta-rules given earlier. They express simply a constraint that should be checked periodically during evolution. A specification of the frequency of the check and of the conditions under which the check should be performed may be in principle added.

Th meta-knowledge expressed by the *USUALLY* formula should be managed by the meta-control MPA. In particular, MPA should consider all constraints that involve one of the elements. In this case, the outcome should be that, whenever the user arrives home, if she/he is undergoing some treatment and should then take medicine, he/she is preemptively warned not to drink.

The initial description of the user can be either hard-wired in the agent program, or more generally be acquired from the agent society. The society will in general provide, initially and later:

- A few mandatory rules.
- Behavioral rules that each agent has the freedom to accept, reject or modify in accordance to its experience and type of user it is supervising.

5 Towards a society of agents

Throughout this paper, for the sake of simplicity we have assumed that learning is achieved via information exchange between sibling agents. However, in our envisaged system architecture, the role of the society is crucial. In fact, we plan to specify a meta-meta-level which is present in every agent which participates in the society. This higher level should be responsible for such information exchange. This could be achieved, for example, by exploiting and developing techniques based on social evaluation and consensus, involving credibility measures and overall preferences.

Thus, in this perspective, a set of rules should not be told directly by an agent to another agent but, instead, it should be acquired by the global agent society which, in turn, will have suitable self-evaluation mechanisms. According to this vision, the society will have the role of proposing behavioral rules (that are socially accepted) to its agents, which have the freedom to accept them in accordance to their experience and to the type of user they are monitoring. It is also plausible that the agent society should have the possibility to enforce mandatory rules.

In this architecture, any time an agent provides its evaluation to the agent society, that agent is responsible for the information it provides. This agent will then be rewarded in case the rule it proposes will be positively evaluated by other agents. Doing so will increase the reputation/trust of the proposing agent with respect to society, and the future rules proposed by it will be accepted with greater strength. On the contrary, agents proposing bad rules will be penalized, and eventually will be socially eliminated or outcast, and eventually replaced by new agents. The resulting agent society is thus not static, but self-evolves by trying to adapt to new situations. For example, it may revise its policy to reward/punish agents, etc.

The function of the society is particularly important in contexts where agents can be dynamically allocated to new "roles". Assume for instance that an agent is required to act as a baby-sitter. The kind of knowledge it will be equipped with can consist for instance of the following.

Mandatory rules (some examples):

- Children cannot drink alcohol. This is to be strictly observed.
- Children have to go to bed "early". Each agent can however interpret what "early" means, according to children's age and family habits.
- Children should not watch too much television. Here, each agent can define what "too much" may mean, also according to circumstances and type of program.

Optional" rules to be interpreted, adapted and possibly ignored or modified (some examples):

- Children should eat healthy food (if available, with the exception of e.g. birthday parties).
- Children should benefit from fresh air and exercise: the agent should find ways of fulfilling this requirement.

6 Concluding Remarks

There are several future directions for the ideas that we discussed and sketched in this initial work.

First, we intend to develop a full realization of these ideas, staring from EVOLP, DALI and KGP agents that provide the main elements and can be exploited in combination in an implementation. We have discussed a semantic framework for such an integration in [3].

Next, we aim at designing the meta-meta level for controlling knowledge exchange. Particular attention should be dedicated to strategies involving reputation and trust for the evaluation of learnt knowledge. The social acceptance of rules can be partly based on existing techniques and algorithms. However, we believe that an extension is necessary because, where learning is concerned, techniques that just measure reputation/trust on the basis of agents' feedback are not sufficient: some kind of economical and efficient evaluation of both the degree of compatibility of the new knowledge with an agent's previous knowledge base and of the performance of the acquired rules with respect to the expected objectives is also required.

REFERENCES

[1] K. R. Apt and R. Bol, 'Logic programming and negation: A survey', *The Journal of Logic Programming*, **19-20**, 9–71, (1994).

[2] A. Bracciali, N. Demetriou, U. Endriss, A. Kakas, W. Lu, P. Mancarella, F. Sadri, K. Stathis, G. Terreni, and F. Toni, 'The KGP model of agency: Computational model and prototype implementation', in *Global Computing: IST/FET International Workshop, Revised Selected Papers*, volume 3267 of *LNAI*, 340–367, Springer-Verlag, Berlin, (2005).

[3] S. Costantini, P. Dell'Acqua, L. M. Pereira, and A. Tocchio, 'Conditional learning of rules and plans in logical agents'. Submitted.

[4] S. Costantini and A. Tocchio, 'A logic programming language for multi-agent systems', in *Logics in Artificial Intelligence, Proc. of the 8th Europ. Conf.,JELIA 2002*, LNAI 2424. Springer-Verlag, Berlin, (2002).

[5] S. Costantini and A. Tocchio, 'The dali logic programming agent-oriented language', in *Logics in Artificial Intelligence, Proc. of the 9th European Conference, Jelia 2004*, LNAI 3229. Springer-Verlag, Berlin, (2004).

[6] S. Costantini, A. Tocchio, and F. Toni, 'A multi-layered general agent model'. Technical Report, 2006.

[7] M. Gelfond and V. Lifschitz, 'The stable model semantics for logic programming', in *Logic Programming, Proc. of the Fifth Joint Int. Conf. and Symposium*, pp. 1070–1080. MIT Press, (1988).

[8] J. J.Alferes, A. Brogi, J. A. Leite, and L. M. Pereira, 'Evolving logic programs', in *Logics in Artificial Intelligence, Proc. of the 8th Europ. Conf., JELIA 2002*, LNAI 2424, pp. 50–61. Springer-Verlag, Berlin, (2002).

[9] A. Kakas, R. Kowalski, and F. Toni. The role of abduction in logic programming, 1998.

[10] A. C. Kakas, P. Mancarella, F. Sadri, K. Stathis, and F. Toni, 'The KGP model of agency', in *Proc. ECAI-2004*, (2004).

[11] Stephen Muggleton and Luc De Raedt, 'Inductive logic programming: Theory and methods', *Journal of Logic Programming*, **19/20**, 629–679, (1994).

[12] P. J. Richardson and R. Boyd, *Not by Genes Alone - How Culture Transformed Human Evolution*, The University of Chicago Press, 2005.

[13] K. Stathis and F. Toni, 'Ambient Intelligence using KGP Agents', in *Proceedings of the 2nd European Symposium for Ambient Intelligence (EUSAI 2004)*, eds., P. Markopoulos, B. Eggen, E. H. L. Aarts, and J. L. Crowley, number 3295 in Lecture Notes in Computer Science (LNCS), pp. 351–362, Eindhoven, The Netherlands, (8–10 November 2004). Springer Verlag.

[14] A. Tocchio, *Multi-Agent systems in computational logic*, Ph.D. dissertation, Dipartimento di Informatica, Universitá degli Studi di L'Aquila, 2005.

Argumentation-based decision making for selecting communication services in ambient home environments [1]

Maxime Morge and **Paolo Mancarella** [2]

Abstract. We propose here an Argumentation Framework (AF) for decision making in order to select services in ambient environments. A logic language is used as a concrete data structure for holding the statements like knowledge, goals, and actions. Different priorities are attached to these items. These concrete data structures consist of information providing the backbone of arguments. In this way, our AF selects some services but also provides an interactive and intelligible explanation of the choices.

1 INTRODUCTION

Service selection is the act of taking several material or immaterial products, and choosing some of them to meet the needs of a given customer. Indeed, when a user identifies her needs and specifies them with high-level and abstract terms, there should be a possibility to choosing some existing services. The related issues are being addressed by ongoing work in the area of the Semantic Web, Business Processes Workflow Management, and MultiAgent Systems (MAS). The latter offers solutions where the service selection could be performed dynamically by agents through negotiation [3].

Service selection requires MAS algorithms for negotiation, in order to select the best services taking account the user's constraints. Negotiation is a form of interaction in which a group of agents, with conflicting interests and a desire to cooperate, try to come to a mutually acceptable agreement [9]. Various decision mechanisms for automated negotiation have been proposed and studied. These include: game-theoretic analysis; heuristic-based approaches; and argumentation-based approaches. The main distinguishing feature of the latter is that it allows for more sophisticated forms of interaction. In this paper we present an Argumentation Framework (AF) for decision-making in order to perform the service selection. A logic language is used as a concrete data structure for holding the statements like knowledge, goals, and actions. Different priorities are attached to these items. These concrete data structures consist of information providing the backbone of arguments. In this way, our AF selects some solutions but also provides an interactive and intelligible explanation of the choices that could enrich the negotiation.

Section 2 introduces the walk-through example. In order to present our Argumentation Framework (AF) for decision making, we will browse the following fundamental notions. First, we define the *object language* (cf Section 3). Second, we will focus on the internal structure of *arguments* (cf Section 4). We present in Section 5 the *interactions* between them. These relations allow us to give a declarative model-theoretic *semantics* to this framework and we adopt a *dialectical proof procedure* to implement it (cf Section 6). Section 7 discusses some related works and draws some conclusions.

2 WALK-THROUGH EXAMPLE

Ambient communication aims at enabling new forms on human communication in ambient home environments. Inspired by [11], we consider here a flexible system for selecting services in ambient communication environments. An agent is in charge of managing requirements and selecting some services.

The agent is responsible for selecting a suitable application, either based on explicit user needs or on context-based rules. She combines situation-specific constraints provided by devices and their knowledge on typical services. The main goal, that consists of selecting the services (g_0), is addressed by four decisions: the selection of the $\mathtt{Audio}$, $\mathtt{Video}$, $\mathtt{Txt}$ and $\mathtt{Content}$ channel. The assistant agent must select, for each decision, one alternative. For instance, $\mathtt{Txt}(x)$ with $x \in \{\mathtt{im}^3, \mathtt{mobile}, \mathtt{mail}, \mathtt{none}\}$. The main goal ($g_0$) is split into sub-goals. The service must be adapted for 'important' (g_1), 'urgent' (g_2), and 'persistent' (g_3) communications. These high-level goals reveal the users'needs. The knowledge about the context is expressed with predicates such as: $\mathtt{Loc}(\mathtt{user}, \mathtt{office})$ (the user is in her office space), or $\mathtt{Ste}(\mathtt{user}, \mathtt{free})$ (the user is free).

Figure 1 provides a simple graphical representation of the decision problem called influence diagram [4]. The elements of the decision problem, i.e. *values* (represented by rectangles with rounded corners), *decisions* (represented by squares) and *knowledge* (represented by ovals), are connected by arcs where predecessors affect successors. We consider here a multiattribute decision problem captured by a hierarchy of values where the abstract value (represented by a rectangle with rounded corner and double line) aggregates the values in the lower level. While the influence diagram displays the structure of the decision, the object language reveals the hidden details of the decision.

3 THE OBJECT LANGUAGE

Since we want to provide a computational model of decision making and we want to instantiate it for our case study, we need to specify a particular logic.

The object language expresses rules and facts in logic-programming style. In order to address a decision making problem, we distinguish:

- a set of *goals*, i.e. some propositional symbols which represent the features that the decision must exhibit (denoted by $g_0, g_1, g_2, \dots$);

[1] This work is supported by the Sixth Framework IST programme of the EC, under the 035200 ARGUGRID project.

[2] Universita di Pisa, Italy, email: {morge,paolo}@di.unipi.it

[3] im stands for instant messaging.

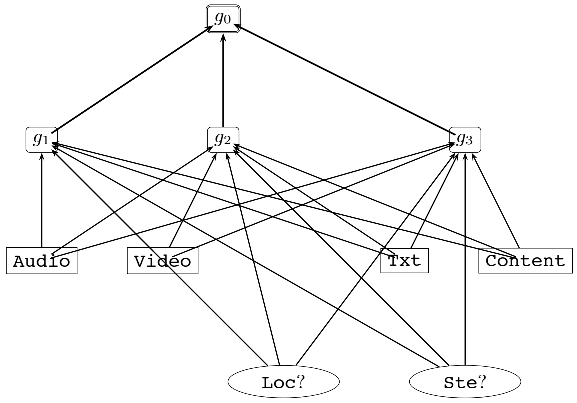


Figure 1. Influence diagram to structure the decision

- a *decision*, i.e. a predicate symbol which represents the action which must be performed (denoted by D);
- a set of *alternatives*, i.e. some constants symbols which represent the mutually exclusive solutions for the decision (eg `im`, `mail`);
- a set of *beliefs*, i.e. some predicate symbols which represent epistemic statements (denoted by words such as `Loc`, or `Ste`).

Since we want to consider conflicts in this object language, we need some form of negation. For this purpose, we only consider strong negation, also called explicit or classical negation[4]. A strong literal is an atomic formula, possible preceded by strong negation $\neg$. $\neg L$ says "L is definitely not the case". Since we restrict ourselves to logic programs, we cannot express in a compact way the mutual exclusion between alternatives. For this purpose, we define the incompatibility relation (denoted by $\mathcal{I}$) as a binary relation over atomic formulas which is transitive and symmetric. Obviously, $L \mathcal{I} \neg L$ for each atom L, and $D(a_1) \mathcal{I} D(a_2)$, a_1 and a_2 being different alternatives. Similarly, we say that a sentence ϕ_1 is incompatible with a set of sentences Φ_2 ($\phi_1 \mathcal{I} \Phi_2$) iff there is a sentence ϕ_2 in Φ_2 such as $\phi_1 \mathcal{I} \phi_2$. A theory gathers the statements about the decision making problem.

Definition 1 (Theory) *A* theory $\mathcal{T}$ *is an extended logic program, i.e a finite set of rules of the form* $R : L_0 \leftarrow L_1, \ldots, L_n$ *with* $n \geq 0$, *each* L_i *being a strong literal. The literal* L_0*, called the* head *of the rule, is denoted by* $L_0 = head(R)$*. The finite set* $\{L_1, \ldots, L_n\}$*, called the* body *of the rule, is denoted by* $body(R)$*. The body of a rule can be empty. In this case, the rule is called a* fact*.* R*, called the name of the rule, is an atomic formula.*

Considering a decision making problem, we distinguish:

- *goal rules* of the form $R : g_0 \leftarrow g_1, \ldots, g_n$ with $n > 0$. Each g_i is a goal. According to this rule, the head goal is reached if the goals in the body are reached;
- *epistemic rules* of the form $R : B_0 \leftarrow B_1, \ldots, B_n$ with $n \geq 0$. Each B_i is a belief literal;
- *decision rules* of the form $R : g \leftarrow D, B_1, \ldots, B_n$ with $n \geq 0$. The head of this rule is a goal and the body includes a decision literal (D) and a possible empty set of belief literals. According to this rule, the goal can be eventually reached by the decision D, provided that conditions $B_1, \ldots, B_n$ are satisfied.

Considering statements in the theory is not sufficient to take a decision, since some priorities between these pieces of information should be taken into account. For this purpose, we consider that the *priority* $\mathcal{P}$ is a (partial or total) preorder on $\mathcal{T}$. $R_1 \mathcal{P} R_2$ can be read "R_1 has priority over R_2". We define three priority relations:

- the priority over *goal rules* comes from their levels of *preference*. Let us consider two goal rules R_1 and R_2 with the same head ($g_0 = \text{head}(R_1) = \text{head}(R_2)$). R_1 has priority over R_2 if the achievement of the goals in the body of R_1 are more "important" than the achievement of the goals in the body of R_2 as far as reaching g_0 is concerned;
- the priority over *epistemic rules* comes from their levels of *certainty*. Let us consider, for instance, two facts F_1 and F_2. F_1 has priority over F_2 if the first is more likely to hold than the second one;
- the priority over *decision rules* comes from their levels of *credibility*. Let us consider two rules R_1 and R_2 with the same head. R_1 has priority over R_2 if the first conditional decision is more credible than the second one.

The goal theory, the epistemic theory (resp. the decision theory) are represented in Table 1 (resp. Table 2).

To simplify the graphical representation of the theories, they are stratified in non-overlapping subsets, i.e. different levels. The *ex æquo* rules are grouped in the same level. Non-comparable rules are arbitrarily assigned to a level. According to the goal theory, the achievement of g_1, g_2 and g_3 is required to reach g_0, but this constraint can be relaxed and the achievement of g_3 is more important than the achievement of g_2 which is is more important than the achievement of g_1 to reach g_0. According to the epistemic theory, the assistant agent does not know where the user is. Due to conflicting sources of information, the agent has conflicting beliefs about the state of the user. Since these sources of information are more or less reliable, $F_3^\beta \mathcal{P} F_1^\beta$. According to the decision theory, the user prefers instant messaging and mobile to mail and no text for urgent communications. However the credibility of these alternatives depends on the context: the location and the state of the user.

Table 1. The goal theory and the epistemic theory

↑	$R^\alpha_{012} : g_0 \leftarrow g_1, g_2, g_3$	↑	$F_1^\beta : \texttt{Ste}(\texttt{user}, \texttt{free}) \leftarrow$
	$R^\alpha_{023} : g_0 \leftarrow g_2, g_3$ $R^\alpha_{013} : g_0 \leftarrow g_1, g_3$ $R^\alpha_{012} : g_0 \leftarrow g_1, g_2$		$F_3^\beta : \neg\texttt{Ste}(\texttt{user}, \texttt{free}) \leftarrow$
	$R^\alpha_{03} : g_0 \leftarrow g_3$		
	$R^\alpha_{02} : g_0 \leftarrow g_2$		
	$R^\alpha_{01} : g_0 \leftarrow g_1$		

We will build now arguments in order to compare the alternatives.

4 ARGUMENTS

Due to the recursive nature of arguments (arguments are composed of subarguments, subarguments for these subarguments, and so on), we adopt and extend the tree-like structure for arguments proposed in [12].

[4] Weak negation considered eg in [8] seems not to be useful in our applications.

Table 2. The decision theory

↑	$R^{\delta}_{21} : g_2 \leftarrow \mathtt{Txt(im)}, \mathtt{Loc(user, office)}, \mathtt{Ste(user, free)}$
	$R^{\delta}_{21} : g_2 \leftarrow \mathtt{Txt(mobile)}, \mathtt{Ste(user, free)}$
	$R^{\delta}_{23} : g_2 \leftarrow \mathtt{Txt(mail)}, \neg\mathtt{Ste(user, free)}$
	$R^{\delta}_{24} : g_2 \leftarrow \mathtt{Txt(none)}$

Definition 2 (Argument) *An argument has a conclusion, top rules, premises, suppositions, and sentences. These elements are abbreviated by the corresponding prefixes. An argument A is:*

1. *a* supposal argument *built upon an unconditional ground statement. If L is a ground literal such that there is no rule R in $\mathcal{T}$ which can be instantiated in such a way that $L = head(R)$, then the argument, which is built upon this ground literal is defined as follows: $\mathtt{conc}(A) = L$, $\mathtt{top}(A) = \emptyset$, $\mathtt{premise}(A) = \emptyset$, $\mathtt{supp}(A) = \{L\}$, $\mathtt{sent}(A) = \{L\}$.*
   or
2. *a* trivial argument *built upon an unconditional ground statement. If F is a fact in $\mathcal{T}$, then the argument A, which is built upon the ground instance F^g of F, is defined as follows: $\mathtt{conc}(A) = head(F^g)$, $\mathtt{top}(A) = F^g$, $\mathtt{premise}(A) = \{head(F^g)\}$, $\mathtt{supp}(A) = \emptyset$, $\mathtt{sent}(A) = \{head(F^g)\}$.*
   or
3. *a* tree argument *built upon an instantiated rule such that all the literals in the body are the conclusion of subarguments. If R is a rule in $\mathcal{T}$, we define the argument A built upon a ground instance R^g of R as follows. Let $\{L_1, \ldots, L_n\}$ be the body of R^g and $\mathtt{subarg}(A) = \{A_1, \ldots, A_n\}$ be a collection of arguments such that, for each $L_i \in body(R^g)$, $\mathtt{conc}(A_i) = L_i$ each A_i is called a subargument of A. Then: $\mathtt{conc}(A) = head(R^g)$, $\mathtt{top}(A) = R^g$, $\mathtt{premise}(A) = body(R^g)$, $\mathtt{supp}(A) = \cup_{A' \in \mathtt{subarg}(A)} \mathtt{supp}(A')$, $\mathtt{sent}(A) = \cup_{A' \in \mathtt{subarg}(A)} \mathtt{sent}(A') \cup body(R^g)$.*
   Moreover, a tree argument must be consistent, i.e. $\mathtt{sent}(A)$ is neither incompatible with itself nor incompatible with $\mathtt{conc}(A)$.

The set of arguments built upon $\mathcal{T}$ is denoted $\mathcal{A}(\mathcal{T})$.

As in [12], we consider *atomic* arguments (2) and *composite* arguments (3). Moreover, we distinguish *supposal* arguments (1) and *built* arguments (2/3). Notice that we add a technically essential constraint on arguments that is commonly assumed in the literature, namely that each argument is consistent. Due to the abductive nature of decision making, we define and construct arguments by reasoning backwards. Therefore, arguments are minimal, i.e. they do not include irrelevant information such as sentences not used to prove the conclusion. Notice that the different premises can be challenged and can be supported by composite arguments. In this way, arguments are intelligible explanations.

Triples of conclusions - premises - suppositions are simple representations of arguments. For example, some of the arguments concluding g_2 are the following:

$-A^2 = \langle g_2, (\mathtt{Txt(im)}, \mathtt{Loc(user, office)}, \mathtt{Ste(user, free)}), (\mathtt{Txt(im)}, \mathtt{Ste(user, free)})\rangle$;
$-B^2 = \langle g_2, (\mathtt{Txt(mobile)}, \mathtt{Ste(user, free)}), (\mathtt{Txt(mobile)}, \mathtt{Ste(user, free)})\rangle$;
$-C^2 = \langle g_2, (\mathtt{Txt(mail)}, \neg\mathtt{Ste(user, free)}), (\mathtt{Txt(mail)}, \neg\mathtt{Ste(user, free)})\rangle$;
$-D^2 = \langle g_2, (\mathtt{Txt(none)}), (\mathtt{Txt(none)})\rangle$.

Let us focus on A^2. This tree argument is built with two supposal arguments and one trivial argument:

$-A = \langle \mathtt{Txt(im)}, \emptyset, (\mathtt{Txt(im)})\rangle$;
$-B = \langle \mathtt{Loc(user, office)}, \emptyset, \emptyset\rangle$;
$-C = \langle \mathtt{Ste(user, free)}, \emptyset, (\mathtt{Ste(user, free)})\rangle$.

Due to their structure/nature, arguments interact with one another.

5 Interactions between arguments

The interactions between arguments may come from their nature, from the incompatibility of their sentences, and from the priority relation between the top rules of built arguments. We examine in turn these different sources of interaction.

Since sentences are conflicting, arguments interact with one another. For this purpose, we define the attack relation.

Definition 3 (Attack relation) *Let A and B be two arguments. A* attacks *B (denoted by attacks (A, B)) iff $\mathtt{conc}(A)\ \mathcal{I}\ \mathtt{sent}(B)$.*

This attack relation, often called *undermining* attack, is indirect, i.e. directed to a “subconclusion”. However, the direct attack, also called *rebuttal* attack, can also be obtained [7]. Since each argument is consistent, it does not attack itself. The attack relation is useful to build an argument which is an homogeneous explanation.

Since arguments have different natures (supposal or built) and the top rules of built arguments are more or less strong, they interact with one another. For this purpose, we define the strength relation.

Definition 4 (Strength relation) *Let A_1 be a supposal argument, and A_2, A_3 be two built arguments. 1) A_2 is stronger than A_1 (denoted $A_2\ \mathcal{P}^{\mathcal{A}}\ A_1$); 2) If $\mathtt{top}(A_2)\ \mathcal{P}\ \mathtt{top}(A_3)$, then $A_2\ \mathcal{P}^{\mathcal{A}}\ A_3$;*

Since $\mathcal{P}$ is a preorder on $\mathcal{T}$, $\mathcal{P}^{\mathcal{A}}$ is a preorder on $\mathcal{A}(\mathcal{T})$. Obviously, arguments built upon the existing knowledge are preferred to supposal arguments. When we consider two built arguments, we adopt the last link principle: the stronger the top rule is, the better the argument is. The strength relation is useful to choose (when it is possible) between homogeneous concurrent explanations, i.e. non conflicting arguments with the same conclusions.

The two previous relations can be combined to choose (if possible) between non-homogeneous concurrent explanations, i.e. conflicting arguments with the same conclusions.

Definition 5 (Defeats) *Let A and B be two arguments. A* defeats *B (written defeats (A, B)) iff attacks (A, B) and $\neg(B\ \mathcal{P}^{\mathcal{A}}\ A)$. Similarly, we say that a set S of arguments defeats an argument A if A is defeated by one argument in S.*

Since A^2, B^2, C^2, and D^2 suggest incompatible alternatives, these arguments attack each other. Since the top rule of A^2 and B^2 (i.e. R^{δ}_{21} and R^{δ}_{22}) have priority over the top rule of C^2 and D^2 (i.e. R^{δ}_{23} and R^{δ}_{24}), A^2 and B^2 defeat C^2 and D^2. If we only consider these four arguments, the agent cannot decide what the best alternative is. However, B^2, which is composed of one supposal argument, is "better" than A^2, which is composed of two supposal arguments. Determining whether a service is ultimately selected requires a complete analysis of all arguments and subarguments.

6 SEMANTICS AND PROCEDURES

We can consider our AF abstracting away from the logical structures of arguments. This abstract AF consists of a set of arguments associated with a binary defeat relation. It can be equipped with various semantics, which can be computed by dialectical proof procedures.

Given an AF, "acceptable" sets of arguments[5] are defined as follows:

Definition 6 (Semantics) *An AF is a pair* $\langle \mathcal{A}, \ defeats \ \rangle$ *where* $\mathcal{A}$ *is a set of arguments and* $defeats \ \subseteq \mathcal{A} \times \mathcal{A}$ *is the defeat relationship*[5] *for AF. For* $A \in \mathcal{A}$ *an argument and* $S \subseteq \mathcal{A}$ *a set of arguments, we say that:*

- A *is* acceptable *with respect to S (denoted* $A \in \mathcal{S}^S_{\mathcal{A}}$*) iff* $\forall B \in \mathcal{A}, \ defeats \ (B, A) \ \exists C \in S \ \ such \ that \ \ defeats \ (C, B);$
- S *is* conflict-free *iff* $\forall A, B \in S \ \neg \ defeats \ (A, B);$
- admissible *iff S is conflict-free and* $\forall A \in S, \ A \in \mathcal{S}^S_{\mathcal{A}};$
- preferred *iff S is maximally admissible;*

The semantics of an admissible (or preferred) set of arguments is credulous, in that it sanctions a set of arguments as acceptable if it can successfully dispute every arguments against it, without disputing itself. However, there might be several conflicting admissible sets. Various sceptical semantics have been proposed for AF [5]. Since an ultimate choice amongst various admissible set of alternatives is not always possible, we adopt a credulous semantics. The decision $D(a_1)$ is *suggested* iff $D(a_1)$ is a supposition of one argument in an admissible set.

Since our practical application requires to specify the internal structure of arguments, we adopt the procedure proposed in [7] to compute admissible arguments. If the procedure succeeds, we know that the argument is contained in a preferred set. We have implemented our AF, called MARGO[6] (Multiattribute ARGumentation framework for Opinion explanation). For this purpose, we have translated our AF in an assumption-based AF (ABF for short). CaSAPI[7] computes the admissible semantics in the ABF by implementing the procedure proposed in [7]. Moreover, we have developed a CaSAPI meta-interpreter to relax constraints on the goals achievements and to make suppositions in order to compute the admissible semantics in our concrete AF. In this section, we have shown how arguments in the framework can be categorized in order to select some services.

7 RELATED WORKS AND CONCLUSIONS

The Belief-Desire-Intention (BDI) model of agency is the most famous model of agents for decision making. However, the simplifying assumptions made to implement modal logic specifications of BDI agents meant that they lack of a strong theoretical underpinning [10]. That is the reason why [6] proposes the KGP model [6] adopting Knowledge, Goals, and Plans as the main component of an agent state. However, this model deals only partially with priorities, as required by service selection, eg preferences between goals, reliability of knowledge, and credibility of possible actions. For this purpose, we have provided here a suitable revised representation of knowledge, goals and actions, Future investigations must make planning abilities available.

[8, 1] focus on AFs for selecting single actions. [1] (resp. [2]) is a mathematical (resp. philosophical) general approach of defeasible argumentation for practical reasoning. To the best of our knowledge, the existing AFs for decision making leave the underlying language unspecified contrary to our AF.

In this paper we have proposed a concrete AF for selecting services which provides an interactive and intelligible explanation of the choices made to reach such selection. Moreover we have implemented this AF and test it for this usecase. A logic language is used as a concrete data structure for holding the statements like knowledge, goals, and actions. Different priorities are attached to these items corresponding to the reliability of the knowledge, the preferences between goals, and the credibility of actions. These concrete data structures consist of information providing the backbone of arguments. Due to the abductive nature of decision making, arguments are built by reasoning backwards. To be intelligible, arguments are defined as tree-like structures. Since an ultimate choice amongst various admissible set of services is not always possible, we have adopted a credulous semantics. Future investigations must explore how to drive argumentation-based negotiations between agents.

References

[1] Leila Amgoud, 'A unifed setting for inference and decision: an argumentation-based approach', in *Proc. of the 5th Workshop on Computational Models of Natural Arguments*, (2005).

[2] Trevor Bench-Capon and Henry Prakken, 'Justifying actions by accruing arguments', in *Proc. of the 1st International Conference on Computational Models of Argument*, pp. 247–258. IO Press, (2006).

[3] Yasmine Charif-Djebbar and Nicolas Sabouret, 'Dynamic service composition and selection through an agent interaction protocol', in *Proc. of International Workshop on Service Composition*. IEEE Computer Society, (2006).

[4] Robert Taylor Clemen, *Making Hard Decisions*, Duxbury. Press, 1996.

[5] Phan Minh Dung, 'On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games', *Artif. Intell.*, **77**(2), 321–357, (1995).

[6] Antonis C. Kakas, Paolo Mancarella, Fariba Sadri, Kostas Stathis, and Francesca Toni, 'The kgp model of agency', in *Proc. of ECAI*, pp. 33–37, (2004).

[7] Francesca Toni Phan Minh Dung, Robert A. Kowalski, 'Dialectic proof procedures for assumption-based, admissible argumentation', *Artif. Intell.*, **170**(2), 114–159, (2006).

[8] Henry Prakken and Giovanni Sator, 'Argument-based logic programming with defeasible priorities', *Journal of Applied Non-classical Logics*, **7**, 25–75, (1997).

[9] I. Rahwan, S. D. Ramchurn, N. R. Jennings, P. McBurney, S. Parsons, and L. Sonenberg, 'Argumentation-based negotiation', *The Knowledge Engineering Review*, **18**(4), 343–375, (2003).

[10] Anand S. Rao, 'Agentspeak (l): Bdi agents speak out in a logical computable language', in *in Agents Breaking Away, 7th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW'96)*, ed., Rudy Van Hoe, number 1038 in Lecture Notes of Computer Science, pp. 42–55. Springer-Verlag, (1996).

[11] Mathieu Vallée, Fano Ramparany, and Laurent Vercouter, 'Using device services and flexible composition in ambient communication environments', in *1st International Workshop on Requirements and Solutions for Pervasice Software Infrastructure (RSPSI)*, (2006).

[12] Gerard Vreeswijk, 'Abstract argumentation systems', *Artificial Intelligence*, **90**(1-2), 225–279, (1997).

[5] Actually, in [5] the defeat relation is called attack.

[6] https://margo.sourceforge.net/

[7] http://www.doc.ic.ac.uk/∼dg00/casapi.html

Affective Smart Environments

Ambient Intelligence (AmI) is an emerging and popular research field with the goal to create "smart" environments that react in an attentive, adaptive and proactive way to the presence and activities of humans, in order to provide the services inhabitants of these environments request or are presumed to need. AmI is increasingly affecting our everyday lives: computers are already embedded in numerous everyday objects like TV sets, kitchen appliances, or central heating, and soon they will be networked, with each other as well as with personal ICT devices like organizers or cell phones. Communication with ambient computing resources will be ubiquitous; bio-sensing will allow devices to perceive the presence and state of users and to understand their needs and goals in order to improve their general living conditions and actual well-being.

According to the *Computers As Social Actors* paradigm, interaction with technology is driven by rules that derive from social psychology. These aspects become even more relevant when media are not boxed in a desktop computer but are integrated pervasively in everyday life environments. An Affective Smart Environment should be able to grasp these factors and adapt its behavior accordingly. Imagining and designing this kind of environment requires combining knowledge and methods of ubiquitous and pervasive computing with those of affective and social computing. And as yet there exists little in the way of coherent models of interaction on which to base our design approaches to such environments.

This symposium is an interdisciplinary meeting focused on methods and techniques for integrating affective and social factors in ambient intelligence, including: non invasive methods for sensing, recognizing and modeling the emotional state of users in 'natural', everyday situations; methods and models for profiling emotion information; methods for building the inhabitants' group profiles from their individual models; methods for learning long-term features, from tracing of interaction histories; methods for inferring how to adapt the environment to the recognized situation; methods for enforcing the sense of trust in the environment; and theoretical approaches to the design of ambiently intelligent interaction.

Berardina Nadja De Carolis, Christian Peter & John Waterworth (Symposium Chairs)

Programme committee: Berardina Nadja De Carolis (Unversity of Bari); Christian Peter (Fraunhofer IGD); John Waterworth (Umeå University); Elisabeth Andre (University of Augsburg); Russell Beale (University of Birmingham); Gerald Bieber (Fraunhofer IGD); Andreas Butz (LFM University); Marco Combetto (Microsoft Research Cambridge); Lola Cañamero (University of Hertfordshire); Giovanni Cozzolongo (University of Bari); Fiorella de Rosis (University of Bari); Rino Falcone (ISTC-CNR); Thomas Kirste (University of Rostock); Catherine Pelachaud (Université de Paris 8); Paolo Remagnino (Kingston University); Thomas Rist (University of Applied Sciences Augsburg & DFKI); Massimo Zancanaro (ITC IRST)

Automatic Generation of Relational Reports for Teamwork

Massimo Zancanaro, Bruno Lepri, Elena Not and Fabio Pianesi[1]

Abstract. Multimodal analysis of group behavior is a relatively recent research area compared to the large body of studies focusing on multimodality as a flexible, efficient, and powerfully expressive means for human-computer interaction. The availability of rich multimodal information makes it possible to explore the possibility of automatically providing services oriented to the group. In this paper, we investigated the feasibility and the acceptability of a functionality inspired by *coaching*: it consists of a report about the social behaviour of each participant of a meeting; the report is generated from multimodal information, and privately delivered. The underlining idea is that the individual, the group(s) they are parts of, and the whole organization might benefit from an increased awareness of participants about their own behavior during meetings.

1. INTRODUCTION

Meetings are more and more important in structuring daily work in organizations. For example, according to a survey in [15] executives spend on average 40%-50% of their working hours in meetings; 50% of that time is unproductive and up to 25% of it is spent discussing irrelevant issues. This situation is determined not only by task related factors (e.g., a difficult in choosing the right items for the agenda, and/or in focusing the attention on relevant issues), but often by the complexity of group dynamics in small groups, which hinders the performance of teams. Different means can be put at work to support dysfunctional teams, ranging from facilitation to training sessions conducted by experts.

In discussing the role of collaboration for teachers and in particular peer coaching, Andersen [1] suggests that coaching sessions provide a scheduled opportunity to step out of the reflexive mode and think reflectively, and that the coaching process allows the externalization of both though contents and processes that are normally internal, making them available to examination. There are three stages in the reflective process [9]: (i) the return to experience (what happened during the meeting?); (ii) attending to feelings (how did I feel, why did I act or react this way?); and (iii) the re-evaluation of the experience (what does it mean?). Reflective thinking can be effectively fostered in coaching relationship: by bringing a different perspective to the relationship, the coach can see both circumstances and possibilities that the coachee cannot [8].

The availability of rich multimodal information when meeting rooms are equipped with technology for audio-visual scene capture makes it possible to explore the possibility of providing some of these services (semi-)automatically. In this paper, we investigated the feasibility, the usefulness and the acceptability of a functionality inspired by *coaching*: it consists of a report about the social behaviour of individual participants that is generated from multimodal information, and privately delivered to them. The underlining idea is that the individual, the group(s) they are part of, and the whole organization might benefit from an increased awareness of participants about their own behavior during meetings.

This work is structured as follow: next section discusses relevant works in the areas of multimodality and group modelling; section 3 discusses the acceptability study conducted as part of the user-centered design and presents part of the user studies conducted to assess the acceptability of such a service; section 4 introduces the coding scheme used to describe the functional roles and its reliability; section 5 presents the multimodal corpus used to train the learning component used to automatically classify roles which is in turn introduced in section 6. Finally, section 7 describes the generation component.

2. PREVIOUS AND RELATED WORK

Multimodal analysis of group behavior is a relatively recent research area compared to the large body of studies focusing on multimodality as a flexible, efficient, and powerfully expressive means for human-computer interaction.

McCowan et al. [23] developed a statistical framework based on Hidden Markov Models to recognize actions that belong to the group as a whole from multimodal features extracted from individuals' actions. For example, "discussion" is a group action which can be recognized from the verbal activity of individuals.

Rienks and Heylen [27] used Support Vector Machines to automatically detect the team members who play a dominating role in a meeting, by relying on a few basic features.

Rienks and colleagues [28] addressed the problem of automatically detecting participant's influence levels in meetings using static models (i.e. SVMs) and a dynamic model, the team-player influence model (a dynamic Bayesian network with a two-level structure: the player level and the team level).

Banerjee and Rudnick [5] proposed a simple taxonomy of participant roles and meeting states. Then, they trained a decision tree classifier to learn roles and states from simple speech-based features. The classifier takes as input a feature representation of a short time window during the meeting (meeting history) and classifies the roles and the states at the end of the window.

[1] ITC-irst, Italy. Email {zancana, lepri, pianesi, not}@itc.it

Finally, Brdiczka and colleagues [10] developed a real-time detector for configurations of interaction groups. The detector is based on a Hidden Markov Model built upon conversational hypotheses and the detector's input is a speech activity vector containing information about who is speaking.

Regarding applications to support actual teamwork, the Neem project [6] explores the use of multimodal context of human-to-human communication by "overhearing" conversations in order to monitor user actions and react to perceived opportunities for augmentation via intelligent system interventions. The shift is from employing multimodality in human-computer interaction [24] to use it to understand the context of interaction. Yet the ultimate goal is still to help the group in having easy access to computerized services.

In the field of CSCW where the focus is often in distributed meetings, the social relationships among the participants of a meeting has been recognized as a fundamental aspect of the meetings' efficacy since the seminal work of Tang [30]. Many different attempts have been made to bring the social dynamics at a "visible" level. For example, Erickson and colleagues in [16] propose the idea of "social translucence", that is graphical widgets that signal cues that are socially salient. The claim is that such a functionality—by supporting mutual awareness and accountability—makes it easier for people to carry on coherent discussions; to observe and imitate others' actions; to engage in peer pressure; to create, notice, and conform to social conventions; and to engage in other forms of collective interaction.

What we propose in this paper is inspired by the works above but it has some characteristic aspects: (i) it is based on actual user studies to assess its acceptability; (ii) it uses multimodal fusion (that is, audio-visual scene analysis) combined with machine learning techniques to monitor the group relation behavior in face to face meeting and (iii) it uses shallow natural language generation techniques to provide an offline service to the group.

3. ACCEPTABILITY STUDIES

Since the coaching relationship is based upon trust and permission [22], we expected problems regarding acceptability to arise. The first step of our research was on assessing the willingness of meeting participants to accept working in an augmented environment whose task is not merely to provide technological support but also that of deliberately keeping trace of their behavior and understanding their role in the team work. In this section we present two studies aimed at understanding the way people would react to the very idea of the automatic coaching.

3.1 Initial Study: Focus Groups

Three focus groups have been organized: the first and second one consisted of five researchers of ITC-irst not involved in the project, while the third had four participants from the clerical staff. The facilitator was the same in all the focus groups. Each focus group was structured in three phases. During the first one the facilitator introduced the general topic (the relational report) and the rules of the discussion. The relational report was explained by first presenting two videos drawn from one of our recorded meetings, followed by a multimodal relational report addressing one of the participants in that meeting, and constructed according to the principles explained above. After this, the second phase started, which was devoted to discussing four specific issues, one at a time. The facilitator introduced each issue by asking a specific question that the focus group would then discuss. The allowed time was approx. 10 minutes per issue. During the third phase the facilitator presented a brief summary of the discussion for the group to briefly discuss.

The issues investigated were the *perceived usefulness* ("what do you think about the usefulness of a report such as the one you have seen? –would you prefer a descriptive or a normative report?), the *reliability* of the report ("what do you think about the reliability of the report?'"), its *intrusiveness* ("What are your opinions about the possible intrusiveness of the report and of the equipment it needs?"), and its *acceptability* ("what do you think about the acceptability of the report? – does it change according to whether the feedback is positive or negative?"). The facilitator never intervened during the discussion, except when needed to keep the discussion to its topic, or explain the question. All the focus groups were video recorded. The facilitator used these recordings to compile a summary after the end of each focus group.

Concerning the *perceived usefulness* of the report, the consensus in all the focus groups was that it could be useful, though the utility was seen as dependent on the disposition of each addressee to consider criticisms. Two participants in the expert groups rejected the usefulness of the feedback, one of them motivating his position with the idea that people are already aware of their own behavior, the other one arguing that behavior is not an important aspect for the meeting success. In both cases, this negative attitude was at least partially determined by the lack of trust towards the computer reliability. One participant in the non-expert focus group, on the other hand, considered the possibility of showing the report to a supervisor, in order to obtain a kind of formative counsel.

As to the *reliability,* it was widely agreed by all groups that the report was reliable. Nevertheless, almost everyone pointed out the need for more audio-video evidence for the statements made. In details, the expert participants suggested adding, both quantitative (e.g., statistics on turn taking, time spent on talking, overlapping speech, etc.) and qualitative information, also suggesting that the report should take into account the official (organizational) role of the addressee. The administratives' focus group emphasized the fact that the report must be "an objective synthesis of the behavior exhibited during the meeting" while lamenting that this goal can be hindered by the incapability of the system to contextualize people's behavior, this possibly leading to inaccurate or wrong reports. Here contextualization does not refer to the immediate context of the interaction, but to the long term relational history of the group and of the individuals composing it, including their official roles and position within the organization.

Regarding *intrusiveness*, being video recorded was acknowledged to be more intrusive and annoying than the individual delivery of the behavioral feedback. There are differences between the expert and non-expert focus groups, though. The latter could not explain precisely why they were annoyed by the video recordings, maintaining that the embarrassment is automatic and not controllable ("it is the very idea of an "eye watching you" that is annoying; this is intrusive by definition"). The experts, on the other hand, explicitly linked their attitudes to privacy problems: people fear that the video recording could be used in an unfair way. Interestingly, these negative feelings seemed to be triggered more by the visual part of the recordings than by the audio one.

The *acceptability* of the report turned out to depend mainly on the trust in the system, and on the subjective disposition and motivations to accept external feedbacks. Thus, the expert groups agreed that the possibility of choosing whether to receive or not the report improves the acceptability (for sure, if you ask for the report, you trust in it). As with usefulness, the acceptability of the report is expected to depend on the quantity and quality of the evidence (quantitative or descriptive) the report comes with; as many put it, this information allows addressees to control the factual basis of the report. The motivations and the cost-benefit balance turned out to be an important aspect determining acceptability (and usefulness). Thus many mentioned, again, the greater acceptability (and utility) the relational report can have for people involved in a formative path. Finally, many lamented the lack of the interaction that a human coach makes available: the possibility of explaining the reasons of the behavior and discussing them would improve the acceptability of the relational feedback.

The two kinds of subjects we used in our focus groups did indeed differ along the computer trust dimension. Whereas the experts confirmed their skepticism about the possibility for a computer to do the job we illustrated by means of our mock-up, this did not appear to be a major issue for the non-experts. On the contrary, these were much more interested in finding the right place to the relational reports in their own working experience and environments. Non-experts admitted that reports could be used to improve individual skills, but they also discussed the possibility of making it available to office managers and heads, to allow them monitoring the relational skills of their people in a better way. In a way, our non-experts seemed to be keen to assimilating the system producing the relational report to one of the official authorities they are used to deal with. This attitude has a number of possible consequences that clearly emerged in the course of the discussion: in the first place, the acceptability of the system as an authority is bound to its being objective; thence the worry that the system considers the extended group and personal context. Secondly, the level of concern for privacy issues, and the felt intrusiveness are lower, given that the cost of being monitored is balanced by the attempt at finding the relational report the proper place in their environment. Finally, it motivates a more constructive approach, manifesting itself, e.g., in a discussion about the means that are more appropriate to convey the report. Hence, some of the non-experts observed that the synthetic face, with its facial expressions, introduce an evaluative aspect that can have an important impact on acceptability. This was widely agreed upon by all the other members of the focus group; the discussion continued addressing the best ways to present the report, and much consensus gained the idea of using only text, in a way, a very objective and aseptic mean.

As anticipated, the (dis)trust factors colored many of the statements of the expert people: they not only do not believe that the machine will (ever) be able to monitor human behavior and meaningfully report about it. They were clearly much worried that this might be the case, appearing much concerned with privacy issues and with the potential intrusion in very delicate issues. Very few were the attempts at finding a place to the system in their environments. Finally, let us observe that besides being motivated by the higher awareness on the limits, defects and advantages of the technology, this general attitude could be related to a more individualistic conception of the work (most of the experts were researchers).

3.2 The Wizard-of-Oz experiment

In order to compare the acceptability of automatically produced relational reports with respect to similar reports produced by a human coach, we organized a Wizard-of-Oz experiment where eleven group of four people where requested to enter in a structured discussion of about half an hour.

In order to provide for as much a uniform context as possible, our groups were engaged in the solution of one of two versions of the Survival Task, a task often used in experimental and social psychology to elicit decision-making processes in small groups [18]. The exercise consists in promoting group discussion by asking participants to reach a consensus on how to survive in a disaster scenario, like moon landing or a plane crashing in Canadian mountains. The group has to rank a number (15 in our case) of items according to their importance for crew members to survive. Consensus decision making scenario was chosen for the purpose of meeting dynamics analysis because of the intensive engagement it requests to groups in order to reach a mutual agreement, this way offering the possibility to observe a large set of social dynamics and attitudes.

The average duration was 25 minutes. All the participants (40% males and 60% females) were clerical people working at ITC-irst. In all cases they knew each other, and had often been involved in common group activities in the past. The average age was 35 years.

Few days after the session, the participants received an individual report elaborated by a social psychologist. Each report described the behaviour of the participant in terms of the functional roles played during the meeting. The psychologist took the roles of the coding scheme (see section 4 below) as a reference, adopting a descriptive style and without mentioning explicitly the role labels (for example, the Orienteer/Protagonist label was paraphrased as *"[…] She initiates the discussion by proposing an importance order, justifying it and using a quiet tone of voice […]"*). In writing the reports, the psychologist considered only behavioural aspects of the participants, such as the posture and the tone of voice, and not aspects related to content such as the individual contributions to the discussion.

Half of the participants were told that their report was automatically elaborated by an intelligent system able to monitor the groups' behaviour, while the other half (i.e. the control group) were told that the report was written by a psychologist.

The attitude toward the report was tested by a seven item questionnaire aimed at assessing the perceived usefulness, its reliability, the perceived degree of intrusiveness and its acceptability. A semantic differential aimed at assessing the appropriateness, the completeness and the clarity of the report (the semantic differential was part of the 6-scale questionnaire proposed by Garrison [17] with a Cronbach alpha of 0.9482) was also used.

The answers to the questionnaire were analyzed with a two-tailed multivariate ANOVA (p=.05), considering 42 questionnaires: half for the "expert source" of the report and half for the (pretended) "system source" of the report. The independent variable was the source of the report, in order to monitor how it affects the dimensions investigated in the questionnaires items (dependent variables). Generally, there were no statistically significant differences among the questionnaire's responses in the two groups. The subjects' attitudes were more positive toward the system source (though not is a significant way) in regard to: (i) the perceived usefulness of the report for improving their own relational behaviour (F(1,40)= .366), (ii) the perceived usefulness for improving interactions in meetings (F(1,40)= .143), and (iii) the willingness to remember the report observations (F(1,40)= .175). Vice versa, the subjects' attitudes were more positive toward the human expert in regard to: (i) the usefulness of the report for stimulating reflection on behavioural aspects not considered in the past (F(1,40)= 2.138), (ii) the completeness of the report in catching relevant behavioural aspects (F(1,40)= .293 for caught and F(1,40)=.518 for not caught aspects), and (iii) the analysis ability of the expert (F(1,40)= 1.675).

Regarding the subscales of the semantic differential, they were analyzed with a two-tailed multivariate ANOVA with p=.05. In this analysis, a more positive attitude toward the human expert emerges, but not at a statistical significant level. The more relevant difference emerged in the appropriateness sub-scale ($F_{(1,40)}$= 4,007, p=.05, only marginally significant), less evident in completeness ($F_{(1,40)}$= 2.079) and clarity ($F_{(1,40)}$= .901) sub-scales.

The results of the two studies might suggest that automatic coaching can acceptable, at least for clearical people, as much as human coaching.

4. THE FUNCTIONAL ROLE CODING SCHEME

The goal of presenting individual profiles to participants suggested that we consider those approaches to social dynamics that focus on the roles members play inside the group, as opposed to approaches that define roles according to the social expectations associated with a given position (as in [20]).

This kind of roles—called *functional roles* [29]— are defined in terms of the behaviour enacted in a particular context and allow exploiting information about what actually happened in the course of the interaction, while reducing the necessity for knowledge about the group' structure, history, position in the organization, etc.

Benne and Sheats [7] provided a list of functional roles for working groups, and collected them into three classes: task-oriented, maintenance-oriented and individual-oriented roles. The first two types of roles are directed toward the group's needs: task-oriented roles provide facilitation and coordination in view of task accomplishment, while maintenance roles contribute to social structure and interpersonal relations in order to reduce tensions and maintain smooth group functioning. The third type of roles, the "individual roles", focuses on the individual and his/her goals and needs rather than the group's. During the interaction, each person can enact more than one role.

Drawing on Benne and Sheats, Bales [2] proposed the Interaction Process Analysis—IPA, a framework to study small groups by classifying individual behaviour in a two-dimensional role space consisting of a Task and of a Social-Emotional area. The roles pertaining to the latter stem from activities that support or weaken interpersonal relationships. For example, complimenting another person is a positive socio-emotional behaviour in that it increases group cohesion and mutual trust; insulting another participant, on the other hand, can undermine social relationships. The other six categories pertain to task-oriented activities, that is, behavioural manifestations relating to management and solution of the problem(s) the group is addressing. Giving and asking for information, opinions, and suggestions related to the problem at hand are examples of task-oriented activities.

Building on Benne and Sheats's functional roles and on Bales' two dimensional approach, and drawing on observations performed on a set of face-to-face meetings, a coding scheme was produced—the Functional Role Coding Scheme (FRCS)—consisting of five labels for the Task Area and six labels for the Socio-Emotional one.

The Task Area consists of roles relating to the facilitation and coordination of the tasks the group is involved in, as well as to the technical skills of the members as they are deployed in the course of the meeting. It includes: the *Orienteer*, who orients the group by introducing the agenda and defining goals and procedures. He/she keeps the group focused and on track and summarizes the most important parts of the discussion and the group's decisions. The *Giver* provides factual information and answers to questions. The *Seeker* requests information. The *Recorder* manages the available resources for the sake of the group. The *Follower* merely listens, and does not actively participate to the interaction.

The Socio-Emotional Area concerns the relationships between group members and roles oriented toward the functioning of the group as a group. It includes: the *Attacker,* who deflates the status of others or expresses disapproval; the *Gate-keeper*, who acts as the moderator within the group; the *Protagonist*, who takes the floor; the *Supporter*, who shows a cooperative attitude; the *Neutral*, who passively accepts the idea of others, serving as an audience in group discussion.

Of course, participants may—and often do—play different roles during the meeting, but at a given time each of them plays exactly one role in the Task Area and one role in the Socio-Emotional one.

The reliability of the scheme was assessed on a corpus consisting of 130 minutes for the Socio-Emotional Area and 126 minutes for the Task Area taken from internal meetings of researcher at our institute (of course, none of the researchers was involved in the project). Two trained annotators coded five participants on the Socio-Emotional Area and five in the Task Area. Cohen's κ was used to assess inter-annotator agreement. The agreement on the roles of the Task Area is good, κ=.71, (N=758, SE=.02, p<.0001), with confidence interval 0.65 – 0.75 (α=.05). The agreement on the roles of Socio-Emotional Area is less high: κ=0.6 (N=783, SE=0.023, p<.0001), with confidence interval 0.56-0.65 (α=.05).

Following Landis and Koch [25], the agreement on the roles of the Task Area is good ($0.6 < k < 0.8$) while the agreement on the roles of Socio-Emotional Area is on the borderline between being good and moderate ($0.4 < k < 0.6$).

5. THE SURVIVAL CORPUS

The multimodal corpus used to automatically code participants' behavior (see section 6) is based on the audio and video recordings of the eleven meetings of the acceptability study described above. Each session was recorded in the specially-equipped room at ITC-irst (see Figure 1), by means of five Firewire—four placed on the four corners of the room and one directly above the table— and four web cameras installed on the walls surrounding the table.

Speech activity was recorded using four close-talk microphones, six tabletop microphones and seven T-shaped microphone arrays, each consisting of four omni directional microphones installed on the four walls in order to obtain an optimal coverage of the environment for speaker localization and tracking.

The following annotations were produced: functional relational roles (task roles and socio-emotional roles), speech activity; and fidgeting activity.

Speech activity refers to the detection of the presence/absence of human speech. Each session was segmented by first automatically labeling by means of VAD (voice activity detector) [11] the speech activity recorded by the close-talk microphones every 330ms. For each speaker, VAD identifies the amount of activity, and produces an output of the form ⟨*participant-code*, *start time*, *end time*, *label*⟩, where *label* takes on the values 'speech' and 'no-speech'. VAD's output was then manually checked to control for cross-talk, and for purifying the annotation from breaths, yawns, coughing, and noises caused by the subjects when touching the microphones.

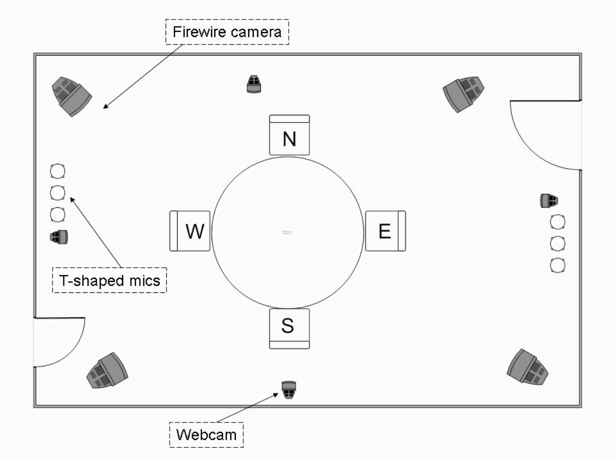


Figure 1. The experimental setting in the CHIL room

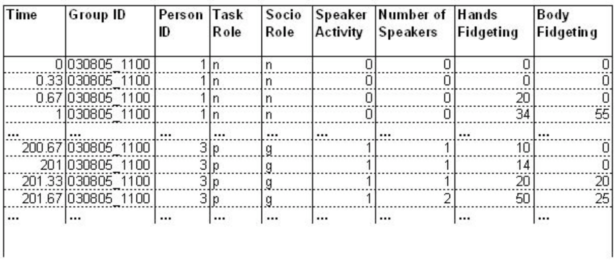
Fidgeting refers to localized repetitive motions such as when the hand remains stationary while the fingers are tapping on the table, or playing with glasses, etc. Fidgeting was automatically tracked by using skin region features, and an MHI of the convex skin polygons and temporal motion as the trigger is used [12]. The values of fidgeting for hands and body were extracted for each participant and normalized on the fidgeting activity of the person during the entire meeting.

Table 1. Distribution of the categories in the corpus (330ms time stamp)

Task Area Roles			Socio-Emotional Area Roles		
Neutral	71147	66.12%	*Neutral*	78427	72.88%
Orienteer	5458	5.07%	*Gate-Keeper*	0	0.00%
Giver	28214	26.22%	*Supporter*	9401	8.74%
Seeker	2789	2.59%	*Protagonist*	19487	18.11%
Recorder	0	0.00%	*Attacker*	293	0.27%
Total	**107608**		**Total**	**107608**	

Functional roles were manually annotated for each participant by considering the participants behavior every 5 seconds and then re-sampled every 330ms to align them with the other features. The corpus was quite unbalanced (see Table 1): *Follower* and *Neutral*—as expected—were the most frequent roles while *Attacker* was quite rare (the participants knew they were observed and perhaps they tended to avoid aggressive or uncooperative behavior). *Recorder* and *Gate-Keeper* roles were never observed.

The corpus consists of 107608 rows each reporting the speech activity of one of the participants during a 330ms interval, his/her hands and body fidgeting, the number of people speaking during that time, and the functional roles the person plays (see figure 2).



Time	Group ID	Person ID	Task Role	Socio Role	Speaker Activity	Number of Speakers	Hands Fidgeting	Body Fidgeting
0	030805_1100	1	n	n	0	0	0	0
0.33	030805_1100	1	n	n	0	0	0	0
0.67	030805_1100	1	n	n	0	0	20	0
1	030805_1100	1	n	n	0	0	34	55
...	...	...	...	...	...	...	...	...
200.67	030805_1100	3	p	g	1	1	10	0
201	030805_1100	3	p	g	1	1	14	0
201.33	030805_1100	3	p	g	1	1	20	20
201.67	030805_1100	3	p	g	1	2	50	25
...	...	...	...	...	...	...	...	...

Figure 2. The annotated Survival Corpus

6. AUTOMATIC DETECTION OF ROLES

In order to train a machine learning system to automatically code the relation behavior, the corpus was first of all reduced by considering only the cases corresponding to time intervals where the participant for whom the roles were to be classified was speaking. We employed sliding windows to take in account the time dimension during our analysis [14]: the classifier works on all the data comprised in the time window to assign a Task area role and a Socio-Emotional area one only at the end of the window. We considered windows of varying size, from 0 to 14 seconds (i.e. 42 rows), placed both to the left the relevant time point, and windows centered on the latter. Initial attempts showed that centered windows are less effective [32]; hence in this paper we report only on the results from left windows. For each window size we built a dataset by adding to each row all the features of the rows before included in the window width.

Therefore, for a given time and a given participant, the information that the classifiers had available to classify his/her roles was the information about his/her speech and fidgeting activity, as well as the number of simultaneous speakers and the information about the speaking activity and the fidgeting of all the other participants, during the window time. Each dataset was then split in two equal parts for training and testing.

We modeled role assignment as a multiclass-classification problem on a relative large and very unbalanced dataset, and used Support Vector Machines as classifier. The choice of these classifiers is due to their robustness with respect to over-fitting: the SVMs in fact find a boundary between instances of different classes such that the distance between this boundary and the nearest instances is maximized [13].

The bound-constrained SV classification algorithm with a RBF kernel $K(x,y) = \exp(-\gamma||x-y||^2)$ was used. The cost parameter C and the kernel parameter γ were estimated with the grid technique by cross-fold validation using a factor of 10[2]. Furthermore, the cost parameter C was weighted for each class with a factor inversely proportional to the class size. SVM were originally designed for binary classification but several methods have been proposed to construct multi-class classifier [19]. The “one-against-one” method [21] was used whereby each training vector is compared against two different classes by minimizing the error between the separating hyper-plane margins. Classification is then accomplished through a voting strategy whereby the class that most frequently won is selected.

The performance of the classification for the Task area roles is rather good. The highest accuracy is reached with window 27 (with a value of 0.90). The max value of macro F is 0.87 and is reached at the largest window size (42, 14 seconds). We prefer to consider window 42 since we value macro F as a better measure of accuracy on our corpus. Table 2 summarizes the precision and recall values for the Task Area roles on that window.

Table 2. Precisions and Recall values for Task roles at window width 42 (14 seconds)

	Neutral	*Supporter*	*Protagonist*	*Attacker*
Precision	0.89	0.89	0.91	0.83
Recall	0.92	0.81	0.91	0.74
F	0.91	0.85	0.91	0.78

The results are similar for the Socio area roles: the highest accuracy is reached with window 27 (0.92), while the max value of macro F is reached with window 42 (0.86). Table 3 summarizes the precision and recall values for the Socio Area roles on window 42.

Table 3. Precisions and Recall values for Socio roles at window width 42 (14 seconds)

	Follower	*Orienteer*	*Giver*	*Seeker*
Precision	0.84	0.93	0.93	0.89
Recall	0.90	0.87	0.91	0.68
F	0.87	0.90	0.92	0.77

Of course, for both Task and Socio roles, the worse classification results were on low-represented classes (the Seeker and the Attacker). Although, it is likely that distribution of roles will always be unbalanced, a different task for the groups’ activity in the future data collections may provide more examples of these classes.

Another direction for improvement is adding more features. We chose to use two simple features of the audio-visual scene: speakers’ activity and body energy since we aimed at setting a baseline before investigating a richer set of features yet in the next steps we plan to add a few more starting from vocal energy, 3D postures and focus of attention and to analyze which of these features have more impact for the automatic detection of group roles.

7. GENERATION OF THE MULTIMEDIA REPORTS

Finally, a generation component has been built to produce the profile reports from the behavior sequences (i.e. roles) extracted from the audio visual analysis as described in the previous section. The report is built according to the task and socio-emotional roles simultaneously assumed by the speaker to whom the report is addressed during the interaction as well as taking into account the roles played by the other participants.

The system is based on a strategic text planner which accesses a repository of declaratively defined discourse schemata which look for specific patterns of roles played by the participants and decide what to say and in which order.

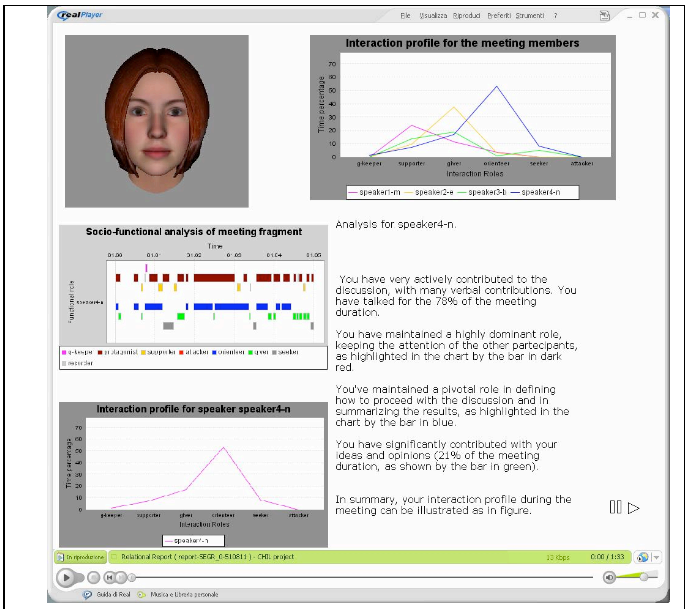
The linguistic realization of the sentences is currently template-based. Notwithstanding the limitations of the current, preliminary, implementation of the generator, a positive response to the technological feasibility of the system has emerged. An example of brief report actually automatically generated from real data is the following (see Figure 3).: “*You have actively contributed discussion, with many verbal contributions. You have talked for the 78% of the meeting duration. You have maintained a pivotal role in defining how to proceed with the discussion and in summarizing the results […]*”.

Each final report has the form of a multimedia presentation where a talking head reads the report (which is also reported in a written form to overcome the limitation of current speech synthesizers). Moreover, the presentation is enriched with short audio-video clips from the meeting that exemplify the information presented, and with graphics that intuitively display and compare the interaction profiles held by each participant during the meeting, according to well-know theories of small-group dynamics [3] [4].

The talking head uses emotional expressions to stress positive and negative behaviors. For example, “*You have actively contributed to the meeting.*” while showing a neutral expression, then, changing into a happy expression “*You’ve helped to focus the discussion on the relevant topics and have provided useful*

[2] We used the BSVM tool [19] available at http://www.csie.ntu.edu.tw/~cjlin/bsvm/.

information and opinions to clarify some of the issues." finally, moving to a sad expression she adds *"You have profitably cooperated with your colleagues, even though a contrast with R.Z. emerged during the first part of the meeting."*



Analysis for speaker4-n.

You have very actively contributed to the discussion, with many verbal contributions. You have talked for the 78% of the meeting duration.

You have maintained a highly dominant role, keeping the attention of the other partecipants, as highlighted in the chart by the bar in dark red.

You've maintained a pivotal role in defining how to proceed with the discussion and in summarizing the results, as highlighted in the chart by the bar in blue.

You have significantly contributed with your ideas and opinions (21% of the meeting duration, as shown by the bar in green).

In summary, your interaction profile during the meeting can be illustrated as in figure.

Relational Report (report-SEGR_0-510811) - CHIL project

Figure 3. An example of a multimedia relational report

Starting from the output of the SVM, a table is compiled for each participant showing the distribution of his/her roles. For each role, we took into consideration the total number of seconds the relevant participant played that role, as well as the role's distribution during the meeting (for the sake of simplicity, we considered the meeting as divided into quarters).

A number of patterns were then elicited and mapped to verbal statements that describe the behaviour. The pattern "high protagonist+orieenteer held for XY part of the meeting" is mapped onto the sentence "You have actively contributed to the XY part of the meeting". Similar patterns were then used to build the rest of the example above.

More elaborate strategies involve reasoning about the behaviour of various participants at a time. For example, should the participant for whom the report is prepared have maintained high percentages of "orienteer" role especially at the beginning and at the end of the meeting with the others being silent or neutral, the report could include a statement like "at the beginning of the meeting you have helped define the agenda and initiate the discussion, summing up the outcome of the meeting at the end". At the opposite, should the considered participant have maintained significant percentages of orienteer and seeker roles for the most part of the meeting, a statement as the following could be included: *"During the meeting you have played a leading role, defining discussion topics and soliciting your colleagues' participation."* In case sequences of seeker-(recorder)-attacker are observed, the report could be complemented with: *"in some cases, however, you have displayed a critical and aggressive behaviour as a response to your colleagues' contribution"*.

8. CONCLUSIONS

In this paper, we presented a multimodal service based on monitoring the audio visual scene of a meeting to produce a profile report that takes into consideration the relational behavior of the participants. In the context of a user-centred design, we first of all investigated the acceptability of such a service with a qualitative study (focus groups) and with a Wizard-of-Oz experiment. The results of these studies suggested that this service might be accepted by clerical users although technical people may be biased by a system that makes judgments. We then described the implementation of the two basic components to realize such a system, namely the component that automatically codes the behavior and the report generator.

The performance of the classification of the machine learning component is rather good yet the differences among the classes are not negligible and the worse classification results were, of course, on low-represented classes (the seeker and the attacker, in particular). Although, it is likely that distribution of roles will always be unbalanced, a different task for the groups' activity in the future data collections may provide more examples of these classes and hopefully bring to better classification.

Finally, although our user studies make as feel confident regarding the acceptability of the relational profile, the question of evaluation of the actual service is not trivial, in particular for what concern the impact of the different media (i.e text, talking head, emotions, ecc.). This aspect will be pursued as further work.

9. ACKNOWLEDGMENTS

This work has been partially supported by the CHIL project.

10. REFERENCES

[1] Andersen, C. A Theoretical Framework for Examining Peer Collaboration in Preservice Teacher Education. In Proceedings of the 2000 Annual *International* Conference of the Association for the Education of Teachers in Science. , January 6-9, Akron, Ohio, Jaunuary 6-9, 2000.

[2] Bales, R. F. Personality and interpersonal behavior. New York: Holt, Rinehart and Winston, 1970.

[3] Bales, R.F. Interaction process analysis : a method for the study of small groups, University of Chicago press, 1976

[4] Bales, R. F. and Cohen, S.P. Symlog : a system for the multiple level observation of groups, Collier Macmillan, London, 1979

[5] Banerjee, S. and Rudnicky, A. I. Using Simple Speech-Based Features to Detect the State of a Meeting and the Roles of the Meeting Participants. In Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech 2004-ICSLP), Jeju Island, Korea, October 2004.

[6] Barthelmess, P. and Ellis, C.A. The Neem Platform: an Evolvable Framework for Perceptual Collaborative Applications. In Proceedings of the International Conference on Cooperative Information Systems (CoopIs 2002), Irvine, CA, October 27- November 1, 2002.

[7] Benne, K. D., Sheats, P. Functional Roles of Group Members. In Journal of Social Issues 4, pp. 41-49, 1948.

[8] Bloom, G., Castagna, C. and Warren, B. More than mentors: Principal coaching. *Leadership*. May/June 2003.

[9] Boud, D., Keogh, R. and Walker, D. (eds.) Reflection: Turning experience into learning. Logan: Kogan Page, 1988.

[10] Brdiczka, O., Maisonnasse, J. and Reignier, P. Automatic Detection of Interaction Groups. In Proceedings of the 7th International Conference on Multimodal Interface. Trento, Italy, October 2005.

[11] Carli, G., Gretter, G. A Start-End Point Detection Algorithm for a Real-Time Acoustic Front-End based on DSP32C VME Board. In Proceedings of ICSPAT, Boston, USA. 1992.

[12] Chippendale P 7th International Conference on Automatic Face and Gesture Recognition - FG2006 (IEEE) Southampton, UK, April 2006.

[13] Cristianini N. and Shawe-Taylor J. Support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge, 2000.

[14] Dietterich T. G. Machine Learning for Sequential Data: A Review. In T. Caelli (ed.) Lectures Notes in Computer Science. Springer-Verlag, 2002.

[15] Doyle M. and Straus D. *How To Make Meetings Work*. The Berkley Publishing Group, New York, NY. 1993.

[16] Erickson T., Halverson C., Kellogg, W. A., Laff M., and Wolf T. Social translucence: Designing social infrastructures that make collective activity visible. Communications of the ACM- – Special Issue on Community, J. Preece (Ed.), 45 (4), pp.2002, 40–-44, 2002.

[17] Garrison, B. The perceived of Electronic Mail in Newspaper Newsgathering. In Proceedings of Communication Technology and Policy Division, Association for Educational in Journalism and Mass Communication Midwinter Conference, March 1, Boulder, Colorado, 2003.

[18] Hall, J. W., Watson, W. H. (1970) The Effects of a normative intervention on group decision-making performance. In *Human Relations*, 23(4), 299-317.

[19] Hsu C.-W. and Lin C.-J. A Comparison of Methods for Multi-Class Support Vector Machines. In *IEEE Transactions on Neural Networks*, 13, pp.415-425, 2002.

[20] Katz, D., and Kahn, R. L. The Social Psychology of Organizations (2nd ed.) New York: John Wiley, 1978.

[21] Kressel U. Pairwise classification and support vector machines. In B. Scholkopf, C. J. C. Burges and A. J. Smola (eds.) Advances in Kernel Methods – Support Vector Learning. MIT Press, Cambridge, MA. 1999.

[22] Kuhn D., Garcia-Milá M., Zohar A., and Andersen C. Strategies of knowledge acquisition. Monographs of the Society for Research in Child Development, 60, 4 (Serial No. 245),1995.

[23] McCowan I., Bengio S., Gatica-Perez D., Lathoud G., Barnard M., and Zhang D. Automatic Analysis of Multimodal Group Actions in Meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI), 27 (3), pp. 305-317, 2005.

[24] Oviatt, S. Multimodal Interfaces. In Handbook of Human-Computer Interaction, (ed. by J. Jacko & A. Sears), Lawrence Erlbaum: New Jersey, 2002.

[25] Landis JR, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.

[26] Pianesi, F. Zancanaro M., Falcon V., Not E. Toward Supporting Group Dynamics. In Proceedings of Artificial Intelligence Applications and Innovations. Athens, Greece. June 2006.

[27] Rienks R. and Heylen D. Dominance Detection in Meetings Using Easily Obtainable Features. In Revised Selected Papers of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms. Edinbourgh, Scotland, October 2006.

[28] Rienks R., Zhang D., Gatica Perez D., and Post W. Detection and Application of Influence Rankings in Small Group Meetings. In Proceedings Int. Conf. on Multimodal Interfaces (ICMI), Banff, Nov. 2006.

[29] Salazar, A. An Analysis of the Development and Evolution of Roles in the Small Group. Small Group Research, 27, 4, pp. 475-503, 1996.

[30] Tang, J. C. Findings from observational studies of collaborative work. *International Journal of Man-Machine Studies*, 34, pp. (1991), 143-160, 1996.

[31] Yang Y. A Study on Thresholding Strategies for Text Categorization. Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01), pp 137-145, 2001.

[32] Zancanaro M., Lepri B., Pianesi F. Automatic Detection of Group Functional Roles in Face to Face Interactions. In *Proceedings of International Conference of Multimodal Interfaces ICMI-06,* 2006.

Physiological Correlates of Emotions

Astrid Oehme[1] and Antje Herbon[2] and Stefan Kupschick[1] and Eric Zentsch[1]

Abstract

Recent sensor development enables wireless capture of context information, i.e., body and environmental data, in an unobtrusive way. Collected data is provided to mobile systems, which in turn respond intelligently, providing meaningful services to the user. Within the European project e-SENSE, the affective state of the users is one component of context capture. To develop algorithms for emotion inference, an experiment was conducted in which five different emotional states were induced. Heart rate, electrodermal activity, breathing rate, and skin temperature were utilized to measure the respective emotional states. Self-assessment ratings were applied for manipulation check and comparison with the collected physiological data. Results show that at least three of the four measures seem promising for detecting differences in affective states and support a dimensional model of affect.

1 INTRODUCTION

1.1 Project Base

The European Integrated Project e-SENSE[1] (Capturing Ambient Intelligence for Mobile Communications through Wireless Sensor Networks) aims at enabling Ambient Intelligence in “Beyond 3G Systems” using wireless sensor networks (WSN) for providing context-rich information to applications and services.
Within e-SENSE three application spaces and themes have been defined that depict the usage of the e-SENSE concept in various situations. These aim at personal life (Personal application space), the community of professional users (Community application space), and industrial applications (Industrial application space). One focus of especially the Personal application space is the measurement of the users’ affective states. Based on emotions, intelligent applications will respond meaningfully, e.g., by offering information, comforting the user or even helping to relive excitement during previously undertaken activities (c.f. [1]). Figure 1 shows one such scenario aiming at personal security. The collection of voice, breathing rate, heart rate, noise, and position is symbolized. The woman depicted is riding a public transport vehicle all by herself (assessed by ambient sensors) when a dangerous person enters and begins approaching her. Fear is inferred from Body-Sensor-Network data and an alert is triggered to inform security staff members at the next stop ([2]).
The experiment reported in the following is a first step on inferring emotional states of potential users of intelligent applications.

Figure 1. Danger Warning Scenario selected from e-SENSE D1.2.1 ([2] p.13)

1.2 Emotion Modeling

Among the theories for categorizing or structuring emotions, two approaches have been widely accepted. The discrete or categorical approach claims the existence of a set of universal, ‘basic emotions’ (e.g., [3,4]) that can be distinguished clearly from one another and form the basis for all other emotions we might experience. Studies performed in search of physiological patterns specific to basic emotions concentrated mainly on activities of the autonomous nervous system (ANS) and characteristic speech signal changes. ANS-related studies (e.g., [5,6,7,8,9,10] and many others) showed very interesting results each on its own, but until now no distinct fixed patterns for the proposed six basic emotions could be found. The results of the studies are controversial and the variables measured do not seem to allow a clear distinction between different emotions.
The other approach proposes two or more major dimensions, which enable the description of different emotions and the distinction between them (e.g., [11]). According to the dimensional view, emotions are mainly characterized by their valence and arousal[2].

[1] HFC Human-Factors-Consult GmbH, Berlin
Koepenicker Str. 325, D-12555 Berlin
[Oehme/Kupschick/Zentsch]@human-factors-consult.de

[2] Technical University Berlin, Zentrum Mensch-Maschine-Systeme
Franklinstraße 28-29, Sekretariat FR 2-7/2, D-10587 Berlin
antje.herbon@zmms.tu-berlin.de

[3] www.e-SENSE.org; e-SENSE is an Integrated Project (IP) supported by the European 6th Framework Programme. The study is a result of a collaboration between Human-Factors-Consult HFC, Berlin and the University of Surrey UniS, Guildford.

[2] While the terms valence and arousal are used here, the exact titling of the two dimensions has been very controversial. C.f. Feldman Barrett and Russell [25] for a discussion on the topic.

Valence is defined by its two poles negative/bad and positive/good, whereas the arousal dimension spans between the two poles sleepy/calm for very low arousal and aroused/excited for very high arousal. Valence and arousal have proven to be the two main dimensions, accounting for most of the variance observed [12]. Cowie et al. [13] proposed the application of additional dimensions for emotions that share the same degrees of arousal and valence, but are perfectly distinguishable in everyday life. For fear and anger a dominance or control dimension would support the distinction between the two emotions. For psycho-physiological studies the dimensional model has a high face validity, since physiological data is continuous and should correspond well to the dimensions proposed. The most commonly used physiological parameters applied in studies based on the dimensional model are skin conductance level (SCL), facial electromyogram (EMG) and heart rate (HR) (e.g., [14,15,16,17,18]), but speech parameters have also been examined (e.g., [18,20,21,22]). Lang [23] found linear increases of Galvanic skin response as an indicator of SCL with the level of overall arousal. Burch und Greiner [24] predict the same for electrodermal responses.

For the emotion inference study described in this paper, SCL, Electrocardiogram (heart rate), Breathing Rate (BR) and Skin temperature were chosen. Emotions were induced using short pre-selected and validated films. The goal of the study was to find correlates of the induced emotions in the physiological data and their corresponding subjective placement in the dimensional model.

2 METHOD

2.1 Participants

The 40 participants (27 male) were recruited from the Center for Communication Systems Research CCSR and Surrey's School of Management. They took part in the experiment voluntarily and were not rewarded. The sample was aged 22 to 54 years with a mean age of 30.1 years. The sample was culturally diverse and included 16 different nationalities.

2.2 Subjective Ratings of Emotions

The self-assessment derived from the dimensional model consisted of three manikin scales (Self Assessment Manikin, SAM) representing the proposed dimensions arousal, valence, and dominance [26]. After each emotion induction phase the participants were asked to fill in each of the three scales. They were first to indicate their level of valence by saying "I felt like …" and inserting the letter that was written underneath. Then the second scale (arousal) and third scale (dominance) appeared on the screen respectively. Figure 2 depicts the three SAM scales.

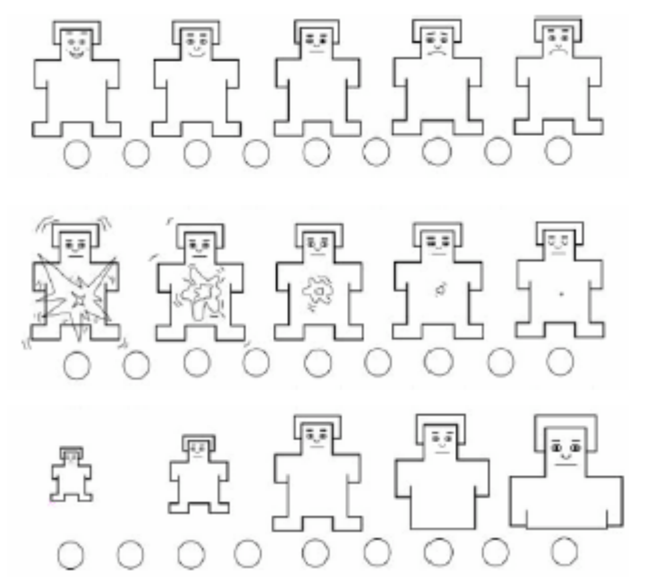


Figure 2. SAM-Scales for valence (top), arousal (middle) and dominance (bottom)

2.3 Physiological Measurement of Emotional Changes

SCL, ECG, BR, and Skin temperature were collected with components of the HealthLab System (Koralewski Industrie-Elektronik oHG). EDA was taken on the palm of the participant's left hand, three ECG electrodes were placed left and right approximately in the 5th intercostal space on the median axillary line, a chest belt measured respiration and the skin temperature was derived from the participant's left index finger.

Data was transmitted from the sensors to a master satellite and then sent via Bluetooth from each master satellite to a laptop, which was equipped with special software (Heally Control, Koralewski oHG).

2.4 Emotion Induction

Five films were constructed to induce emotions in all four quadrants of the coordinate system spanned by the dimensional model (see Figure 3). During the construction phase, the films were pre-tested in two trials of 5 subjects each at HFC, Berlin and a third trial of seven subjects at UniS, Guildford.

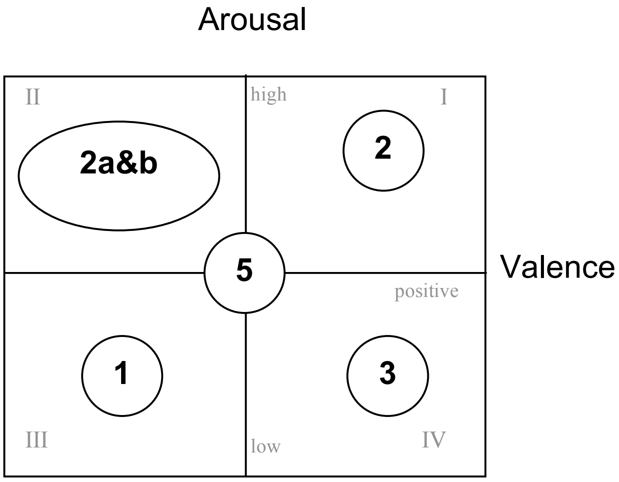


Figure 3. Films in all four quadrants of the coordinate system. Note that number 5 indicates a 2.5-minute-baseline-phase before the beginning of the first movie during which all physiological parameters were recorded.

Film 1 was to induce a positive arousing emotion, i.e. something similar to happiness. For this purpose, a short, supposedly funny cartoon was chosen. Pretests proved the validity of this film. For evaluation, three dimensional ratings (degrees of valence, arousal and dominance) were collected from the participants. As expected, film 1was mainly rated positive and arousing (Figure 4).

There are two emotions that – if rated according to their valence and arousal – are both located in quadrant 2, which are anger and fear. To distinguish between the two emotions, the additional dominance dimension with the poles strong and weak has been included. Anger is usually supposed to be of high, while fear is supposed to be of low dominance. Most of the effort in the material construction phase was put into the film that was to induce anger. As far as dimensional ratings are concerned, the film can be considered valid (see Figure 5).

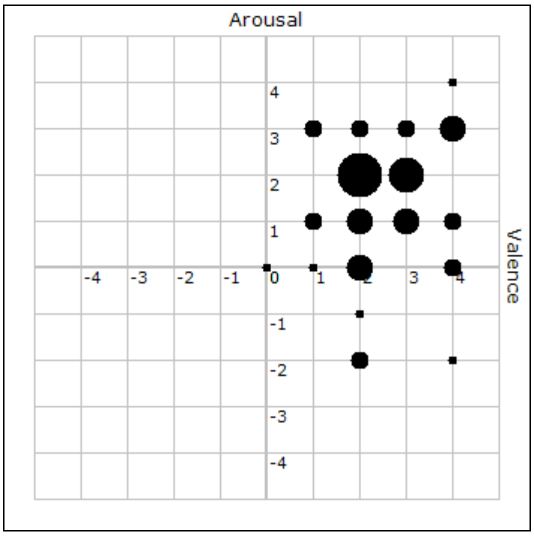


Figure 4. Film 1: dimensional ratings, target: happiness

Emotional self-assessment of Film 2a, which aimed at anger, was mainly located in the second quadrant, which was the target quadrant of this emotion.

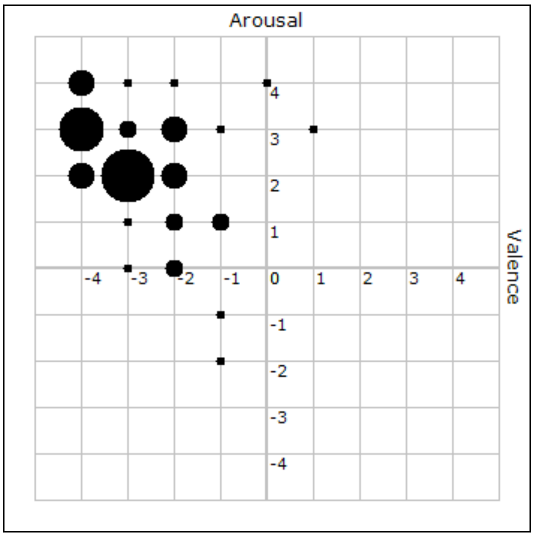


Figure 5. Film 2a: dimensional ratings, target: anger

Ratings of film 2b were mostly situated in the second sector, which was the respective target sector.

Film 3 ratings showed a diverse pattern: The resulting emotion was rated as negative, as was targeted, but ratings were located in two quadrants, namely quadrant two and quadrant three, while only quadrant three was targeted. Thus, about half of the subjects rated the emotion that resulted from Film 3 as arousing and the other half as calming.

Ratings for Film 4, which aimed at contentment, were also distributed over two quadrants, namely quadrants one and four, while quadrant four was targeted.

2.5 Setting and Procedure

The experiment was performed at the I-Lab of the University of Surrey in Guildford, England over a period of five days. The subjects sat in a separated 3.2mx4.5m windowless test room in 2m distance of a 2.4mx1.35m screen (projector solution: 1920x1080 pixels). Stereo 2.1-sound was provided by two front speakers. The investigators observed the trial via a glass pane from a separate room. Microphones ensured that test-leaders and subjects could talk to each other in case of any problem. Subjects were welcomed to the I-Lab, received a written instruction sheet about the experiment and were equipped with the HealthLab System. The rating scales were explained to them once they had taken a seat in the laboratory they were asked to sign a consent form for participation and filled in a first demographical questionnaire. The subject was then left alone and a 2-minute-baseline was recorded. Films were presented in randomized order to avoid sequence effects. Rating scales appeared on the screen after the end of each of the films in the same order: valence, arousal, dominance. Once the rating had been completed, the light was switched on again and a 1.5-minute break followed.

3 RESULTS

3.1 Correlates of Self-Assessment and Physiology

SAM-ratings were correlated with physiological measures to identify evidence for an underlying two- or three-dimensional model. Table 1 shows the results of this analysis[3]. Note that for this correlation over all films no coherences between dominance and any of the physiological parameters were found, whereas this dimension correlated quite substantially with valence. Skin temperature did not correlate significantly with any of the dimensions.

Table 1. Significant overall correlations for SAM-Ratings and physiological parameters.

	Valence	Arousal	HR	BR	SRL
Valence	1				
Arousal	-.200	1			
Heart Rate (HR)	-.252		1		
Breathing Rate (BR)		.196	-.160	1	
Skin Resistance Level (SRL)		-.339		-.245	1
Dominance	.591	-.143			

In a second step, correlation analyses were conducted for all films separately to identify coherences that might be specific to certain emotional states.

Inter-dimensional correlations increased when films 1 and 2a were analyzed separately. Arousal and dominance showed correlation coefficients of up to -.400. The analysis of film 2a additionally led to the identification of stronger coherences between all three dimensions and breathing rate and between dominance and skin temperature (see Table 2).

Table 2. Significant correlations for Film 2a

	Valence	Arousal	Dominance	BR	ST
Valence	1				
Arousal		1			
Dominance	.586	-.400	1		
Breathing Rate (BR)	-.346	.383	-.416	1	
Skin Temperature (ST)			-.328		1

The separate consideration of film 3 resulted in the finding of higher coherences between the dimension of arousal and skin resistance level with a correlation coefficient of -.409. When film 4 and 5 were analyzed separately, all significant coherences vanished.

Regression analysis was conducted to investigate substantial inter-dimensional correlations further. Valence and arousal account for 35% of the observed variance in the dominance dimension, while valence and arousal accounted for 4% of each others variance.

3.2 Analysis based on Stimulus Films

Repeated-measures ANOVAs were conducted for the respective physiological parameters. The ANOVAs confirmed significant main effects for the factor “film” in all parameters except temperature (BR: F(4,36)= 11,4; HR: F(4,36)=8,9; SCL: F(4,36)=16,9). A significance level of p<.10 was used for the computations, because error probability was considered less critical than rejecting assumed differences in physiological data based on the film perception.

Post-hoc single comparisons between pairs of films were conducted for the remaining three physiological parameters.

The following tables (**Table 3Table 4**, and**Table 5**) illustrate the results of these computations. The first column “Pair of films” shows which films were compared. The second column “Mean” is calculated as the difference between the mean values of the physiological parameter in both films. All comparisons cited in the tables are significant with p<.10.

Table 3. Paired Samples Tests for Breathing Rate.

Pair of films	Mean	T
Fear – Contentment	0,68	3,592
Fear – Sadness	0,71	4,498
Anger – Contentment	0,57	2,956
Anger – Sadness	0,59	2,878
Contentment – Happiness	-1,11	-4,802
Sadness – Happiness	-1,11	-5,596

Table 4. Paired Samples Test for factor Skin-Conductance Level

Pair of films	Mean	T
Fear – Contentment	-1,23	-5,757
Fear – Sadness	-1,23	-5,365
Fear - Anger	-1,11	-6,908
Fear - Happiness	-0,33	-2,378
Anger – Happiness	0,78	5,203
Contentment – Happiness	0,89	5,212
Sadness – Happiness	0,76	3,756

Table 5. Paired Samples Test for factor Heart Rate

Pair of films	Mean	T
Fear – Sadness	-0,69	-3,801
Fear - Anger	-0,85	-4,193
Anger - Contentment	0,52	2,713
Anger – Happiness	1,00	5,127
Contentment – Happiness	0,48	2,462
Sadness – Happiness	0,85	3,573

4 DISCUSSION

Correlations between arousal and SCL found in the past (e.g., [18]) could be replicated.

An interesting finding of this first analysis was a moderate correlation between dominance and valence that was supported further with a moderate prediction of dominance variance by valence and arousal in the regression analysis. Although this implicates that it might not be necessary to consider dominance in future studies, it cannot be answered clearly at this point if correlations between physiological parameters (i.e., BR and skin temperature) that were observed during film-specific analyses can

[3] Data was significant on a .05-level.

be explained by valence-dominance and arousal-dominance correlations alone.

Independently from self-reported affective states, results based on the film comparison show significant differences for breathing rate, heart rate, and skin-conductance. Differences in breathing rate were found between higher and lower arousing films. It can be assumed that, e.g., the parameter breathing rate could be a good indicator for different levels of arousal. This also corresponds with the correlation results regarding the subjective data. SCL could be used for the identification of fear-situations, because significant differences to all other films appear. The correlations of HR and valence as well as the larger differences between films corresponding to positive and negative situations suggest this parameter for valence inference.

Summarizing the results, at least three physiological measures can separate between very different emotional states and will therefore be included in future studies.

5 FUTURE PROSPECTS

This study was in line with previous ones concerning the identification of specific physiological changes for certain emotions. Significant differences were found.

Especially further investigation on different parameters of electrodermal activity seems worthwhile to reach a larger granularity regarding affective state differences [27,28]. Emotion-based scenarios like excitement revival that rely mainly on arousal correlates seem less problematic concerning data integration and overall severity of the user's situation. Especially for safety-related applications as depicted above, stronger coherences of physiological data and affective states have to be reached to prevent systems from sending warning signals in non-dangerous situations, i.e., to produce false alarms. Happiness, e.g., is similar to fear as far as excitement is concerned and anger is even similar in both valence and arousal. However, situations in which these affective states occur differ from one another a lot and should not be misinterpreted by an application, since products without sufficient detection-accuracy of inference algorithms will never find their way into the market.

Future research can not rely on single parameters, but will instead have to follow a multimodality-approach, combining different physiological parameters and including additional non-physiological parameters like content information as investigated further within the e-SENSE project.

6 REFERENCES

[1] Forest, F., Oehme, A., Yaici, K. & Verchère-Morice, C. (2006). *Psycho-Social Aspects of Context Awareness in Ambient Intelligent Mobile Systems.* 15th IST Mobile & Wireless Communication Summit, Myconos, http://www.ist-esense.org/index.php?id=149

[2] Oehme, A., Yaici, K., Forest, F. & Bourbiaux, G. (2006). D1.2.1 – Scenarios and Audio Visual Concepts. Deliverable Report, http://www.ist-esense.org/index.php?id=33

[3] Plutchik, R. (1980). A general psychoevolutionary theory of emotion In: Plutchik, R., Kellerman, H. (Eds.), *Emotion: Theory, Research, and Experience: vol. 1. Theories of Emotion* (pp. 3–33). New York: Academic Press.

[4] Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion 6* (3/4).

[5] Ax, A. (1953). The physiological differentiation between fear and anger in humans. *Psychosomatic Medicine, 55 (5)*, 433–442.

[6] Ekman, P., Levenson, R.W., & Friesen, W. (1983). Autonomic nervous system activity distinguishes among emotions. *Science 221.*

[7] Palomba, D. & Stegagno, L. (1993). Physiology, perceived emotion and & Huber Publishers.

[8] Palomba, D., Sarlo, M., Agrilli, A., Mini, A., Stegagno, L. (1999). Cardiac response associated with affective processing of unpleasant film stimuli. International Journal of Psychophysiology 36, 45–57.

[9] Fredrickson, B.L., Mancuso, R.A., Branigan, C., & Tugade, M.M. (2000). The undoing effect of positive emotions. *Motivation and Emotion, 24 (4)*, 237–257.

[10] Christie, I.C. (2002). *Multivariate Discrimination of Emotion-Specific Autonomic Nervous System Activity*. MSc Thesis, Virginia Polytechnic Institute and State University.

[11] Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161–1178.

[12] Russell, J.A. (1983). Pancultural aspects of the human conceptual organization of emotions. *Journal of Personality and Social Psychology, 45 (6)*, 1281–1288.

[13] Cowie, R., Douglas-Cowie, E., Apolloni, B., Taylor, J., Romano, A., & Fellenz, W. (1999). What a neural net needs to know about emotion words. *CSCC'99 Proceedings*, 5311–5316.

[14] Bradley, M., Greenwald, M.K., & Hamm, A.O. (1993). Affective picture processing. In: Birbaumer, N., Öhman, A. (Eds.), *The Structure of Emotion* (pp. 48-65), Toronto: Hogrefe & Huber Publishers.

[15] Detenber, B.H., Simons, R.F., & Bennett, G.G. (1998). Roll 'em!: the effects of picture motion on emotional responses. *Journal of Broadcasting and Electronic Media, 21*, 112–126.

[16] Anttonen, J., & Surakka, V. (2005). Emotions and heart rate while sitting on a chair. *CHI'05 Conference Proceedings*. ACM Press, New York, pp. 491–499.

[17] Branco, P., Firth, P., Encarnacao, L.M., & Bonato, P. (2005). Faces of emotion in human–computer interaction. *CHI'05 Extended Abstracts*, 1236–1239.

[18] Herbon, A., Peter, C., Markert, L., van der Meer, E. & Voskamp, J. (2005). Emotion studies in HCI—a new approach. *Proceedings of the 2005 HCI International Conference, Las Vegas*.

[19] Pereira, C. (2000). Dimensions of emotional meaning in speech. *ISCA Workshop on Speech and Emotion*.

[20] Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. *Proceedings of Eurospeech 2001, Aalborg, 1,* 87–90.

[21] Küstner, D., Tato, R., Kemp, T., & Meffert, B. (2004). Towards real life applications in emotion recognition; comparing different databases, feature sets, and reinforcement methods for recognizing emotions from speech. In: Andre´ et al. (Ed.), *Affective Dialogue Systems, Proceedings of the Kloster Irsee Tutorial and Research Workshop on Affective Dialogue Systems. Lecture Notes in Computer Science, 3068*. Berlin: Springer.

[22] Laukka, P., Juslin, P.N., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion, 19 (5)*, 633–653.

[23] Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. American Psychologist, 50(5):372–385.

[24] BURCH, N. R, and GREINER, T. H. Drugs and human fatigue: GSR parameters. Journal *of Psychology, 45:3,* 1958.

[25] Lang, P.J. (1980). Behavioral treatment and bio-behavioral assessment: computer applications. In: Sidowski, J.B., Johnson, J.H., & Williams, T.A. (Eds.), *Technology in Mental Health Care Delivery Systems* (pp. 119-137). Norwood, NJ: Ablex.

[26] Feldman Barrett, L. & Russell, J.A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology, 74 (4)*, 967–984.

[27] Kilpatrick, D. G. (1972). "Differential responsiveness of two electrodermal indices to psychological stress and performance of a complex cognitive task." Psychophysiology 9(2): 218-26.

[28] Boucsein, W. (1988). Elektrodermale Aktivität. Grundlagen, Methoden und Anwendungen. Berlin, Heidelberg, New York: Springer.

A Computational Study on Emotions and Temperament in Multi-Agent Systems

Daria Barteneva, Nuno Lau and Luís Paulo Reis[1]

Abstract. Recent advances in neurosciences and psychology have provided evidence that affective phenomena pervade intelligence at many levels, being inseparable from the cognition-action loop. Perception, attention, memory, learning, decision-making, adaptation, communication and social interaction are some of the aspects influenced by them. This work draws its inspirations from neurobiology, psychophysics and sociology to approach the problem of building autonomous robots capable of interacting with each other and building strategies based on temperamental decision mechanism. Modelling emotions is a relatively recent focus in artificial intelligence and cognitive modelling. Such models can ideally inform our understanding of human behavior. We may see the development of computational models of emotion as a core research focus that will facilitate advances in the large array of computational systems that model, interpret or influence human behavior. We propose a model based on a scalable, flexible and modular approach to emotion which allows runtime evaluation between emotional quality and performance. The results achieved showed that the strategies based on temperamental decision mechanism strongly influence the system performance and there are evident dependency between emotional state of the agents and their temperamental type, as well as the dependency between the team performance and the temperamental configuration of the team members, and this enable us to conclude that the modular approach to emotional programming based on temperamental theory is the good choice to develop computational mind models for emotional behavioral Multi-Agent systems.

1 INTRODUCTION

Emotions are part of our every day lifes. They help us focus attention, remember, prioritize, understand and communicate. The possibility of computation of emotions has interested researchers for many years. The emotions influence decision-making processes, socialization, communication, learning and many other important issues of our life. Implementation of emotions in an artificial organism is an important step for different areas of intervention, since academical inquiry [1-10], education [13-15], communication [11, 16], entertainment and others [12, 17-19, 29, 30]. Researchers have focused on the functions of emotion for computational models trying to describe some of behavioral responses to reinforcing signals, communications which transmit the internal states or social bonding between individuals, which could increase fitness in the context of evolution. Among some models of emotions that are described through the computational process exists different approaches to the proper concept of emotion. Each model results of the definition that is given to the emotional process. Since analysis of needs/satisfactions of the human being [24, 25], passing through the analysis of characteristics of the superior nervous system [26, 28], physiological changes [23, 31], neurobiological processes [27], appraisal mechanism and analysis of the psychology of individual personality [20, 21].

The most important questions we made in this project are: what is emotion? How can we represent emotions through computational model? Many authors have tried to categorized emotions. As Marvin Minsky said [22] our culture sees emotions as a deep and ancient mystery. He also reinforce that the psychology has not even reached a consensus on which emotions exists. He describes emotions as a complex rule schemes which we develop during our mental grow submitted to external influences like learning. A more detailed approach was made by António Damasio [27] who studied emotions from a neurobiological perspective. He divided emotions in two groups: primary and secondary. In first group he included emotions which depends on our limbic system and these are innate reactions. In the second group Damásio included emotions we develop during our life based on our experience. Finally he defined emotions as a process of mental evaluation, simple or complex, with disposal responses directed to the body and the brain resulting in additional brain alterations.

For our project we define emotions as a set of external and internal responses which depends on the set of rules based on agent beliefs, desires and intentions. To proceed with development of our emotional model we need to use some kind of quantitative measure to evaluate emotional state. But what is this “Emotional state”? Interesting definition was given by Mehrabian [21] for this concept. He defined it as transitory conditions of the organism – conditions that can vary substantially, and even rapidly, over the course of a day. He also defined “emotional traits” (i.e. Temperament) as conditions that are stable over periods of the year or even a lifetime. As described in Pavlov's theory [28], all human and animal behaviors are coordinated by the Central Nervous System (CNS). Therefore we can't study emotional agents without considering the particularities of the CNS and, consequently, the particularities of temperamental theory.

[1] A manuscript submission in January 8, 2007.

Luís Paulo Reis is Auxiliar Professor at the Faculty of Engineering of the University of Porto and Researcher at LIACC (NIAD&R) - Artificial Intelligence and Computer Science Lab. (Distributed Artificial Intelligence and Robotics Unit), Portugal (phone: +351 22 5081829; fax: +351 22 5081443; e-mail: lpreis@fe.up.pt).

Nuno Lau is Auxiliar Professor at the Department of Electronics and Telecommunications of the Aveiro University and Researcher at IEETA – Institute of Electronic Engineering and Telematics of Aveiro, Portugal (phone: +351 34 370500; fax: +351 34 370545; e-mail: lau@det.ua.pt).

Daria Barteneva is Master of Science Student at the Faculty of Engineering of the University of Porto, Portugal (e-mail: daria.barteneva@gmail.com).

The classical definition for “Temperament” follows: it is a specific feature of Man, which determines the dynamics of his mental activity and behaviour. Two basic indexes of the dynamics of mental processes and behaviours at present are distinguishable: activity and emotionality. In this project we will analyze and develop an emotional model for the agents with temperament. We will use a complex approach to emotion/temperament concepts: based on physiological (CNS) characteristics and on psychological characteristics of the agents.

Our computational model of emotion is inspired on appraisal theory and on superior nervous system characteristics. Most appraisal theories [32, 33] assume that beliefs, desires and intentions are the basis of reasoning and thus of emotional evaluation of the agents situation. In order to create a more flexible and efficient emotion-based behavior system, the appraisal model is implemented in mixture with Pavlov's temperamental theory [28] which studies the basic reasons for different temperamental behaviors and Eysenck's [26] neurophysiological interpretation of the basic measurements of temperament.

We choose to mix these different theories to study the agent teams and to evaluate the team performance when we define different temperamental configuration of the team.

A simulation environment based on a Cyber-Mouse [34, 35] competition simulator was used to test and evaluate the strategies used on the work. A set of robotic experiments was conducted in order to test the performance of the system.

The paper is organized as follows. Section 2 presents the Emotions & Temperament approach distinguishing the physiological and psychical perspectives to the subject of our study in order to explain the fundamentals for the implementation performed for this project. Here we present Pavlov's theory about superior nervous system, Eysenck's scale and Mehrabian's PAD model. Section 3 describes the simulated environment (Cyber-Mouse) we used. Section 4 describes implementation of physiological and psychical layers we have defined for this project. Section 5 describes describes the evaluation experiences we performed to validate our work. Finally Section 6 presents the conclusion we made and future work.

2 EMOTIONS AND TEMPERAMENT

As we already have refered, for constructing our emotional model we studied two subjects: emotional states which characterize the immediate emotional condition of the agent and emotional trait (temperament) which define the personality characteristics and behaviors of the agent and influence his emotional state changes. We decided to approach the study of emotions from different perspectives: physiological and psychical, creating double layer architecture for emotional model to increase the system performance. Let us examine each perspective of our approach.

A General framework for describing central nervous system and Pavlov's theory

In order to find physiological reasons to emotions appearance we started by studying the Pavlov's theory [28] about superior nervous system activity. Activity can be expressed in different degrees of tendency to act, to participate in the diverse challenges. It is possible to note two extremes: from one side, high energy, fervency and swiftness in the mental activity, the motions and the speech, while another - passiveness, sluggishness, the apathy of mental activity, motion and speech. The second index is dynamicity and it is expressed in different degrees of emotional excitability, in the velocity of appearance and the force of the emotions of man, and in the emotional sensitiveness (receptivity to the emotional actions). Based on this characteristics four basic forms of temperament may be distinguished, which were named as follows: sanguine (living), phlegmatic (slow, calm), choleric (energetic, passionate) and melancholic (locked, inclined to the deep experiences).

The definite scientific explanation of temperaments was given by Ivan Pavlov's theory about the types of higher nervous activity. Pavlov described three properties of the processes of excitation and braking [37]:

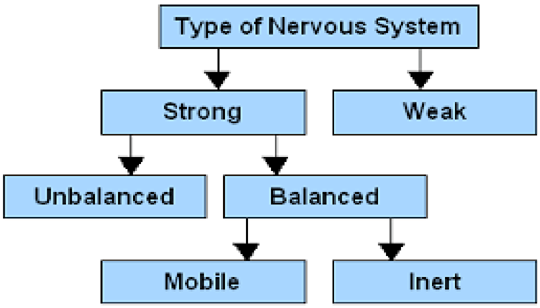


Figure 1. Classification of higher nervous system

- The **force** of the processes of excitation and braking;
- The **steadiness** of the processes of excitation and braking;
- The **mobility** of the processes of excitation and braking.

The combinations of the properties of nervous processes indicated were assumed as basis to determinations of the type of higher nervous activity. Depending on the combination of force, mobility and steadiness of the processes of excitation and braking four basic types of higher nervous activity are distinguished.

Pavlov correlated the types of nervous systems with the psychological types of temperaments isolated with it and revealed their complete similarity. Thus, temperament is a manifestation of the type of nervous system into the activity. As a result the relationship of the types of nervous system and temperaments appears as follows (fig. 1):

- Strong, balanced, mobile type - sanguine temperament;
- Strong, balanced, inert type - phlegmatic

temperament;

- Strong, unbalanced, with the predominance of excitation - choleric temperament;
- Weak type - melancholic temperament.

Eysenck methodology

One of the things Pavlov tried with his dogs [37] was conflicting conditioning - ringing a bell that signalled food at the same time as another bell that signalled the end of the meal. Some dogs took it well, and maintain their cheerfulness. Some got angry and barked like crazy. Some just laid down and fell asleep. And some whimpered and whined and seemed to have a nervous breakdown.

Pavlov believed that he could account for these personality types with two dimensions: On the one hand there is the overall level of arousal (called excitation) that the dogs' brains had available. On the other, there was the ability the dogs' brains had of changing their level of arousal - i.e. the level of inhibition that their brains had available.

- Lots of arousal, but good inhibition: sanguine.
- Lots of arousal, but poor inhibition: choleric.
- Not much arousal, plus good inhibition: phlegmatic.
- Not much arousal, plus poor inhibition: melancholy.

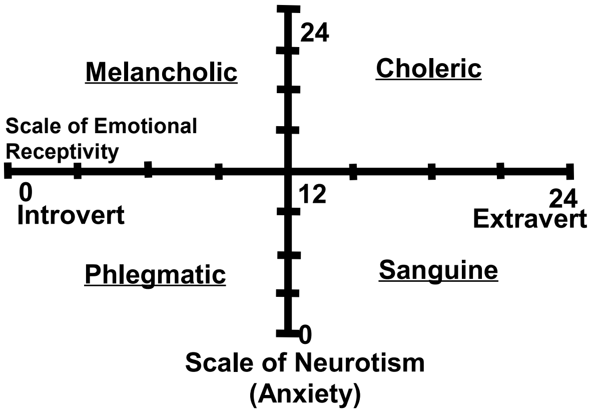


Figure 2. Scale of temperamental classification by Eysenck

Arousal would be analogous to warmth, inhibition analogous to moisture. This became the inspiration for Hans Eysenck's theory [26].

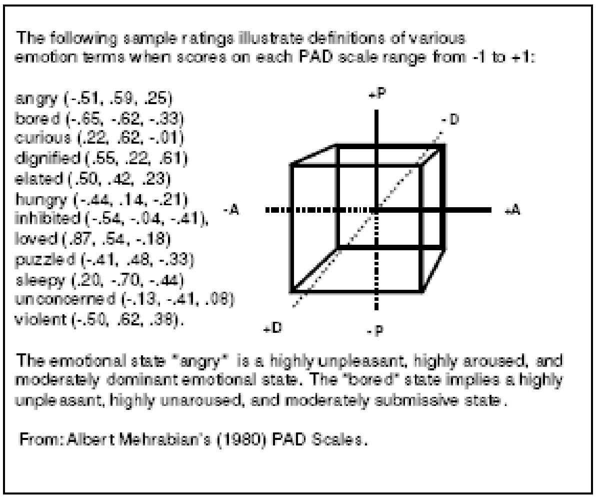
In the works of Eysenck a neurophysiological interpretation of the basic measurements of temperament was given, among which were separated - the factor of extraversion -introversion and the factor of neurotism.

Using the methodology of Eysenck [26] we can perform the personality test to describe the temperament of the individuals by Introvert/Extravert characteristic and Anxiety (Fig. 2).

Analysis of personality factors in terms of the PAD temperamental model

Analysis of emotional states leads to the conclusion that the human emotions such as anger, fear, depression, elation, etc. are discrete and we need to define some kind of measures to have a basic framework to describe each emotional state using the same scale. After studing the appraisal theory we find Mehrabian model [20, 21] more suitable for computational needs since it defines three dimensions to describe each emotional state and provides an extensive list of emotional labels for points in the PAD space (Fig 3) gives an impression of the emotional meaning of combinations of Pleasure, Arousal and Dominance (PAD).

The three dimensions of the PAD temperament model define a three-dimensional space where individuals are represented as points, personality types are represented as regions and personality scales are represented as straight lines passing through the intersection point of the three axes. Mehrabian uses +P, +A and +D to refer pleasant, arousable and dominant temperament. Respectively, and by using -P, -A and -D to refer unpleasant, unarousable and submissive temperament, respectively. Since most personality scales load on two or more of the PAD temperament dimensions, Mehrabian define them using the four diagonals in PAD space as follow:



The following sample ratings illustrate definitions of various emotion terms when scores on each PAD scale range from -1 to +1:

angry (-.51, .59, .25)
bored (-.65, -.62, -.33)
curious (.22, .62, -.01)
dignified (.55, .22, .61)
elated (.50, .42, .23)
hungry (-.44, .14, -.21)
inhibited (-.54, -.04, -.41),
loved (.87, .54, -.18)
puzzled (-.41, .48, -.33)
sleepy (.20, -.70, -.44)
unconcerned (-.13, -.41, .08)
violent (-.50, .62, .38).

The emotional state "angry" is a highly unpleasant, highly aroused, and moderately dominant emotional state. The "bored" state implies a highly unpleasant, highly unaroused, and moderately submissive state.

From: Albert Mehrabian's (1980) PAD Scales.

Figure 3. Mehrabian PAD temperamental scale.

- Exuberant (+P+A+D) vs Bored (-P-A-D)
- Dependent (+P+A-D) vs Disdainful (-P-A+D)
- Relaxed (+P-A+D) vs Anxious (-P+A-D)
- Docile (+P-A-D) vs Hostile (-P+A+D)

In the Analysis of Big-Five Personality factors in terms of PAD temperamental model [38] Mehrabian find the relationship between five temperamental types and the PAD scale. He describe this relationship using linear regressions. The resulting equations are given below for standardized variables with a 0.05 significant level:

Extraversion =	0.24P		+0.72D
Agreeableness =	0.76P	+0.17A	-0.19D
Conscientiousness =	0.29P		+0.28D
Emotional Stability =	0.50P	-0.55A	
Sophistication =		+0.28A	+0.60D

Big Five personality factors

The Big Five factors and their constituent traits can be summarized as follows:

- *Openness to Experience (or Sophistication)* - Appreciation for art, emotion, adventure, unusual ideas; imagination and curiosity.

- *Conscientiousness* - A tendency to show self-discipline, act dutifully, and aim for achievement (spontaneousness vs planned behavior).
- *Extraversion* - Energy, surgency, and the tendency to seek stimulation and the company of others.
- *Agreeableness* - A tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others (individualism vs cooperative solutions).
- *Neuroticism* (or *Emotional Stability*) - A tendency to easily experience unpleasant emotions such as anger, anxiety, depression, or vulnerability (emotional stability to stimuli).

We will use this result to determine the emotional state of the agents depending on their temperamental type.

3 CYBER-MOUSE SIMULATION ENVIRONMENT

We choose the Cyber-Mouse Simulation environment to implement and test our model because it offers an open, modular and flexible platform permitting unlimited applications and a fully configurable simulation system.

Cyber-Mouse is a modality included in the Micro-Mouse competition organized by Aveiro University (Portugal). This modality is directed to teams interested in the algorithmic issues and software control of mobile autonomous robots. This modality is supported by a software environment, which simulates both robots and a labyrinth [34].

The simulation system possesses a distributed architecture where some types of applications communicate among each other, nominated, a simulator, an application for each agent and a viewer application. The architecture is client-server, where the simulator acts as the server and both the agents and the viewer, acts as clients. This architecture is similar to the Simulation League of RoboCup [36].

The simulator shapes all the components of the robots hardware and the labyrinth. The simulation is executed in discrete time, cycle by cycle. In the beginning of each cycle of simulation the simulator sends to all robotic agents in test, the measures of its sensors, and to all viewers the positions and robots information. The agents can answer with the power values to apply to the engines that command the wheels.

All robots in test have the same physiological characteristics. All have the same sensors and the same engines.

Each robot (fig. 4) is equipped with the following sensors [34, 35]:

- 3 sensors of proximity guided to the front and 60º for each side.
- Beacon sensor that indicates which is the difference between robots direction and the beacon direction.
- Ground sensor, active when robot enters in the arrival zone.
- Compass sensor that allows robot to know which its absolute orientation in the labyrinth is.
- Collision sensor, asset in the case of robot collision.
- Vision sensor, works by identifying other robots and their emotional state.

The measures of the sensors include some noise added for the shape simulator in order to simulate real sensors.

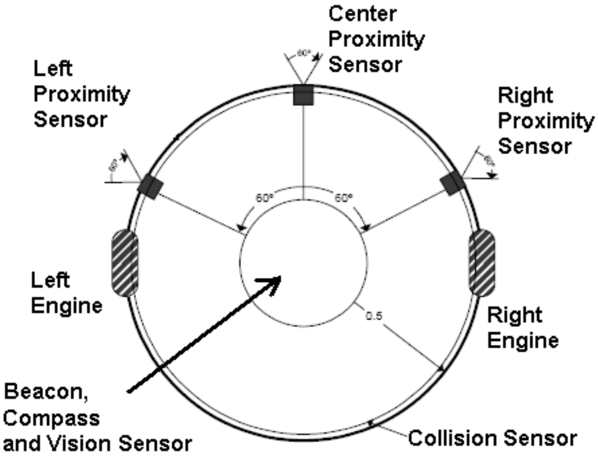


Figure 4. Virtual Agent Diagram

In order to detect the beginning of the test and possible interruptions each robot has 2 buttons:

- Start, active when is initiated the test.
- Stop, active when a test interruption exists.

In terms of virtual engines robots is constituted by:

- 2 wheels for 2 independent motors, one on the left and one on the right;
- LED of finishing, to light when reached the arrival zone.

In each cycle of simulation the agents receive the values measured by all its sensors and must decide which power to apply in each motor. The perception that a robotic agent has from the exterior environment is limited and noisy transforming him into the most appropriate tool to perform our work with almost realistic precision.

4 A DUAL LAYER MODEL OF EMOTION

As we reference in previous chapters, we choose the approach to emotional programming through the implementation of artificial personalities and the integration of the emotional decision model based on the appraisal theory. The innovation of our approach consists in the duality of our emotional character: it processes the information and gives the output using two different engines, physiological and psychological. In our model the temperament of the agent is defined as the configuration of his mechanical engines and the personality functions which simulates his psyche as the decision mechanism. The emotional response of the agent possesses a dual mechanism: it is physiological (such as motor and sensor force, face expression, mobility) and psychical (such as a vector which defines his internal emotional state).

We also need to emphasize the difference between the agent's temperamental state and agent's emotional state. Temperament, as we already defined, is the steady characteristics of the agent which is "innate" and do not suffer alterations during the agent's life. On the other side, the emotional state of the agent is the dynamic set of values which depends on the external influences, and on the agent's temperament.

We can define emotion as a short episode triggered by an (internal/external) event composed of

- subjective feelings
- inclinations to act
- facial expressions
- cognitive evaluation and
- physiological arousal.

And emotions have a role of heuristic relating events to goals, needs, desires, beliefs of an agent and evaluate their personal relevance and help decision-making.

So, for instance, two agents with different temperaments and the same emotional states on some temporal period, which receive the same external input will have different responses on both, the physiological and the psychical mechanism. We also define different sets of needs and motivations for each temperamental type by the influence of the agent's performance and stimuli on the team work. This modular, but complementary approach, is the core of the innovation of our emotional system and our aspiration of its usability.

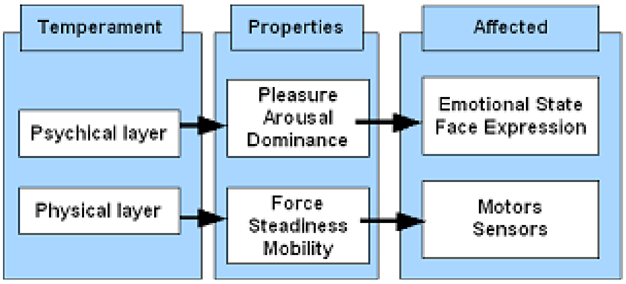


Figure 5. Temperamental architecture.

We assume a two layer architecture (Fig. 5) for our emotional model. One layer is physiological and describes the superior Nervous system from the Pavlov perspective. The other layer is psychical and works with the appraisal model created by Mehrabian.

In our temperamental architecture we have not implemented any of dependency between physiological and psychical layers and we are trying to discover some kind of influence that one layer could have on the other layer through the temperamental configurations or common goals implementation. Psychical layer controls the emotional state of the agent through PAD values, and the physiological layer control the engine configuration (motors, sensors, etc...) and the group interaction, based on temperamental needs of the agent (like extroversion/introversion or emotional stability).

Fig. 6 presents the diagram which describes the relationship between the Simulation environment, Decision layer and Data layer. Decision layer works in order to process the inputs received by the agents from simulation environment, determine the outputs that agent return and update his emotional state. Physiological Bank contains the fuzzy measures of force, mobility and steadiness. Values for motors and sensors are archived in Physiological Bank for each temperamental type.

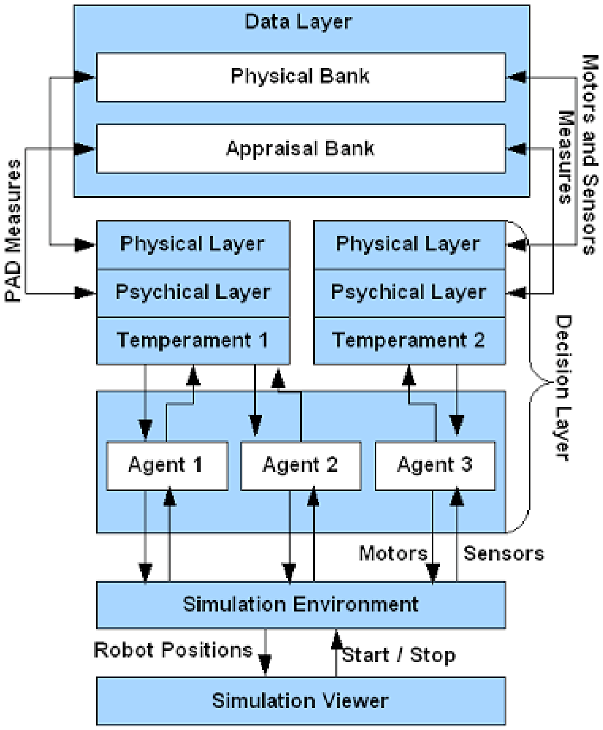


Figure 6. Cyber-Mouse Simulation Architecture vs Appraisal Model and Central Nervous System Mechanism

This figure presents a scheme for two temperaments and three agents, but as we already explain we use four temperaments for physiological layer and 5 temperaments for psychical layer:

- Choleric, Sanguine, Phlegmatic and Melancholic defined in **physiological layer** with different Fuzzy set's of values for Force, Steadiness and Mobility.
- Extraversion, Agreeableness, Conscientiousness, Emotional Stability and Sophistication in **psychical layer** which are described using Pleasure, Arousal and Dominance. In order to simplify our model we are using just Extraversion and Emotional Stability in this project.

Physiological layer

As we show on previous chapter, Pavlov's theory defines the temperamental model based on characteristics of the superior nervous system, but at the same time there are no pure temperamental types in nature, but there are mixtures of different properties which characterize one or another unique temperamental type. So, as we see, one person can have all temperamental types in different ratios. The different proportion of values: force, mobility and steadiness of processes of excitation and braking defines the unique temperamental type for each person. Based on this uncertainty we use Fuzzy Logic to describe and monitorize the temperamental types in our project [39]. In the beginning of the simulation we generate the values which will define the unique combination of temperamental type of the agent, but then these characteristics are changing in run-time in order to adapt the agent state to the external influences. We define the fuzzy intervals for each temperamental variable which define the temperamental characteristics (Force, Mobility, ...) and the value of this variable increases in stressful situations (close threat, wall-shock, etc...) and decreases in calm situations. The speed of the increase and decrease depends on agent's Arousal.

Force

In our multi-agent system the force of excitation and braking processes is represented by the force of the motor and reach of the sensors. We define the superior limit for the force value in order to obtain a better simulation of the real world.

Mobility

The mobility of the agent is represented by its "persistence" to reach the goal and avoid negative emotions. For instance if some agent is "comfortable" in some place, and his mobility is low, he will not look to move to search other places. He will slow his motors and just stay in the same place until the environment changes forces him to move quickly. At the same time, one agent who has a high mobility will search new places and new directories even if he is comfortable enough in some temporal phase. According to Mehrabian [21] arousal is highly correlated with activity and alertness so changing the Arousal we can control the Mobility of the agent.

Steadiness

The steadiness of the agent is the velocity of his emotional state variation. For example, more balanced agents have a slow variation of emotional state. For this we introduce the variable called Anxiety which is used to increase or decrease the Pleasure variable. The value of Anxiety depends on the temperament of the agent. We choose the values for anxiety based on the Eysenck test [26].

Emotional receptivity

This variables were based on the Eysenck test described on the second section. The Melancholic and Phlegmatic temperamental types are included in the Introverts group and the Sanguine and Choleric types are included in the Extroverts group. We will evaluate how they performance to reach the beacon, conditioned by their temperamental needs.

Psychical layer

Our approach does not prescribe a specific set of appraisal dimensions. We have chosen the Pleasure, Arousal, Dominance (PAD) personality-trait and emotional-state scales by Albert Mehrabian [21] because these dimensions are generally not considered to be appraisal dimensions. He argues that any emotion can be expressed in terms of values on these three dimensions, and provides extensive evidence for this claim [20]. This makes his three dimensions suitable for a computational approach. Mehrabian also provides an extensive list of emotional labels for points in the PAD space [21] and gives an impression of the emotional meaning of the combinations of Pleasure, Arousal and Dominance. The emotional-state of an agent can thus be understood as a continuously moving point in an n-dimensional space of appraisal dimensions.

Appraisal Banks

The appraisal bank defines the needs, motivations and stimulus of the agent as a set of subjective measures, called appraisal dimensions. First, a simple instrumentation based on appraisal bank that emotionally evaluates events related to survival. Second, a more complex instrumentation based on two appraisal banks, one related to survival the other related to reach the beacon and satisfy temperamental needs. In both banks we have used event-encoding to simulate emotional meaning of events. We now describe how events are interpreted by the two appraisal banks.

As we work with BDI agent's, their thinking are based on **beliefs**, **desires** and **intentions**. Their possess basics to which appraisal based emotions can be added. Lets describe beliefs, desires and intentions of the agents in our simulation system:

Beliefs:

- angry agents are dangerous;
- wall collisions are painful;
- happy agents are friendly and nice;

Desires:

- reach the beacon;
- satisfies personal (temperamental) need like necessity of company of other agents or necessity of loneliness;
- don't get hurt;

Intentions:

- avoid threats (angry agents);
- avoid wall collisions;
- follow happy agents;

We define **Pleasure** as the *conductance of the goal*. For instance if the agent sees the beacon and no obstacle is present his pleasure is high, while if he sees a threat or looses the goal this is highly unpleasant. **Arousal** is the *amount of attention each event needs*, for instance to avoid threats the attention of the agent is needed and lose it needs no attention. **Dominance** is a measure that defines the *amount of freedom of the agent*. For example, if it sees a wall the dominance decreases but if it sees no obstructed way to the goal the dominance increases. It is not a subject of this paper the detailed definition of the appraisal banks.

Appraisal-results are integrated using following formula,

$$E_{t+1} = E_t + \sum_{i=0}^{n} \Delta PAD_{ti}$$

where E_t is the emotional-state at time t, E_{t+1} is the new emotional-state, n is the number of appraisal banks and ΔPAD_{ti} the appraisal-result vector of bank *i* at time *t*.

5 EXPERIMENTAL RESULTS

Tests were performed to assess the system regarding its performance, temperamental characteristics and emotional behaviors analysis.

To perform our tests, we evaluated the agent's performance on reaching the goal. We also evaluated the appraisal values modifications during the simulation time. We performed the evaluation of an entire team of nine agents, in order to compare their performance with other teams of agents. During these evaluations we tried to analyse the difference between distinct temperamental teams and compare them in general terms (PAD scale and emotion valence), as well as their performance on reaching the goal.

We perform the evaluation on three different simulation scenarios:

- First scenario has few obstacles (walls) and a small arena. These conditions enable fast detection of the beacon to the agents, but force the interaction between agents across the simulation arena.
- Second scenario has a lot of obstacles (walls) and a larger arena. In this scenario the goal is very difficult to accomplish and the agents have enough space for team interaction (grouping or isolation).
- Third scenario has a lot of obstacles (walls) and a small arena. In this scenario the goal is very difficult to accomplish, and the agents have fewer space for team interaction (grouping or isolation).

To perform our tests we have defined the same simulation time for each team (180 seconds), but if some agent have accomplished the goal during simulation cycle, his simulation time is considered as a time he spend to accomplish the goal. For example, if we have some team with two agents and one of them has accomplished the goal in 200 seconds and other haven't accomplished the goal, the medium time of the team will be calculated as (1800+200)/2=1000 seconds and the best time will be considered as 200 seconds.

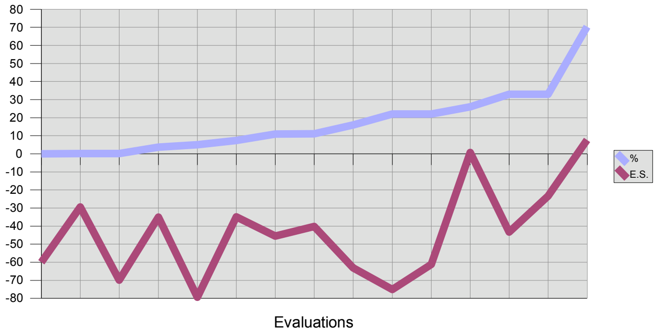


Figure 7*:* Team Performance vs Appraisal emotional state

We analyse our agent's performance on these three simulation scenarios and try to discover the advantages and disadvantages of using temperamental agents with this kind of simulation. We perform 10 interactions for each team/simulation scenario. The PAD values are presented in [-10, 10] interval instead of [-1, 1].

We perform the tests with Choleric, Sanguine, Melancholic and Phlegmatic homogeneous teams and with heterogeneous multi-temperamental team of agents.

Figure 7 shows the dependencies between Team Performance and Emotional State of the agents. In our architecture the performance of the agents doesn't depend on appraisal mechanism which only controls the psychical layer of the agent and only influences his PAD values and the emotional state. The agents performance only depends on temperamental (physiological) configuration of the agent (motors, sensors, anxiety, etc..) and his decision layer based on extrovert/introvert characteristics. So, we can see that the temperamental decision mechanism clearly influence the emotional state of the agent during the simulation.

In order to explain the nature of this influence we think that the implementation of the Pavlov's temperamental theory results in agents executing different grouping activities to satisfy their temperamental needs, so, this could influence the PAD values and consequently the Emotional State of the agent. Also we can analyse the influence of the system goals on the Emotional State of the agent (from the Appraisal Bank), and as we describe in Section 4, the decision temperamental mechanism works in order to accomplish some of these goals (avoid the walls or reach the beacon, for instance). So, even having no direct dependence between the implementation of these two layers, there are similar goals which are defined and this could explain the dependence between the agent performance and his Emotional State.

6 CONCLUSIONS AND FUTURE WORK

A main goal of this project was to develop and test a new model of a computational emotional mind using two different temperamental theories. For tackling this domain, we first had to study the basics of psychology and different approaches to evaluate emotional life of personality from temperamental perspective. As a rule, the theories of emotions can say too little about the role of emotions in the development of personality and about their influence on thought and action. The majority of the researchers of emotions are connected only with one of the components of the emotional process. Although some theories develop the separate aspects of the interrelations of emotion and reason, actions and personality, much still must be done both on the theoretical and on the empirical levels.

In this work we have tried to implement emotional agents using two different approaches: appraisal theory with the PAD model and the central nervous system theory. We approach the concept of emotion from the physiological and psychical perspective, defining the personality of the agent and analysing the different components of agent behaviors. We have simulated a kind of homogeneous and non-homogeneous society with different personalities and analysed their group and individual performances.

We can conclude that our approach produce very good results showing the dependence between two different layers (physiological and psychical) which where implemented independently. So, as it already has been proved theoretically from psychological perspective, which define that our emotional process are dependent on our temperamental type, we could state that our architecture is consistent and show the same dependence between two layers. This let us a large room for future improvement and research on this area.

The modular approach to emotion programming is a very promising way to integrate emotions into multi-agent systems with different goals and configurations. Temperament helps to support agent's decision making and with proper use can improve the agent's performance and the global teamwork. Also, our system helps us analyse the configurations we could choose to implement the personality in the system with different and particular characteristics, helping us to select the variables and functions of personality with better fitness to the specific system.

We are planning to implement in this project different search algorithms for evaluate the impact of emotions and temperament on search strategies. Other development is the introduction of visual emotional feedback using the face expressions such as proposed by the Russel [19]. Also we are aiming at the introduction of additional objects in the simulation environment with different degree of thread/satisfaction.

REFERENCES

[1] Hille, K., Synthesizing Emotional Behavior in a Simple Animated Character, *Artificial Life*, 2001, Vol. 7, No. 3, pp. 303-313.

[2] Sloman, A., Croucher, M., *Why Robots will have Emotions*, IJCAI, 1981.

[3] Sarmento, L., *An Emotional-Based Agent Architecture*, Msc Thesis, Faculty of Science of Porto University, Portugal, 2004

[4] Picard, R. W., *Affective Computing,* MIT Press, 1997

[5] Velásquez, J., *Modeling emotion-based decision-making*. In Dolores Cañanero, editor, Proceedings of the 1998 AAAI Fall Symposium Emotional and Intelligent: The Tangled Knot of Cognition, pages 164–169, Orlando, FL, USA, 1998. http://www.ai.mit.edu/people/jvelas/papers/Velasquez-FS98.ps.

[6] Cañamero, L., *Designing emotions for activity selection.* Technical Report TR - DAIMI PB 546, Univesity of Arhaus, 2000. http://citeseer.nj.nec.com/384344.html.

[7] Cañamero, L., *A hormonal model of emotions for behavior control.* Presented as poster at the Fourth European Conference on Artificial Life (ECAL ' 97), Brighton, UK, 1997. http://citeseer.nj.nec.com/canamero97hormonal.html.

[8] Cañamero, L., *Issues in the design of emotional agents.* In AAAI Fall Symposium on Emotional and Intelligent: The tangled knot of cognition, pages 49–54, Menlo Park, CA, 1998. http://citeseer.nj.nec.com/437878.html.

[9] Frijda, N. H., *The Emotions.* Cambridge University Press, 1986.

[10] Frijda, N. H., *Emotions are functional, most of the time*. In Paul Ekma and Richard The Nature of Emotion - Fundamental Questions. Oxford University Press, 1994.

[11] Gebhard, P., ALMA – A Layered Model of Affect, *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'05)*, 29-36, Utrecht, 2005

[12] Pereira, D., Oliveira, E., Moreira, N., Sarmento, L., "Towards an Architecture for Emotional BDI Agents", *In IEEE Proceedings of EPIA*, Covilhã, December 2005

[13] Dias, J., Paiva, A., "Feeling and Reasoning: a Computational Model for Emotional Characters", in EPIA Affective Computing Workshop, Covilhã, Springer, 2005

[14] Prada, R., *Teaming Up Humans and Synthetic Characters* - PhD Thesis, UTL-IST, Lisboa, Dezembro de 2005.

[15] Prada, R., Paiva, A.: *Believable groups of synthetic characters*. AAMAS2005: 37-43 ACM-Press

[16] Bartneck, C., *How convincing is Mr. Data's smile: Affective expressions of machines. User Modeling and User-Adapted Interaction,* 11, pp. 279-295. 2001

[17] Ortony, A., Clore, G. L., Collins, A., "*The Cognitive Structure of Emotions*." Cambridge University Press, Cambridge, UK, 1988.

[18] Petta, P., "*The role of emotions in tractable architectures for situated cognizers*." In Robert Trappl, Paolo Petta, and Sabine Payr, editors, Emotion in Humans and Artifacts. MIT Press, Cambridge, MA, 2002.

[19] Russel, J.A., A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 1980, No. 39, pp. 1161-1178.

[20] Mehrabian, A. Pleasure-Arousal-Dominance: A General Framework for Describing and Measuring Individual Differences in Temperament. *Current Psychology: Developmental, Learning, Personality, Social,* Winter, 1996, Vol. 14, No. 4, 261-292.

[21] Mehrabian, A., Basic Dimensions for a General Psychological Theory, *Cambridge: OG&H Publishers*, 1980.

[22] Minsky, M., The Society of Mind, First Touchstone Edition, 1988.

[23] Darwin, C., "*The Expression of the Emotions in Man and Animals*", 1872

[24] Izard, Carroll E., *Human Emotions*. Plenum Press, New York, 1977.

[25] Izard, C. E., Libero, D. Z., Putnam, P., & Haynes, O. M., "Stability of emotion experiences and their relations to traits of personality." *Journal of Personality and Social Psychology*, 64, 847-860, 1993

[26] Eysenck, H. J. , *The Scientific Study of Personality*, L., 1952

[27] Damásio, A. R., *Descartes's Error: Emotion, Reason and the Human Brain.* Gosset/Putman, New York, 1994.

[28] Павлов, И.П., *Общие типы высшей нервной деятельности животных и человека*, Москва, 1927/ Pavlov, I. P., *General Types of Superior Nervous System in Animal and Man*, Moscow, 1927

[29] Gadanho, S. C., *Emotional and cognitive adaptation in real environments.* In the symposium ACE'2002 of the 16th European Meeting on Cybernetics and Systems Research, Vienna, Austria, 2002.

[30] Gadanho, S. C., Hallam, J. *Emotion triggered learning for autonomous robots.* In Dolores Cañamero, Chisato Numaoka, and Paollo Petta, editors, SAB '98 Workshop on Grounding Emotions in Adaptive Systems, pages 31–36, August 1998.

[31] James, W., "*What is an emotion?",* Mind, 9(34):188–205, 1884.

[32] Gratch, J., Mao, W., Marsella, S., "Modeling Social Emotions and Social Attributions", *Cognition and Multi-Agent Interaction,* Cambridge University Press, pp. 219-251, 2005

[33] Broekens, J., DeGroot, D., *Formalizing Cognitive Appraisal: From Theory to Computation*, Proceedings of the 18th European Meeting on Cybernetics and Systems Research (EMCSR 2006, Vienna, Austria) (pp. 595-600), ASCS: Vienna, 2006

[34] Lau, N., Pereira, A., Melo, A., Neves, A., Figueiredo, J., Ciber-Rato: Um Ambiente de Simulação de Robots Móveis e Autónomos, *Revista do DETUA* / Ciber-Mouse: A Simulation Environment for Mobile and Autonomous Robots, *DETUA Journal*, September, 2002, Vol. 3, No. 7, pp. 647-650.

[35] Lau, N. *Manual do Simulador Ciber-Rato./ Manual for Ciber-Mouse Simulator*, Aveiro University, Portugal, 2001

[36] RoboCup Official Site, http://www.robocup.org

[37] Павлов И. П. Полное собрание сочинений. Изд. 2-е. Т. 3 — 4. М.; Л., 1951.

[38] Mehrabian, A., Analysis of the Big-five Personality Factors in Terms of the PAD Temperamental Model. *Australian Journal of Psychology,* 1996, Vol. 48, No. 2, pp. 86-92.

[39] Barteneva, D., Lau, N., Reis, L. P., "Implementation of Emotional Behaviors in Multi-Agent System using Fuzzy Logic and Temperamental Decision Mechanism", *In Proceedings of the Fourth European Workshop on Multi-Agent Systems*, pp. 5-11, Lisbon, 2006

A Mathematical Model to Analyse the Dynamics of Gesture Expressivity

Ginevra Castellano, Antonio Camurri, Barbara Mazzarino and Gualtiero Volpe [1]

Abstract. In this paper we focus on video analysis of movement and gesture as indicators of an emotional process. We propose a novel approach for the analysis of human emotional behaviour based on a mathematical model explaining the dynamics of expressive motion cues. Our method consists of defining indicators conveying information about the dynamics of the gesture expressivity: starting from the temporal profiles of expressive cues automatically extracted, information about their shape are calculated. This approach allows us to obtain indicators of gesture expressivity that can be used for automated emotion recognition. We report a pilot study that shows how our model is used to differentiate emotional motor behaviour in music performances.

1 INTRODUCTION

One critical aspect of the human-machine interfaces is the ability to communicate with users in an expressive way [14]. In the ambient intelligence and smart environments field there is an increasing attention on designing systems able to recognise and interpret users' emotional states and to communicate expressive-emotional information to them.

In this context a central role is played by automated video analysis aiming to extract and describe information related to the emotional state of individuals. In this paper we focus on analysis of movement and gesture as indicators of an emotional process.

Several studies focus on the relationships between emotion and movement qualities, and investigate expressive body movements (see [17], [19], [8], [3], [20], [5], [15]). Nevertheless, modelling emotional behaviour starting from automatic analysis of visual stimuli is a still poorly explored field.

Camurri and colleagues (see [4] and [7]) classified expressive gesture in human full-body movement (dance performances) and in motor responses of subjects exposed to music stimuli: they identified cues deemed important for emotion recognition and showed how these cues could be tracked by automated recognition techniques. Other studies show that expressive gesture analysis and classification can be obtained by means of automatic image processing and manual annotation techniques (see [2], [9], [12]).

Our goal is to define a model to recognise emotions from video analysis of movement and gesture. Deriving such a model may enhance smart environments, since they would be capable of inferring the emotional states of users in a non-intrusive way. The possibility to use movement and gesture as indicators of the state of individuals provides a novel approach to quantitatively observe and evaluate the users in an ecological environment and to respond adaptively to them. Combining the naturalness of expressive body movement with the communicative power of affective, emotional non-verbal communication may then provide new interaction strategies between humans and machines.

This study aims to identify the movement cues explaining expressivity in gestures and to define indicators useful for automated emotion recognition. In this paper we propose a novel approach for analysis of emotional behaviour based on a mathematical model explaining the dynamics of expressive cues. In the following sections we report an overview of the cues explaining expressivity in movement that we automatically extract. Further, we describe our model and report a case study that shows how our approach works.

2 EXPRESSIVITY IN MOVEMENT

In a cross-disciplinary perspective, research on expressive cues describing emotional aspects of human motion and gesture can build on several bases, ranging from biomechanics, to psychology, to theories coming from performing arts. For example, in our work we considered theories from choreography like Rudolf Laban's Theory of Effort [10][11], theories from music and composition like Pierre Shaeffer's Sound Morphology [16], works by psychologists on non-verbal communication in general [1], on expressive cues in human full-body movement [3][20], on components involved in emotional responses to music [18].

Starting from the above mentioned theories, we (i) identified a collection of expressive cues, i.e., descriptors providing information about the emotional/expressive content conveyed by users through movement and gesture, (ii) developed techniques (computational models, algorithms) for extracting expressive cues from visual input, (iii) developed software modules implementing such techniques.

Software modules have been implemented as modules for the EyesWeb XMI platform for enhanced processing of multimodal streams (see www.eyesweb.org). Such modules are included in the EyesWeb Expressive Gesture Processing Library.

In this section we provide a brief overview of the most relevant expressive cues we automatically extract and use in our research (for more details see [4][5]). The profile of such cues along time provides the input data on which we apply the dynamic approach described in the following section.

Two major expressive cues we frequently use in our work are *Quantity of Motion* and *Contraction Index*.

Quantity of Motion (QoM) is the amount of detected movement in a gesture or movement sequence. It is based on the Silhouette Motion Images. A Silhouette Motion Image (SMI) is an image carrying information about variations of the silhouette shape and

[1] InfoMus Lab, DIST, University of Genova, Italy, emails: {ginevra.castellano; antonio.camurri; barbara.mazzarino; gualtiero.volpe}@unige.it

position in the last few frames. SMIs are inspired to motion-energy images (MEI) and motion-history images (MHI). They differ from MEIs in the fact that the silhouette in the last (more recent) frame is removed from the output image. QoM is computed as the area (i.e., number of pixels) of a SMI. It can be considered as an overall measure of the amount of detected motion, involving velocity and force. Algorithms were developed to compute both the overall QoM and the QoM internal to the body silhouette.

Contraction Index (CI) is a measure, ranging from 0 to 1, of how a movement/gesture is contracted (i.e., performed near to the body) or expanded (i.e., performed with a wide use the space surrounding the body). CI is related to Laban's "personal space". It can be calculated in two different ways: (i) considering as contraction index the eccentricity of an ellipse approximating the body trunk, (ii) using a technique related to the bounding region, i.e., the minimum rectangle surrounding the dancer's body: the algorithm compares the area covered by this rectangle with the area currently covered by the silhouette.

Another expressive cues related to Laban's Theory of Effort is *Directness Index*. Directness Index (DI) is a cue related to the geometric features of a movement trajectory. It is a measure of how much a trajectory is direct or flexible. In the Laban's Theory of Effort it is related to the Space dimension. In the current implementation the DI is computed as the ratio between the length of the straight line connecting the first and last point of a given trajectory and the sum of the lengths of each segment constituting the given trajectory. Therefore, the more it is near to one, the more direct is the trajectory. In a recent development we also defined and extracted Contraction Index for trajectories, i.e., starting from the geometric features of a motion trajectory as for DI. Briefly, a motion trajectory is considered as contracted if its points are located in a limited space around the barycentre of the trajectory.

Two more expressive cues we automatically extract from video sequences of human motion are *Fluency* and *Impulsiveness*.

Impulsiveness is a first example of expressive cue computed from the temporal profiles of other expressive cues. It is computed from the shapes of the motion bells obtained from QoM: for each motion phase the ratio is calculated between the peak of the QoM in the motion phase and the duration of the motion phase. Impulsive movement will be characterized by short motion phase with high QoM, while sustained movement will be characterised by longer motion phases with low variance of the QoM.

Fluency is estimated starting from an analysis of the temporal sequence of motion and pause phases. A gesture performed with frequent stops and restarts (i.e., characterised by a high number of short pause and motion phases) will give less Fluency than the same movement performed in a continuous, "harmonic" way (i.e., along a few long motion phases). The hesitating, bounded performance will be characterised by a higher percentage of acceleration and deceleration in the time unit (due to the frequent stops and restarts). Actually, fluent movements are also continuous with few and short breaks, while more rigid, strong movements are emphasized by more frequent and long pauses. In a recent research related to believability of virtual characters, we compared human movement with the reconstructed movement of a virtual character performing the same motion sequence [13]. Reconstruction was performed starting from the same data obtained from the human. In this framework another approach to Fluency was undertaken, addressing the difference in the profiles of the motion of the upper and lower part of the body.

In another recent work we applied automatic description of the above-described expressive cues for classification of hand gesture according to the dimensions of Laban's Theory of Effort. An array of expressive cues including impulsiveness, trajectory contraction index, quantity of motion etc. was used to classify gestures as quick or sustained (Time dimension of Laban's Theory) and as direct or flexible (Space dimension).

3 A DYNAMIC APPROACH FOR GESTURE EXPRESSIVITY ANALYSIS

The automated extraction of the expressive cues described in the previous section allows to characterize a gesture or sequence of movements with temporal series correspondent to the profiles over time of the same expressive cues. In order to obtain proper inputs for machine learning algorithms for emotion recognition it is necessary to convert these temporal series into useful indicators. Our approach consists of defining dynamic indicators conveying information about the dynamics of the gesture expressivity: starting from the temporal profiles of the expressive cues automatically extracted, information about their shape are calculated.

In the following, we propose a mathematical model for the analysis of gesture expressivity dynamics.

Given with $\vec{y} = \lfloor y_{1,\dots,} y_N \rfloor$ the sequence of values of an expressive cue over time in a gesture (Nk = number of samples for an expressive cue in the k-th gesture), $\vec{m} = \lfloor m_{1,\dots,} m_M \rfloor$ (Mk = number of relative maxima in the k-th gesture) the sequence of the maximum values of an expressive cue in the gesture, ordered over time, we define the following features aiming to explain the dynamics of an expressive cue.

In the following, we call 'gesture' the temporal profile of an expressive cue.

Initial slope of the gesture's temporal profile

This feature is computed as the slope of the straight line joining the first maximum value of the gesture and its initial value

$$\begin{cases} \dfrac{(m_1 - y_1)}{\Delta t} & \text{if } m_1 \neq y_1 \\ 0 & \text{otherwise,} \end{cases} \qquad (1)$$

where Δt is the temporal distance between the two points.

Final slope of the gesture's temporal profile

This feature is computed as the slope of the straight line joining the last maximum value of the gesture and its final value

$$\begin{cases} \dfrac{(y_N - m_M)}{\Delta t} & \text{if } m_M \neq y_N \\ 0 & \text{otherwise,} \end{cases} \qquad (2)$$

where Δt is the temporal distance between the two points.

Initial slope of the main bell
This feature is related to the gesture's main bell containing the gesture's absolute maximum. It is computed as the first derivative between two significant points of the bell, that is the absolute maximum of the gesture and the minimum value preceding it

$$\frac{(M-\min_{pre})}{\Delta t}, \qquad (3)$$

where $M=\max\{y_i, i=1..N\}$and $\min_{pre}$ is the minimum value preceding the absolute maximum and Δt is the temporal distance between the two points.

Final slope of the main bell
This feature is related to the gesture's main bell containing the gesture's absolute maximum. It is computed as the first derivative between two significant points of the bell, that is the absolute maximum of the gesture and the minimum value following it

$$\frac{(\min_{post}-M)}{\Delta t}, \qquad (4)$$

where $M=\max\{y_i, i=1..N\}$and $\min_{post}$ is the minimum value following the absolute maximum and Δt is the temporal distance between the two points.

Maximum value of the gesture's temporal profile
This feature is related to the main peak of the gesture and is represented by the absolute maximum
($M=\max\{y_i, i=1..N\}$).

MaximumValue/GestureDuration
This feature is related to the ratio between the maximum value of the gesture's temporal profile and its whole duration

$$\frac{M}{GD}, \qquad (5)$$

where $M=\max\{y_i, i=1..N\}$ and GD is the gesture temporal duration.

Mean value of the gesture's temporal profile
This feature is related to the mean behaviour of the gesture and is represented by the mean value of the cue over time (μ).

MeanValue/MaximumValue
This feature is related to the relationship between the mean and the maximum value of the gesture's temporal profile

$$\frac{\mu}{M}. \qquad (6)$$

BellShapeDuration/RemainingGestureDuration
This features is related to the relationship between the duration of the main bell of the gesture and the duration of the remaining parts of the gesture

$$\frac{BD}{RD}, \qquad (7)$$

where BD is the duration of the main bell and RD the duration of the remaining parts of the gesture.

Centroid of energy
This feature gives an estimation of the baricenter of energy. At each step, it takes into account the value of the gesture cue at the current instant (frame). The output of the system is a temporal instant (frame).

$$\frac{\sum_k k y_k}{\sum_k y_k} \qquad (8)$$

Distance between absolute maximum and the centroid of energy
This feature refers to the temporal distance between the absolute maximum of the gesture's temporal profile (M) and its centroid of energy.

It also gives an indication on the relative position of the two points (e.g., the absolute maximum follows or precedes the centroid of energy in the gesture's temporal profile).

Symmetry Index of the gesture's temporal profile
This feature gives an estimation of the symmetry of the curve related to gesture. It is computed by means of the following steps: (1) identifying the centre of the curve ($\bar{x}$), (2) turning over and overlapping the two hemi-parts of the curve, (3) calculating the difference of the areas below the two hemi-parts of the curve and (4) dividing by the total area below the whole curve to normalize between 0 and 1

$$\frac{|\sum_{i=\bar{x}}^{N} y_i - \sum_{i=1}^{\bar{x}} y_i|}{\sum_{i=1}^{N} y_i}, \qquad (9)$$

where y_i are the values of the cue (samples), Δt the temporal distance between the samples, $(\Delta t \sum_{i=\bar{x}}^{N} y_i)$ the area below the right part of the curve with respect to the centre, $(\Delta t \sum_{i=1}^{\bar{x}} y_i)$ the area below the left part of the curve with respect to the centre and $(\Delta t \sum_{i=1}^{N} y_i)$ the total area below the curve of the gesture.

Note: area=sum of the rectangles with Δt as basis and amplitude (y_i) as height.

(0=max symmetry; 1=min symmetry)

Shift Index of the main bell of the gesture

This feature gives an estimation of the position of the main bell of the gesture with respect to its centre. It is computed by means of the following steps: (1) calculating the position of the absolute maximum of the gesture, (2) subtracting the duration of the gesture on the right and that one on the left with respect to the absolute maximum and (3) normalizing with respect to the total duration of the gesture.

Notes: (1) Shift Index <0 → main bell overbalanced on the right of the curve; Shift Index >0 → main bell overbalanced on the left. (2) |Shift Index| = 0 → main bell not overbalanced; |Shift Index| = 1 → main bell overbalanced

$$\frac{D_{right} - D_{left}}{GD}, \qquad (10)$$

where GD is the duration of the gesture, D_{right} the duration of the gesture on the right with respect to the position of the absolute maximum and D_{left} the duration of the gesture on the left with respect to the position of the absolute maximum.

Number of peaks

This feature refers to the number of peaks associated to the main bells of the gesture's temporal profile.

Number of peaks preceding the main one

This feature refers to the number of peaks associated to the main bells preceding the one containing the absolute maximum.

4 CASE STUDY: ANALYSIS OF EMOTIONAL BEHAVIOUR IN MUSIC PERFORMANCES

4.1 Description of the experiment

In order to test the dynamic approach for gesture expressivity analysis we focused on video analysis of emotional behaviour in music performances.

We conducted an experiment, in collaboration with Geneva Emotion Research Group, in which two musicians, a pianist and a cello player, were asked to play a fragment of the same piece (an excerpt from the Sonata no 4 op 102/1 for piano and cello from L. van Beethoven) in different emotional interpretations, respectively "sad", "allegro", "serene", "personal" and "over-expressive". Over-expressive refers to an interpretation that exaggerates an underlying emotional theme in the music to an almost stereotypical level. By "personal interpretation" we requested the emotional interpretation that the two musicians thought was most appropriate to the respective passage and that they would adopt in giving a concert of the piece.

The performances were recorded with three video cameras (SONY DSR-PDX10) with constant shutter, manual gain and focus, at 25 fps, in the repetition hall of the Orchestre de la Suisse Romande in Geneva. We did video recordings from three sides: lateral and top views for pianist and frontal view for cello player.

In this study, we aim to identify the movement cues explaining expressivity in gestures used by performers to communicate (consciously or unconsciously) emotional expression intentions. Therefore, in this work, our goal is to verify whether different music performances of the same melody played in different emotional interpretations can be classified and distinguished only starting from the different motor behaviour of the musicians. The questions we address here are the following: (1) How do movement cues covary with different emotional interpretations by music performers? (2) If this is the case, which of these movement cues are most affected by emotional quality?

4.2 Analysis and results

We performed a movement analysis of the pianist by extracting motion cues from the video recordings. Motion cues were extracted using EyesWeb [6]. Movement analysis focused on the quantity of motion (QoM) of the body and the velocity of the head (Velocity). We applied the dynamic approach for gesture expressivity analysis to the temporal profiles of QoM and Velocity, in order to verify if the dynamics of the pianist's gestures vary in the different emotional conditions.

In the analysis, we considered the first musical phrase of the excerpt (measures: end of 28 - beginning of 30). We identified the three distinct gestures forming the whole phrase. We obtained this segmentation from the QoM of the pianist and we considered the separation between two consecutive gestures given by a local minimum. We calculated the features reported in chapter 3 both for QoM and Velocity, with the following difference: we extracted the initial and final slope of the gesture's temporal profile for the QoM and the initial and final slope of the main bell for the Velocity. This choice is due to the typical shape of the Velocity profiles: Velocity has no peaks until the central part of the shape of the gesture; on the contrary, the QoM has several peaks, for example due to the preparation of the gesture.

The analysis of the extracted values of the features for QoM and Velocity of the baricenter of the pianist suggests some general observations: (1) "sad" gestures have lower values of initial and final slope of the main bell for the Velocity; this can be considered as an indicator of impulsiveness; (2) the MaximumValue/GestureDuration parameter explains the differences according to gesture shapes, mainly in the "overexpressive", "personal" and "sad" gestures; (3) "sad" gestures are characterised by lower number of peaks; (4) a high percentage of gestures (80%) have the main peak in the second half of the gesture; (5) the 66% of "sad" gestures have the highest values of the BellShapeDuration/RemainingGestureDuration parameter, that is, sad gestures have a prevalent main bell in each gesture.

We can conclude that the dynamics of QoM and Velocity differentiate between the emotional conditions of the pianist. In particular, the parameter MaximumValue/GestureDuration explain differences in the temporal variation of QoM and Velocity. Furthermore, "sad" gestures distinguish well from the other ones according to slopes of the main bell of Velocity, the number of peaks and the position and the prevalence of the main bell both in QoM and Velocity. Sad performances are then characterized by temporal profiles of QoM showing few and large peaks and temporal profile of Velocity showing few, large and smooth peak slopes, thus "Sad" performances result best characterised by the selected cues.

4.3 Validation with subjects

To complete the study we set up an experiment with a group of subjects, in order to verify whether the audio-visual stimuli we used to analyse the movement of the pianist were communicative, i.e., whether people are able to recognise emotional states of the pianist in the different performances of the same melody.

4.3.1 Participants

A group of thirty-six people (twenty male and sixteen female) from twenty-three to fifty-seven years old participated to the experiment. Thirteen of them were musicians.

4.3.2 Procedure

Participants were divided in three groups, each one made up twelve people. Each group was asked to fill in a questionnaire to indicate the emotional characterisation of the five performances of the pianist (each time subjects had six different options: the five emotional labels related to the emotional characterisation of the performances and the option "nothing of the previous responses"), in three different conditions:

1) The first group could see the video of the performances and hear the related audio.
2) The second group could only see the video.
3) The third group could only hear the audio.

4.3.3 Analysis and results

Analysis of the questionnaires highlighted that:

- the first group obtained a recognition of the 18,33% of the emotional characterisations of the performances and that the most recognised was the "sad" emotion (41,67%);
- the second group obtained a recognition of the 30% of the emotional characterisations of the performances and that the most recognised was the "sad" emotion (58,33%);
- the third group obtained a recognition of the 30% of the emotional characterizations of the performances and that the most recognized was the "over-expressive" emotion (50%), followed by the "sad" emotion (41,67%);
- musicians were more successful then non musicians (with a mean of 1,6 correct responses for person against a mean of 1,1 correct responses for person);
- in all the three groups the "sad" emotion was classified mainly as "sad", contrary to the other ones: there are two cases out of three where the "allegro" emotion is mistaken for the "over-expressive" one and where the "serene" emotion is mistaken for the "allegro" one.

From this experiment we can conclude first that the "sad" emotion is the most recognisable in the three different conditions proposed to participants (video+audio, video only and audio only); secondly, contrary to the expectations, two channels of information (video and audio) do not allow subjects to increase their performance in the recognition of the emotions, suggesting that the cues in the two modalities are redundant.

5 CONCLUSION

The goal of this study is to explore whether it is possible to use automated video analysis of movement and gesture to extract and describe information related to the emotional state of individuals.

We proposed a mathematical model for the analysis of expressivity in gestures that focuses on the dynamics of expressive cues. We defined a set of features derived from visual expressive cues automatically extracted that allow to manage information about emotional states focusing on the movement dynamics. The features proposed are strictly linked to the shape of the curves of the motions expressive cues analysed over time. Following this approach, in the case study we reported, we highlighted that there is a subset of the features explaining the dynamics of specific expressive cues (QoM and Velocity) that differentiates between the emotional conditions of the pianist. The "sad" emotional condition appears to be the most recognisable, result confirmed by the perceptual experiment carried out with subjects.

We consider the results of this pilot study promising from different points of view. First they confirm that the features we defined in our model and that we extracted from automatic video analysis convey information about the emotional states of individuals. Second, they go into the direction of the explanation of the dynamics of an emotional process, investigated by means of an analysis of motor behaviour, and of the definition of a model to recognise emotions from video analysis of movement and gesture.

.

AKNOWLEDGEMENTS

The research work has been realised in the framework of the EU-IST Project HUMAINE (Human-Machine Interaction Network on Emotion), a Network of Excellence (NoE) in the EU 6th Framework Programme (2004-2007). We thank Klaus Scherer and Tanja Baenziger for their contributions to the experiment and the recordings described in the paper.

REFERENCES

[1] M. Argyle, 1980. Bodily Communication. Methuen & Co Ltd, London.

[2] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, 2005. Emotion Analysis in Man- Machine Interaction Systems. Samy Bengio, Hervé Bourlard (Eds.), Machine Learning for Multimodal Interaction, Lecture Notes in Computer Science, Vol. 3361, pp. 318 - 328, Springer-Verlag.

[3] R.T. Boone, and J.G. Cunningham, 1998. Children's decoding of emotion in expressive body movement: the development of cue attunement. Developmental Psychology 34, 1007–1016.

[4] A. Camurri, I. Lagerlöf, and G. Volpe, 2003. Recognizing Emotion from Dance Movement: Comparison of Spectator Recognition and Automated Techniques, International Journal of Human-Computer Studies, 59(1-2), pp. 213-225, Elsevier Science, July 2003.

[5] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and G.Volpe, 2004, Multimodal analysis of expressive gesture in music and dance performances, in A. Camurri, G. Volpe (Eds.), Gesture-based Communication in Human-Computer Interaction, LNAI 2915, Springer Verlag, 2004.

[6] A. Camurri, P. Coletta, A. Massari, B. Mazzarino, M. Peri, M. Ricchetti, A. Ricci, and G. Volpe, 2004, Toward real-time multimodal processing: EyesWeb 4.0, in Proc. AISB 2004 Convention: Motion, Emotion and Cognition, Leeds, UK, March 2004.

[7] A. Camurri, G. Castellano, M. Ricchetti, and G. Volpe, 2006. Subject interfaces: measuring bodily activation during an emotional experience of music. In S. Gibet, N. Courty, J.F. Kamp (Eds.), Gesture in Human-Computer Interaction and Simulation, Volume 3881, pp. 268-279, Springer Verlag, 2006.

[8] M. DeMeijer, 1989. The contribution of general features of body movement to the attribution of emotions. Journal of Nonverbal Behavior (13), 247 - 268.

[9] A. Drosopoulos, T. Mpalomenos, S. Ioannou, K. Karpouzis, and S. Kollias, 2003. Emotionally-rich Man-machine Interaction Based on Gesture Analysis. Human-Computer Interaction International 2003, 22 - 27 June, Crete, Greece, vol. 4, pp. 1372 – 1376.

[10] R. Laban, and F.C. Lawrence, 1947. Effort. Macdonald&Evans Ltd., London.

[11] R. Laban, 1963. Modern Educational Dance. Macdonald & Evans Ltd., London.

[12] J.-C. Martin, G. Caridakis, L. Devillers, K. Karpouzis, and S. Abrilian, 2006. Manual Annotation and Automatic Image Processing of Multimodal Emotional Behaviours: Validating the Annotation of TV Interviews, 5th Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, May 2006.

[13] B. Mazzarino, M.J. Peinado, R. Boulic, M. Wanderley and A. Camurri, 2006. Improving Human Movement Recovery Using Qualitative Analysis. In Proc. Intl. Conference Enactive 2006, Montpellier, France, November 2006.

[14] R. Picard, 1997. Affective Computing. Boston, MA: MIT Press.

[15] F.E. Pollick, 2004. The Features People Use to Recognize Human Movement Style, in A. Camurri, G. Volpe (Eds.), Gesture-based Communication in Human- Computer Interaction, LNAI 2915, pp. 20-39, Springer Verlag, 2004.

[16] P. Schaeffer, 1977. Traité des Objets Musicaux. Second Edition, Editions du Seuil, Paris, France, 1977.

[17] K.R. Scherer, and H.G. Wallbott, 1985. Analysis of Nonverbal Behavior. HANDBOOK OF DISCOURSE: ANALYSIS, Vol. 2, Cap.11, Academic Press London, 1985.

[18] K.R. Scherer, K.R., 2003. Why music does not produce basic emotions: pleading for a new approach to measuring the emotional effects of music. In Proc. Stockholm Music Acoustics Conference SMAC-03, 25-28, KTH, Stockholm, Sweden.

[19] H.G. Wallbott, and K.R. Scherer, 1986. Cues and Channels in Emotion Recognition. Journal of Personality and Social Psychology, 1986, Vol.51, No.4, 690-699.

[20] H.G. Wallbott, 1998. Bodily expression of emotion. European Journal of Social Psychology, Eur. J. Soc. Psychol. 28, 879-896.

Rate of speech and mental processes in emotional and cognitive regulation

Marco Tonti [1]

Abstract.

This study aims to investigate the relationship between emotional-cognitive regulation and rate of speech, proposing a connection between the levels of emotional and cognitive dynamics as expressed by a subject through the words used, and the speed of his or her utterance.

A single, completely transcribed psychotherapy session has been analyzed using the Therapeutic Cycle Model (TCM) in order to identify different patterns of emotional-cognitive regulation (Emotion-Abstraction Patterns — EAPs). The Rate of Speech (ROS) has been additionally computed on the text under analysis.

This study supports the idea that simultaneous high levels of emotional involvement and cognitive load would slow down the rate of speech.

This work offers a way to evaluate the emotional-cognitive regulation within a dialogue without the need of a transcript, using a value that could be derived by the physical emission of the speech, the ROS.

1 INTRODUCTION

The Therapeutic Cycle Model (TCM) [10] is a computer-aided text analytic tool which aims to investigate the emotional-cognitive regulation within the therapeutic process, making use of sessions' transcripts. While the TCM is rooted in the field of the research in psychotherapy, the underlying theoretical aspects are suitable for application in many different fields. Emotional-cognitive regulation is seen in terms of the dynamics of two fundamental psychological processes: emotion and cognition. On the basis of the frequency of respectively *emotional* and *abstract* words used by a subject the method is able to measure the intensity of the different patterns of cognitive-emotional regulation. This piece of information is of great value for psychotherapy research because it allows the identification of *insight* moments. The concept of insight is central in psychotherapy, because it represents the acquisition of new understanding from the patient [3]. From an information-processing point of view, it can be considered as an integration between emotional and cognitive processes, which allows a person to reflect upon his or her own emotions enhancing the therapeutic work.

The TCM requires the availability of transcripts, that must be produced by an human operator due to the limits of the automatic speech recognition systems. Moreover, the implementation of TCM requires a certain methodological sophistication. Is it possible to develop tools which allow the investigation of emotional-cognitive regulation using parameters different from those of TCM? Would these tools speed up the process of analyzing emotional-cognitive regulation, to the point of making it in real-time?

The present work aims at the development of a tool for a text-based analysis of a subject's rate of speech and explores the hypothesis of a relationship between different kinds of emotional regulation within a subject and his or her rate of speech. More exactly, that a simultaneous emotional expression and cognitive elaboration will be related to a slow down in the speech, respectively due to hesitations and to "computation" of what is going to be said.

After having briefly presented the theoretical assumptions of the TCM, will be presented a study which tries to establish an empirical relationship between different kinds of emotional-cognitive regulation and the rate of speech. The results will be finally discussed with reference to the theoretical assumptions of the TCM.

The paper then proposes a purely acoustical approach to the measurement of the rate of speech that could allow an easy integration into a system built to evaluate the internal (emotional and cognitive) state of a subject speaking in an environment or in a telephone call.

2 THE THERAPEUTIC CYCLE MODEL

The TCM theory was developed by Erhard Mergenthaler [7, 8, 9, 10] to give a formalization to the "good therapeutic hour" defined by Kris [4]. The definition of a "good hour" relies heavily on the phenomenon of insight, that can be described as a moment in which the subject is able to look inside himself or herself, to relive past events and, through a rational and cognitive work, to elaborate and overcome the often painful emotions connected to the event. An insight is thus an inextricable blend of the emotion of remembering, and of the conscious comprehension of the emerged memories.

Mergenthaler proposed a computational definition of insight. This definition grounds on the two main concepts of *emotion tone* and *abstraction* (meant as cognitive work) as described in the previous paragraph. In the TCM theory these values are distilled from the transcripts of psychotherapeutic sessions. Emotional and abstract words are considered as "markers" respectively of underlying emotional and cognitive parallel processes. Their frequency, over a congruous block of words, give a measure of the intensity of each process. "With the therapeutic cycle model [. . .] it is the first time that the clinical concept of emotion is brought together with the phenomenon of abstraction [. . .]. It is expected to allow one to operationalize and to measure the most important concept of psychoanalysis, emotional insight, in a transparent way." (Mergenthaler [10], page 1308).

The TCM theory has been applied mainly to the analysis of psychotherapies. Casonato & Gallo [2] for example applied the TCM to the transcripts of a TV interview with Donato Bilancia, an Italian serial killer. Even in such a peculiar context the result of the TCM

[1] Università di Bologna, Italy, email: tonti@cs.unibo.it

analysis is coherent with the psychiatric evaluation of the subject and with the theoretical assumption that the TCM offers a measures of the capability of the subject to keep connected emotional and reflective processes. Some studies have been conducted also for different fields, for example in relation to non-verbal behavior [6]. Still, the application of the TCM to a context different from the psychotherapeutic one is a point that needs to be evaluated more thoroughly from a psychological standpoint.

2.1 Emotional and abstract dictionaries

The dictionary of emotional words for the English language was taken from already compiled corpora[2]. The resulting words were revised in order to meet the following criteria:

- exclude concrete words referring to a sensory representation of emotions (e.g. “heart” or “warm”);
- exclude words that can not be classified in the following dimensions: pleasure / displeasure, approval / disapproval, attachment / disattachment, surprise;
- exclude from the dictionary the polysemic words, even if emotionally tinged (e.g. “like”, “mean”, “kind”, “well”).

These requirements were originated by the need of detaching concreteness from emotions (the reason will be clear in a while) and not to introduce biases due to the dependency of the polysemic words from their context. For other languages the words are added in the dictionaries by trained judges on the basis of the transcripts.

The dictionary of abstract words is much more simple to compile. The abstractness of a word is denoted by its suffix, for example “-ness” and “-ity”[3]. The mental work of producing an abstract word is considered as a marker for cognition.

2.2 The computation of indexes

The text of the transcript is segmented into *word blocks*. The size of a block must be large enough to allow statistical significance to the frequency of the contained words, compared to the natural frequency of emotional and abstract words. The size could vary with the language of the transcript, but its value is usually from 100 to 200 words. The blocks have the same constant size during the analysis, that usually is 150 words. If the last block size is less than the required length it is joined to the preceding block. The method counts the emotional and abstract words for each block, computing their frequency.

The TCM method studies the *variation* of the frequencies and their mutual meaning. Absolute values are not useful to this goal, so the values are standardized as Z-scores. The standardization allows a straightforward confrontation of the two values of word frequencies. From now on, the standardized frequency of the emotional and abstract indexes will be denoted, respectively, as ET (*Emotion Tone*) and AB (*Abstraction*). In the figure 1 it is possible to see a sample result of the elaboration of data.

2.3 Emotion-Abstraction patterns

The AB and ET values are floating-point numbers. Their fluctuation can be meaningful from a qualitative standpoint, but to simplify the analysis it would be useful to have *binary* values that can compactly denote the nature of each word block. This can be obtained observing that an ET or AB value greater than 0 means that the relative word block has a frequency of relevant words greater than the average of the whole transcript. This procedure allows the definition of labels that can be used to mark the nature of a block. The symbols that will be used are:

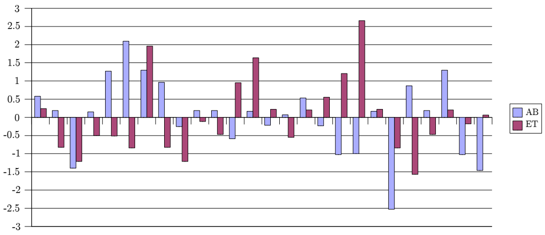


Figure 1. Sample graphic of a session. Word block = 150 words.

$$ET^- \Leftrightarrow ET < 0 \qquad ET^+ \Leftrightarrow ET \geq 0$$

$$AB^- \Leftrightarrow AB < 0 \qquad AB^+ \Leftrightarrow AB \geq 0$$

A word block marked with AB^+ and ET^+ is a container of an insight phase. But this operation generates a by-product: the other three combinations of the binary values. Each of these combinations are called an *Emotion-Abstraction Pattern* (EAP), and yields a precise psychological meaning. The four patterns are shown in the figure 2.

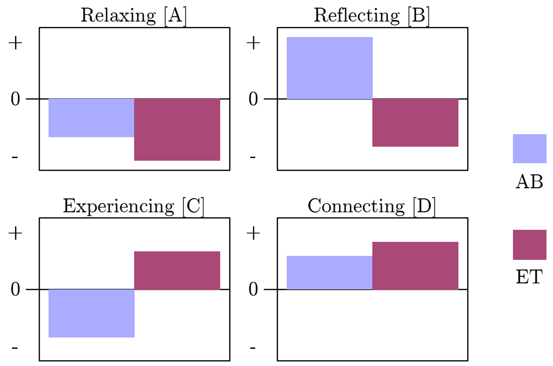


Figure 2. The four emotion-abstraction patterns.

Relaxing (A) AB^- and ET^- In this phase the patient talks about topics that are not connected to the issues for which he or she is in therapy. It is a phase of generic talking that represents a rest, a pause during the course of the session.

Reflecting (B) AB^+ and ET^- The topics offered by the patient are connotated by an high grade of abstract thinking. This may be the expression of an intellectualizing defense mechanism.

Experiencing (C) AB^- and ET^+ The patient is experiencing the emotions connected to past experiences, but without cognitive elaboration of them.

[2] The “emotion” category of the Regressive Imagery Dictionary by Martindale, and the Ulmer Textbank.

[3] The complete list of suffixes recognized by the TAS/C program is: “-ity”, “-ness”, “-nce”, “-ment”, “-any”, “-ncy”, “-ship”, “-dom”, “-ing”, “-ion”, and their plural forms.

Connecting (D) AB^{+} **and** ET^{+} While accessing the emotional experience of conflicting themes, the patient can reflect on the emerged material. This phase denotes the key-moment (insight) during the psychotherapeutic session.

2.4 Cycle of patterns

In the TCM theory it is assumed some kind of regularity in the sequence of patterns: "This is introduced as the *therapeutic cycle* model, consisting of five phases. It is based on the assumption that, in the course of a psychotherapy or within a psychotherapy session, emotion-abstraction patterns do not occur by chance: rather, a periodic process for the underlying variables emotion tone and abstraction is assumed. In explaining this, not only psychic but also biological factors may be taken into account (e.g., endorphins)." ([10], page 1308). The *prototypical cycle* proposed in the theory is shown in figure 3.

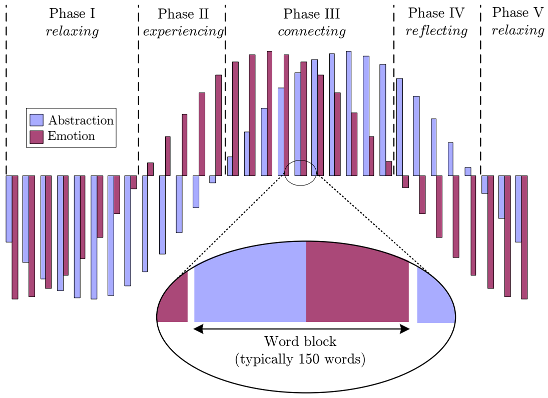


Figure 3. Prototypical cycle of emotion-abstraction patterns within the course of a psychotherapy session.

While the prototypical cycle is defined on the basis of deep theoretical reasons (Kris [4]), a full explanation of these reasons is not relevant for the sake of this document. Still the graphic will be used as a reference for the concepts being exposed forward.

3 TCM AND RATE OF SPEECH

Expression of thoughts pervaded by emotions is often dotted with hesitations. People hardly feel comfortable while expressing their emotions, and this difficulty is evidenced by prolongations of words, pauses etc. A similar effect is expected to take place for intensive cognition phases, but for different reasons: the person is thinking about what he or she is going to say. A person can hardly talk fluently and think at the same time.

On the basis of this consideration, it is reasonable to think that a variation of the rate of speech would correspond to a variation in the emotional and cognitive state of a subject, as measured by the TCM. In particular emotion and abstraction independently can influence the utterance, thus their contribution to the slowdown of speech is considered to be the simple sum of the two values. The grater the sum, the slower the speech.

Here is summarized the idea to be proven:

> **The rate of speech value varies consistently with the values of the Emotion and Abstraction indexes as defined in the TCM theory, presenting an inverse curve in respect to the sum of the** AB **and** ET **indexes.**
> **From another point of view, the rate of speech value is expected to be slower during the *Connecting* phase, faster during the *Relaxing* phase, and around the average during the *Experiencing* and the *Reflecting* phases.**

In figure 6 is shown a graphical representation of the idea exposed.

3.1 Defining the rate of speech

The first issue to face is to give a definition of what "Rate of speech" (ROS) is. Several indexes can be used to measure the ROS, words per second (WPS) or syllables per second (SPS) for example. Other measures can be derived from the phonetic analysis, as syllabic nuclei per second, or phones per second. If a transcript is available another measure could be characters per second (CPS).

The idea proposed is not easy to be proven, because it describes subtle and elusive phenomena. For this reason the first approach is to follow the simplest and most straightforward measures of ROS: words per second and character per second. These are easily computed using the transcripts needed for the TCM part of the study.

In future developments purely acoustical algorithms will be tested, but at this time the main goal is to give a sound ground to the idea that exists a relationship between the physical aspect of the utterance and its contents.

4 METHOD

The TCM theory is essentially based on the analysis of transcripts of psychotherapeutic sessions. Unfortunately a transcript does not hold the information about the *timing* of the utterance, which is exactly the kind of data needed to run the proposed analyses. To overcome this limit is necessary to extract that data from another source, and this source could not be anything else than the original audio recording from which the transcript has been written down.

The operation needed to obtain the timing of text is called *alignment*: the text must be somehow put in relation to the actual speech of the participants. For example, a sentence must be put between the time boundaries obtained from the recording. This allows the computation of the ROS values in terms of words per second, or characters per second. An example of alignment is given in figure 4.

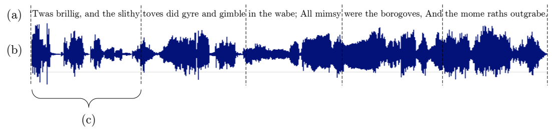


Figure 4. A sample alignment of text (a) and recording (b) into segments (c).

4.1 Segmentation

The rules given by the TCM to transcribe the sessions [11] are sometimes quite complex but they reflect the complexity of the natural language. In the transcripts the paraverbal elements must be transcribed (like *a-ah* and *eheh*), the areas of overlapping speech, the silences, the notes not belonging to the utterance, the pauses, the incomprehensible words, the prolongation of final vowels must be marked .

Many of these elements must be taken into account while aligning the recording to the corresponding text.

The recording is split into small segments lasting from 0.5 to 2 seconds, each one containing from 1 to 6 words. This offers a great flexibility that is necessary for the forthcoming analyses. As already said a TCM word block is usually composed by 150 words, but this number is somehow arbitrarily defined, and should not be taken as a fixed value. The word block needed for the TCM part is obtained by compounding as many segments as necessary to reach the required dimension. This can be done for any word block dimension, giving a great flexibility in the statistics to be exposed.

On the other hand, such a thin segmentation of the recording allows a very precise evaluation of the rate of speech. The ROS computed in this way is suitable for a large scale analysis based on the whole word block, and also for more subtle analyses that will be exposed in detail more forward. The operation of splitting the recording into segments, and of the computation of the ROS for each segment, is shown in figure 5.

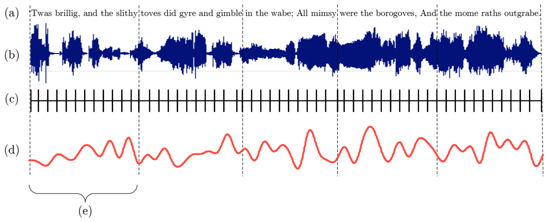

(a)
Twas brillig, and the slithy toves did gyre and gimble in the wabe; All mimsy were the borogoves, And the mome raths outgrabe.
(b)
(c)
(d)
(e)

Figure 5. Example of the process of segmentation. (a) Transcript (b) Waveform of the recording (c) Sequence of the segments (0.5–2 seconds and 1–6 words each) (d) Value of the ROS for each segment (e) TCM word block.

The software tool employed to align the transcript to the recording is Transcriber[4].

4.2 Nonparametric correlation

The first approach to the validation of the idea is to put in correlation (a negative correlation) the values extracted using the TCM technique with the rate of speech. The expected rate of speech should be lower when the values of abstraction and emotion are both high, and *vice versa*. The graphic 6 shows the idea.

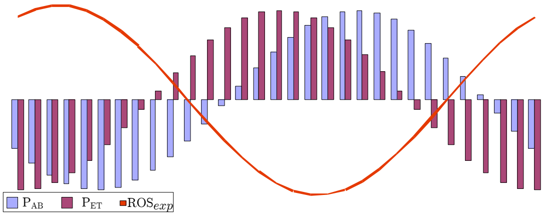


Figure 6. Expected overall rate of speech (ROS)

Each pair of bars represent a word block (recall figure 3) and the continuous line the expected rate of speech, also standardized as Z-scored values. A statistic should reveal a negative correlation between the rate of speech and the sum of the two values of abstraction and emotion.

A big drawback of the correlation approach is that it tends to mask the micro-variation of the ROS value, allowing just for a coarse-grained analysis. This happens because the ROS value is computed over a large block of words, and not at the segment level. Another issue is that for each psychotherapeutic session the number of word blocks is often quite small (20–30 depending on the actual size of the word block) and this makes difficult to obtain very significant results.

4.3 Disaggregation of word blocks

A technique that revealed itself very useful is the one here called "disaggregation", that exploits the structure employed so far. The TCM value of a word block (but here should be considered not the numerical value of the indexes, but rather the pattern that they identify — see figure 2) is based on the aggregation of many small segments, thus is possible to infer that *each one* of these segments holds the same property of the word block in which is contained. In other words, the segments composing the word block inherit the Emotion-Abstraction Pattern (EAP) defined at word block level. The idea is represented in figure 7.

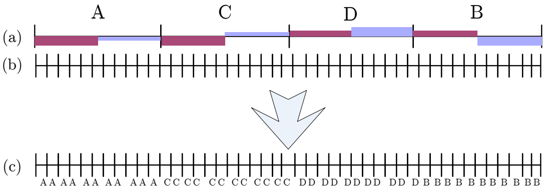

A
C
D
B
(a)
(b)
(c)
AA AA AA AA AA A CC CC CC CC CC CC DD DD DD DD DD D B B B B B B B B B BB

Figure 7. Process of disaggregation of the word blocks. (a) Some word blocks with the pattern as defined by the TCM (b) the segments, obtained from the alignment, composing the blocks (c) the segments "labeled" with the emotion-abstraction pattern of the containing word block.

This technique permits to downsize the word block, making it of the dimension of the small 2-seconds segments. This offers a double positive outcome: it allows for a fine-grained analysis of the rate of speech, and gives a much larger data pool on which compute the statistics.

4.4 Kruskal-Wallis test

Using the disaggregation technique, a second statistical approach can be used. This is done by grouping the small segments (now marked with the corresponding Emotion-Abstraction Pattern derived from the containing word block) into three sets, defined after the hypothesis: High, Medium and Low rate of speech. In the High set are grouped the segments belonging to Relaxing blocks (A), in the Medium the blocks from Experiencing and Reflecting phases (B and C), and in the Low set the segments from Connecting phases (D).

If the hypothesis is correct, the average value of these sets should be significantly different, following the appropriate order (High $>$ Medium $>$ Low).

4.5 Variation of ROS

Another possible hypothesis to test could be the one connected to the elaboration needed to think about what to say, distinct by the actual act of saying that. If a phrase is "ready" to be uttered, probably the

[4] http://trans.sourceforge.net

utterance would be more smooth, without great variations in the rate of speech due to hesitations or elaboration.

This hypothesis can be studied through the use of the *coefficient of variation*, that is defined as

$$\mathrm{VarCoeff}(x) = \frac{\sigma_x}{\bar{x}}$$

The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from each other. This value is often reported as a percentage, by multiplying it by one hundred. An higher value of this coefficient means a greater variance of the data in the set.

The idea here is that while there is a process computing *what to say*, that process slows down the rate of speech, but when the content is uttered this should diminish the variation of the ROS, because the contents are already elaborated. This idea is illustrated in figure 8 with separated values for ROS and VarCoeff, while in figure 9 the two effects are put together in the same curve.

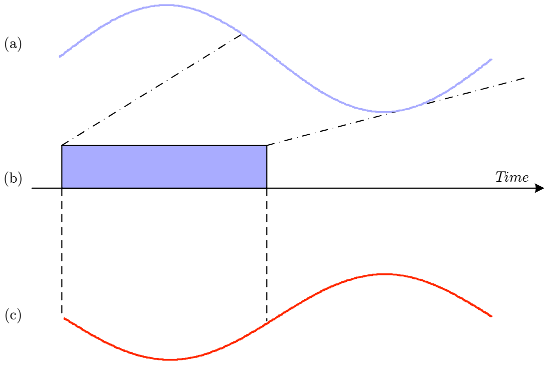


Figure 8. (a) The amount of variation of the rate of speech (b) the process of elaboration (c) the averaged rate of speech.

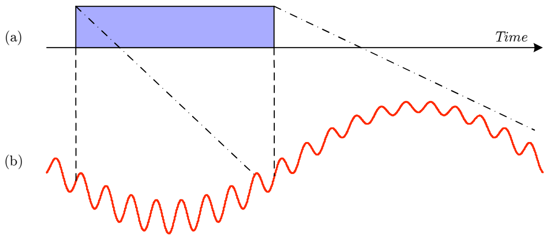


Figure 9. (a) the process computing the contents to be said (b) the expected effects of the computation on the actual rate of speech.

The processes of elaboration are of course always present during the speech, but for the TCM theory they are more frequent (or more intense) in the blocks where the AB value is positive. The process depicted in the figures ideally should last for no more than one or two seconds, and the amount of abstraction should be connected to the frequency and the intensity of the processes occurring in that phase.

5 RESULTS

This research consists of a pilot case study based on only one psychotherapeutic session. The patient is aliased Romina, she is an Italian 36 years-old woman in therapy for eating disorders. The therapy is a psychodynamic therapy continuing for about 8 months, one session per week. The psychotherapist is the psychoanalyst Roberto De Ponte Conti. This session is part of a study exposed in Casonato & Gallo [2], to which the reader is forwarded for further details.

The chosen session has been the 4th one, because of a low grade of interruptions (the therapist is a "verbally active" one) and the meta-analysis, described in [2], evidenced that this session is characterized by a prevalence of the *Connecting* (D) pattern. Furthermore in this session the subject describes her feelings about the death of her father, of which the day of the session is the anniversary.

The silences longer that 0.5 seconds have been totally excluded from the computation of the rate of speech.

5.1 Correlation

In graphic 10 is shown the rate of speech (in WPS, words per second) compared to the sum of AB and ET. The word block size in this case is 159 words. The Spearman's correlation is -0.71 which is highly significant ($p < 0.0001$).

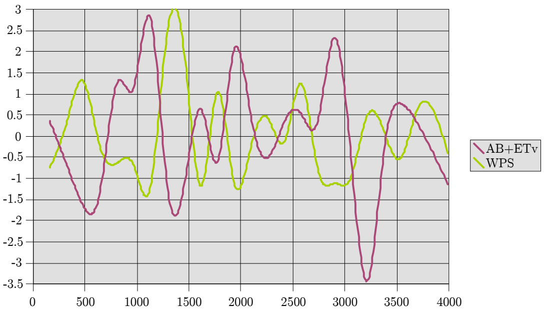


Figure 10. AB+ET and WPS. The WPS value has been Z-scored to fit better into the graphic. The X axis represents the words actually spoken by the subject. Must be remarked that in the first half of the figure the two graphics have almost the very same symmetric shape.

Significant correlation results have been obtained for almost every size of word block between 100 and 200 words.

5.2 Kruskal-Wallis test

For a word block made by 175 words, and measuring the ROS (in CPS), the Kruskal-Wallis test evaluates the difference in the means of the of the three groups of segments, divided by the expected ROS (High, Medium and Low). Those means are different with a $\chi^2 = 11.56, df = 2, p = 0.003$ which is very significant being lower than 0.01 . Furthermore the means of the three sets are as expected $15.19(H) > 14.64(M) > 13.53(L)$.

5.3 Coefficient of variation

The statistical results support the idea proposed in section 4.5, because the *minimum* difference (over any word block size) between the VarCoefficients for high and low abstraction phases if 10%. (In high abstraction phases there is less variability)

Must be remarked that this result hold only for characters per second, probably because of the particular numerical structure of the ROS in terms of words per second computed on the segments: the number of words is quite restricted (from 1 to 6) and also the time dimension of a segment is bounded to about a second and a half. Probably the values of WPS do not have a great variability for this reason: their values are often quite similar.

6 DISCUSSION

The results of the statistical tests, although for a pilot case study, tend to confirm the hypothesized connection between emotional/cognitive regulation and the rate of speech. The statistics followed three different approaches to the problem. The first one confirmed a highly significant negative correlation between the ET+AB value (Emotion+Abstraction) and the rate of speech at word blocks level. This means that when ET+AB increases, the rate of speech decreases.

The second test grouped the segments composing the different Emotion/Abstraction patterns. The three sets showed a very significant difference in their mean values, and the values of the three sets are coherent with the hypothesis. In particular, for example, the mean ROS of the set grouping the segments belonging to *Connecting* phases is lower then the the mean ROS of the other sets, as expected.

The third approach is similar to the second one, but measures how much the rate of speech varies (on a segment basis) in function of the Abstraction value of the utterance. Also this third test confirmed a difference in the values, depending on the cognitive load of the subject. When the subject is an high cognitive moments, the variation of the rate of speech is 10% lower.

These three results support the proposed hypothesis from three different standpoints.

7 ACOUSTICAL COMPUTATION OF ROS

Many algorithms are commonly used to measure the rate of speech on a purely acoustical basis. During the development of this research two of them have been employed to test the validity of the hypothesis without the need of a transcript. These two algorithms are Mermelstein's [12] (based on syllabic nuclei) and Andre-Obrecht's [1] (detecting the phones). Unfortunately, while the phenomenon seems to be verified for text-based measures of ROS, the use of these algorithms did not bear the expected results. Possibly the main reason is that the quality of the recording employed was very low (it was taken using a microphone embedded in a camera).

Further studies are in progress, but once proved the main connection between TCM and ROS this should be a matter of technical refinement.

8 CONCLUSION

The ideas expressed in this work are useful for affective environments under more than one profile. First, the application of the technique developed for the TCM offers a deep insight into the internal dynamics of a person following a sound psychological approach. Second, the study presented a method to overcome the difficulty introduced by the TCM about the needing of the transcript of the spoken words. Once stated and proved a precise relationship between rate of speech (and its variation) and TCM, techniques based only on acoustical methods could be developed and employed. This relationship is supported by the statistical results achieved so far, even if only for a single session pilot study. An ideal application could be to analyze the voice flow in a telephonic conversation.

This technique together with other studies connected to the detection of emotions (see for example Sarracino & Campanelli *et al.* [13] and Laukka, Juslin & Bresin [5]) could offer a simple yet powerful theoretical background to the interpretation of acoustical cues produced by a speaker.

ACKNOWLEDGEMENTS

I would like to thank Prof. Dr. Erhard Mergenthaler, Dr. Fabio Tamburini and Dr. Marco Casonato for their invaluable support in the development of this study. Thanks also to Omar Gelo for his precious hints.

REFERENCES

[1] R. Andre-Obrecht, 'A new statistical approach for the automatic segmentation of continuous speech signals', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **36**(1), 29–40, (1988).

[2] M. Casonato and I.F. Gallo, *L'analisi computerizzata del processo terapeutico*, QuattroVenti, Urbino, 2005.

[3] *Insight in psychotherapy*, eds., L.G. Castonguay and C.E. Hill, American Psychological Association, 2006.

[4] E. Kris, 'On some vicissitudes of insight in psychoanalysis', *International Journal of Psychoanalysis*, **37**, 445–455, (1956).

[5] P. Laukka, P.N. Juslin, and R. Bresin, 'A dimensional approach to vocal expression of emotion', *Cognition and emotion*, **19**(5), 633–653, (2005).

[6] L. Di Marino and E. Mergenthaler, 'Change processes and states of mind in a systemic family therapy', in *Proceedings of the 34th SPR International Meeting, Weimar, Germany*, (June 2003).

[7] E. Mergenthaler, *Textbank Systems: Computer Science Applied in the Field of Psychioanalysis*, Heidelberg: Springer-Verlag, 1985.

[8] E. Mergenthaler, 'Emotion/abstractness as indicators of "hot spots" in psychotherapy transcripts', in *Society for Psychotherapy Research 23 rd "Annual International Meeting", Berkeley, CA.*, (1992).

[9] E. Mergenthaler, 'Computer-assisted content analysis', in *ZUMA — Nachrichten Spezial*, eds., C. Zuell, J. Harkness, and J.H.P. Hoffmeyer, ZUMA, Mannheim., (1996).

[10] E. Mergenthaler, 'Emotion-abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes', *Journal of Consulting and Clinical Psychology*, **6**(64), 1306–1315, (1996).

[11] E. Mergenthaler and C. Stinson, 'Psychotherapy transcription standards', *Psychotherapy Research*, **2**(2), 125–142, (1992).

[12] P. Mermelstein, 'Automatic segmentation of speech into syllabic units', *Journal of the Acoustic Society of America*, **58**(4), 880–883, (1975).

[13] D. Sarracino, L. Campanelli, E. Iberni, S. Degni, and A. De Coro, 'Acoustic analysis of psychotherapist and patient speech during the psychotherapy: a preliminary study'. 37th annual meeting of the Society for Psychotherapy Research. Edinburgh 21–24 June 2006.

[14] F. Tamburini, *Fenomeni prosodici e prominenza: un approccio acustico*, Bologna: Bononia University Press, 2005.

[15] M. Tonti, 'The influence of emotional and cognitive processes in the definition of speech rate'. 37th annual meeting of the Society for Psychotherapy Research. Edinburgh 21–24 June 2006.

An adaptive rule-based inference engine for realising reasonable behaviour of smart environments

Michael Hellenschmidt [1]

Abstract. This paper presents the implementation of an inference engine for modelling reliable behaviour of smart environments. Here especially user's individual variables are taken into consideration. By discussing example scenarios in smart meeting rooms the requirements for such an inference engine are given and its implementation as well as its underlying semantical principles is discussed in detail. The inference process here establishes some novel metaphors for describing reasonable behaviour with rules: the consideration of event sequences and thus the consideration of subsequent user activities as well as the possibility to freely define variables that should be considered during the inference processes. This kind of adaptiveness makes the adaptation of the system to various application scenarios possible and allows an individualization of smart environment's behaviour.

1 Introduction

The concepts of Ambient Intelligence represent a vision of devices and environments that are able to act in a smart and reasonable fashion with a special respect to the user's current activities. Especially in the domain of intelligent houses, smart living rooms or smart conference rooms (resp. bureau environments) remarkable results have been achieved in the past. Prominent examples are the results of the Easy Living project of Microsoft [1], the results of the Aware Home project of the Georgia Institute of Technology [2] or the Intelligent Room in the context of the Oxygen-initiative of the Massachusetts Institute of Technology [3, 4]. The living room that is realized in the Easy Living project is able to control and coordinate different input and output devices like keyboards, computer monitors and projectors by using the information about the current user and her position. This is done by two applications called ”World Model” and ”Rule Engine”. The World Model application is collecting all information that is available by the input components whereas the Rule Engine is inferring which devices should be used for which tasks. The implementations of the Aware Home and the Intelligent Room are bearing some resemblances to this approach. Here, a ”Room Manager” and a monolithic controller respectively are used. Michael Coen, one of the developers of the intelligent room, calls this controller a ”big messy C program” [5] because this program was - at a very early stage of development - too complex to allow a fast integration of new environment functions or even a change of system behaviour. This notice can probably be made for the other well-known implementations as well. Another aspect is obviously the missing link to individual user needs or user's current emotional status. This kind of affective and adaptive behaviour is often impossible because of the usage of static and/or generalized behaviour rules. Often all kind of users are treated equally or when users are treated individually their emotional status is assumed as ”neutral”.

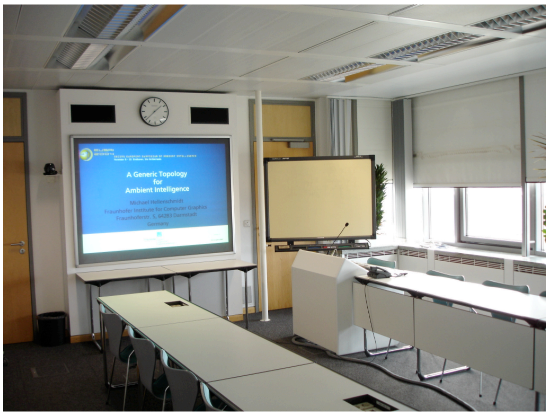


Figure 1. Typical environment we want to be smart. Lights and the shutter are adjusted with respect to the given situation, devices like projectors are controlled and adaptors are switched in a reasonable fashion.

Typical application scenarios for smart environments are combining observations of the user's interaction with general context information to infer reasonable behaviour of the devices that are present and available. Here, the user's interaction could be implicit as well as explicit [6]. An example is a meeting room, where a user starts a presentation on her personal laptop and goes to the presentation board and the standing desk. The room reacts by switching the projector on, activating the VGA-adaptor that corresponds to the user's laptop position, switching the microphone of the standing desk and adjusting the lights and the shutter with respect to the current time and the current lighting conditions (see figure 1). But this kind of proactive room behaviour should be user specific. Furthermore it should be adaptive to the user's emotional status (e.g. stress, impatience) and her individual preferences (e.g. wants to have ambient lights) or handicaps.

In this paper, we discuss the realization of an inference engine as well as the setup of a meeting room that allows experiments considering the reasonable behaviour of the room's internal devices. The further structure of this paper is as follows: In the following section, we define our requirements for smart and affective room behaviour and for our realization of an inference engine for the interpretation of user's interactions and the environment's variables with respect to individual user needs and preferences. After that our implementation is described in detail. Here, the underlying methodology as well as the inference technology itself and its embedment within an autonomous agent are explained. In section 4 some examples are given and our

[1] Fraunhofer-Institute for Computer Graphics, Fraunhoferstr. 5, 64283 Darmstadt, Germany, email: michael.hellenschmidt@igd.fraunhofer.de

experimental setup of a meeting room is illustrated. Finally we give a comparison of our approach with other activities and technologies and outline the next steps and the future work.

2 Requirements

Imagine a participant of a meeting starts a presentation on her own personal laptop. After that he passes some pressure plates (installed in the floor of the room) that indicate his way from the seat to the standing desk. Obviously he wants to present some content by using the room's devices, such as projector, microphones and loudspeakers. In order to infer reasonable room behaviour not only context information but also information about user's interaction (either explicit or implicit) has to be considered. Context information can be defined in accordance with the "What, Who, Where, When and How"-methodology that can be found in [7]. User's interaction can not only be a single event (like pushing a button) but also a concrete sequence of events. In our example scenario it makes quite a difference whether a participant stands up and walks to a window (maybe to open or close it) or he stands up and walks to the standing desk (in order to give a presentation). Here the event sequence that indicates a walk from a certain seat to the standing desk determines the goal of a user that a presentation from exactly the laptop that corresponds to the user's seat should be used for giving the presentation and that the room should be technically prepared for an appropriate presentation ambience. Also the current mood of the user should affect room's behaviour. Thus scenarios that look similar at first sight have to be treated differently with respect to user's individual preferences resp. user's current mood and feelings. Another important criterion has to be considered for any inference processes: that is the addressing of functions that determine the room behaviour. Different approaches are possible: functions that are totally independent from each other (for instance, a person enters a room would mean to open the shutters AND to switch on the room lights) or functions that depend on each other (for instance, a person walks to the presentation desk would mean to dim the room lights first, then to switch on the microphone and finally to start the presentation by using the vga-adaptor that corresponds to the user's seat). Obviously in situations where the execution of functions is dependent on each other, a cascade of functions has to be cancelled if one function is not executable. In order to realise this kind of proactive behaviour of the environment we found the following requirements for implementing an inference mechanisms for smart environments is:

- Environment variables like context information (e.g. time, day, number of persons in the room, lighting status, etc) and user information (e.g. identification number, preferences, mood, name, etc.)
- and user interactions - both in explicit and in implicit fashion - as single events or sequences of events have to be considered during the inference process.
- Functions (or cascades of functions) should be definable that should be executed if a certain definition of environment variables and/or user interaction events are evaluated to true.

And finally the configuration of the room - and thus the intelligence - should be feasible in an easy and intuitive fashion. This means that a developer of a smart environment should edit the environment variables, user status variables and the event sequences that should be true for a certain environment behaviour[2].

3 Inference Technology

This section describes in detail the realized inference engine and the fundamental definitions of the environment's variables as well as the underlying inference mechanisms the syntax of the rules, which can be freely defined.

3.1 The Environment

The environment we consider in the inference process is defined by an object that consists of both information about context variables and user variables. Context as well as users are syntactically described in the same fashion. An environment variable is defined by a variable name (could be considered as a key) and a variable value. The variable value is furthermore defined by its data type. Values are Strings, Numbers, or Sets.

Some examples for context variables are:

```
("time", Num 43200)
("uid", Num 768)
("season", String "summer")
("lighting_condition", String "dim")
("persons", Set [Num 768,
          Num 123, Num 345])
```

These variables indicate that it is 12 o'clock in the noon, the active person is the one with the identification number 768, it is summer but the lighting conditions are dim and the persons in the room are those with the IDs 768, 123, and 345.

Some examples for user variables might be:

```
[(768,[("name", String "John Doe")
     ,("age", Num 26)
     ,("gender",String "male")
     ,("expert", Set [String "Java",
          String "SQL", String "AI"])
     ,("mission", String "professional")
     ,("mood"', Set [String "busy",
      String "stressed"])
     ])
,(123,[("name", String "Frank Miller")
     ,("age", Num 35)
     ...
```

These variables indicate that the person with the identification number 768 is John Doe. He is an expert in different computer science technologies like Java or Artificial Intelligence. His current activities are job-related and he seems to be in a busy and stressed mood. In the same way an unlimited number of other persons can be described. Because the inference engine itself holds the information about the environment's variables it is possible at any time - while the system is running - to extend the set of variables by defining new key-value-pairs (see section 3.3). Those key-value-pairs are defined, for instance, by input applications and are forwarded to the inference engine that includes this new information into its own persistent data memory. Also the modification of existing variables (e.g. the current active person, or the mood of individual persons) is possible. Consequently all context and user variables can be defined freely. Possible variables for describing users within a multi-modal environment can be found in [8].

[2] Here it is obvious that our smart environment should not behave intelligently in terms of artificial intelligence, but reasonable and affective in terms of rationality and know-how of the responsible developers.

3.2 The Inference Engine

The configuration of reasonable system behaviour, which is affective with respect to individual users, their individual preferences and moods, in terms of rules seems to be most appropriate when considering the given application scenarios. Thus we decided to implement an inference engine that allows the definition of rules in an ”if-then”-alike fashion. The most important definitions for defining a rule set are:

- a rule is composed of a rule name, a conditional part (left side) and an action part (right side).
- the conditional part is defined as a list of expressions that is separated by an AND- or an OR-conjunction.
- the action part is defined by function identifier(s) that are itemised in the same way they are executed. This kind of interdependence of functions is indicated by the separation element →.
- an expression is either a boolean expression (e.g. user.mission $==$ ”private”) or a relation expression (e.g. context.time $<$ 12:00:00) or the indication of a sequence of events that should occur in a subsequent order.
- for relation expressions it is possible to use $<$, $<=$, $>$, $>=$, $\sim$, and $! \sim$. The relation $user.expert \sim "Java"$ thus means that the user’s variable expert contains the term Java (among other things). Boolean expressions are indicated by $==$ and $! =$.
- a sequence of events that should occur that a rule is evaluated to true is syntactically written in the form $sequence(a, b, c, ...)$ where a,b, and c are representing the value of events that have happened. The number of arguments is not limited.
- finally a function that should be executed can be defined in the action part of a rule in the form $func(functionname(arg1, arg2, arg3, ...))$ where function name indicates the function that should be executed. Additionally a list of arguments can be assigned to the function call. Consequently it is possible to define function calls like ”switch(context.time, on)”. Here context.time will be automatically exchanged by the value of the context variable time when the action part is executed.

The left part of a rule is either evaluated to

- $true$, if all boolean expressions resp. relation expressions are evaluated to true and sequences that are defined within a rule are completed.
- $false$, if any boolean expression resp. relation expression is evaluated to false or any sequence that is defined within a rule is failed.
- $hold$, if all boolean expressions resp. relation expressions are evaluated to true, but any sequence that is defined within a rule is neither completed nor failed.

The third possible result of an inference process should be described in a more detailed fashion: For instance, a sequence defines the way of a user from a certain seat to the standing desk. Going this way the user has to pass three different floor plates that emit the events bp1 (for base plate 1), bp2, and bp3. This means the $sequence(bp1, bp2, bp3)$ has to be fulfilled. Consequently after the event $bp1$ is emitted a rule that contains this sequence is neither true nor false. Not until the next event(s) is (are) emitted is can decided whether the function part of the rule has to be executed or not. This kind of rule interstage is internally handled by the inference engine by indicating the rule as *hold*. We implemented the inference engine by using a Yacc-compiler (stands for yet another compiler compiler, see [9]). With Yacc parsers can be implemented that are based on an analytic grammar. Such kind of parsers are able to analyse syntactical structures and are therefore able to build up object trees that can be evaluated in further processing steps. Consequently the inference engine that is described here is not fixed to a certain set of rules but only to a defined syntactical grammar. The rules that adhere strictly to this specified syntax are edited in text files. The inference engine is interpreting the rules at system run time, whereas the number of rules that can be evaluated is in principle unlimited. For the implementation we used the BYACC/J that is available from [10].

3.3 The Inference Agent

We realized an autonomous application that holds the inference engine (see section 3.2) as well as the internal representation of the environment’s variables (see section 3.1). This application is implemented as an agent that is able to communicate with other applications. Consequently the inference agent is able to get information about context and user variables that are forwarded by applications like input components or context management applications and integrates them into its own internal environment representations. Events are received in the same way. After getting a context variable (resp. user variable or event) it will be included into the internal environment representation. The inference process of all rules starts automatically after each incoming value with respect to the dynamical behaviour of the current environment setting. The result of an inference process is a list of rules that are evaluated to true. Here the resulting function calls (right side of the rules) are extracted and sent as a message to those applications that are responsible for executing the room’s and the devices’ functions. The inference agent is additionally responsible for handling interdependent functions. Here, the inference agent waits till a function is successfully executed and only then sends the subsequent function calls. If a function call fails the inference agent cancels the following calls. For handling the communication processes we applied the software infrastructure that is described in [11]. This software infrastructure is able to handle an undetermined number of agents in a self-organized fashion by applying certain conflict-resolution mechanisms while forwarding messages from one agent to another. Thus the receiver of a message is determined by the underlying communication mechanisms itself. This feature releases the inference agent from finding appropriate receiver applications while sending a function call.

4 Examples

This section gives some examples and illustrates how rules are syntactically described as well as how rules can be defined that are specific with regards to a given application scenario. Furthermore the examples will be extended to demonstrate the way attentive and affective rules can be defined for both individual users and their current situation.

```
Rule 1:
sequence(bp1,bp2,bp3)
     & context.time > 11:59:00
=> func(projector(on))
     -> func(roomlights(0.5))
     -> func(shutter(0.5))
     -> func(vga_source(user.seat));

Rule 2:
sequence(bp1,bp2,bp3)
```

```
     & user.mission == professional
     & context.time > 11:59:00
=> func(projector(on))
     -> func(roomlights(0.0))
     -> func(frontlights(0.5))
     -> func(shutter(0.6))
     -> func(vga_source(user.seat));
```

Rule 1 defines that four functions should be scheduled, if the event sequence bp1, bp2 and bp3 occurs and the time is past noon. The first function switches on the projector, then the room lights should be dimmed with a value of 0.5, then the shutters should be lowered and after that the vga-source (for the projector) should be set to transmit the signal from the laptop connector that belongs to the user's seat. If the inference agent (by means of the inference engine) evaluates this rule to true, it will send the first function call as a message to the underlying device applications, will wait for the answer (either this function is successfully executed or not) and will then continue with the subsequent function calls.

Rule 2 illustrates how the first rule can be specified in more detail if the user's current activities are job-related. In a professional situation the room should react in contrast to rule 1 by switching off the room lights, and switching on the front lights. The expression that relates to the context variable time illustrates that rules can be defined with respect to any granularity that is appropriate. The reference to the specific user identification number is necessary if rules should be adapted individually to specific persons. This is the case if different persons are assumed to expect different room behaviour even though they are in the same situation. One notice: the current implementation of the inference engine only takes the current active user into consideration, whose identification number is tagged in the environment's variables[3].

```
Rule 3:
sequence(bp1,bp2,bp3)
     & user.id == 768
     & user.mission == professional
     & user.mood == stressed
     & context.numberOfpersons > 5
=> func(projector(on))
     -> func(roomlights(0.0))
     -> func(shutter(0.6))
     -> func(vga_source(user.seat));

Rule 4:
sequence(bp1,bp2,bp3)
     & user.id == 123
     & user.mission == professional
     & user.prefs ~ ambient
     & context.numberOfpersons > 5
=> func(projector(on))
     -> func(microphone(on))
     -> func(roomlights(0.1))
     -> func(shutter(0.2))
     -> func(vga_source(user.seat));
```

Rule 3 and 4 bear resemblances to rule 1 and rule 2 but with slight differences. Modifications of some relation expressions make the adaptation of system behaviour to any conceivable environment variables possible. Rule 3 and 4 illustrates how room behaviour can be adjusted according to the different needs of different users. The following rules demonstrate the application of rules for the person considering her different emotional conditions at very different situations.

```
Rule 5:
user.mission == professional
     & user.mood == stressed
     & context.time < 16:00:00
=> func(projector(on))
     -> ...
     // prepare professional ambience here

Rule 6:
user.mission == private
     & context.time > 16:00:00
=> func(projector(on))
     -> ...
     // prepare private ambience for
     // home cinema here
```

Rule 5 and rule 6 illustrate how the same room is able to provide different application sets considering possible different situation of the user. Configuring rules in this respect makes the adaptation of system behaviour in an attentive and adaptive fashion possible.

4.1 Experimental Setup

For the realisation of a smart meeting room (see figure 1) we implemented applications according to the component architecture illustrated in figure 2. The component structure follows the input-interpretation-execution metaphor that is described in a more detailed fashion in [12]. The room offers a server that understands commands, which are sent using the http-protocol. This makes it possible to control all devices that are available in the room. These are: the lightings in the front of the room, the room lightings, the projector, four different VGA-adaptors on the tables and the shutter. Consequently we implemented different applications (device agents) that are able to control the devices' functions by using the appropriate http-protocol commands. The different device applications are offer-

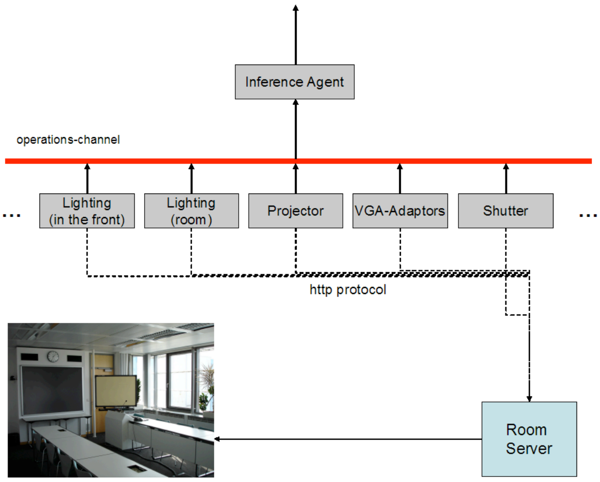


Figure 2. Illustration of the component architecture of our experimental setup of a smart meeting room.

[3] The inference engine described here does not specify the sources of environment's variables like context and user data. Any environment information that is represented by a key-value-pair as illustrated in section 3.1 can be used for the inference process.

ing certain function calls that are representing their internal abilities. For instance, the agent that is controlling the room lights offers the function $roomlights(int \quad intensity)$ whereas the vga-adaptor agent offers the function $vga\_source(int \quad x)$ with $0 < x < 5$. These functions can be used for defining the rules that determine the meeting room's smart behaviour. After a successful inference process the inference agent (see figure 2) sends the function calls to the level of device agents. Please consider that the direction of the connections of the different agents as well as the message bus follows the definitions of the underlying software infrastructure that were made in [11, 12]). Figure 2 does not present the level of input components that are needed for gathering environment variables and events that occur by implicit and explicit user interaction. Because the installation of any physical sensors (e.g. pressure plates, buttons, temperature sensors, etc.) is expensive - and once installed not easy to alter and to exchange - we decided to implement a three-dimensional visualization (see figure 3 of the room to simulate user events).

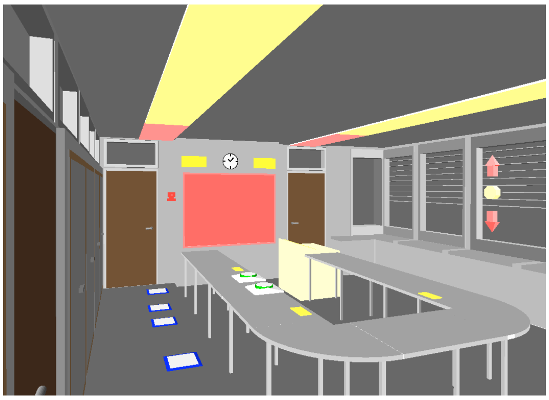


Figure 3. Three-dimensional simulation of our smart meeting room with different virtual installations of pressure plates on the floor and some virtual buttons on the tables.

This simulation allows the virtual installation of different kinds of event sources (buttons, pressure plates and magnet switches) virtually. Activated with the mouse they emit a freely definable event that can be read by the inference agent and be processed by its inference engine. In the same way other applications for defining context and user variables are realized. Consequently this arrangement makes it possible to experiment with different rule sets and the inference mechanisms without the need for expensive physical room installations. Figure 1 shows the room behaviour while using the experimental setup as described here and executing rule 1 that is defined in this section.

5 Related work

The realisation of intelligent behaviour came into the focus of scientific work again with the advent of Ambient Intelligence and the increasing efforts to realize the setup of smart environments[4]. There are the technologies that are based on connectionism and heuristics. These technologies, e.g. neural networks are used for stationary optimisation problems. Current works are concentrating on the combination of neural networks with technologies for semantic knowledge representation to model intelligent behaviour [13]. There are techniques that are based on symbolic approaches (e.g. planning systems) to solve given problems in a recursive fashion (in general by back-propagation mechanisms) or by applying a (stationary) rule set (e.g. expert systems). Expert systems are executing a so-called forward interpretation of context data. Well-known examples are the OPS5-system [14], the Clips-system that is based on LISP and the fast-forward planning-system of Hoffmann and Nebel [15]. The JESS - Java Expert System Shell is a Java-version of Clips. According to [16] this implementation of Clips was used for the realisation of the Intelligent Room that was discussed in this paper at the beginning. In contrast to Clips JESS is using an extended Rete-algorithm that was first described in [17]. Based on these principle technologies other approaches are known. The CommonRules-system [18] allows the definition of prioritisations of rules that means that rules can overrule other rules. The KGP-model (for knowledge, goals and plans) establishes a two-stage inference mechanisms [19]. After the first inference process possible goals are inferred from the current environment status. These goals are further processed to concrete plans which functions have to be used for changing the environments. Also technologies that are using Bayesian networks are already published [20].

The interactive Context-aware Application Prototyper (iCAP) [21] is a system that allows end-user to design rules by using graphical interfaces. This system supports the definition of rules that corresponds to if-else-commands and to relationship-based actions (e.g. ”If I leave the house, turn off the lights.”). iCAP concentrates on the realisation of a system that offers the users the possibility to configure it without the necessity to write any code. But it seems that more complex scenarios are not in the focus of this work. Rather in contrast to the concept of ”goal based interaction” that is presented by Heider and Kirste [22]. Here - similar to the KGP approach - the inference process is divided into the ”intention analysis” part and the ”strategy planning” part. The technique generates strategies automatically to achieve user-defined goals. Goals are declarative represented by using positive and negative literals. Also the current environment conditions are described as a set of literals as well as the available device and environment functions. In principle this method is based on a typical planning system on top of a back-propagation mechanism. In contrast to iCAP this technology can be used for the implementation of very complex scenarios. Also dynamic scenarios - like adding and removing of device functions - seem to be possible.

Recapitulating the technologies for the interpretation of environment variables to infer reasonable system behaviour three different observations can be made. Very powerful techniques that allow the modelling of a wide variety of scenarios are not applicable (or even modifiable or editable) in an easy fashion. Those technologies are mainly meant for self-organizing scenarios. This means that the user is not expected to take corrective action in a running system (e.g. back-propagation techniques, techniques based on Bayesian networks). Other techniques allow explicitly the configuration by the user. Here it seems that the variety of possible scenarios is limited (e.g. iCAP). Other kinds of approaches can be adjusted to special scenarios in a perfect manner (e.g. CommonRules). But this causes complexity and often side-effects that complicates any reconfiguration or modifications.

6 Summary and Outlook

This paper describes an inference engine for the realisation of reasonable behaviour of smart environments, like smart living rooms or

[4] Here, not technologies for realising real artificial intelligence are in the main focus. Rather technologies that allow the configuration and implementation of reasonable conclusions have to be examined and developed.

smart meeting rooms, and its adaptation within attentive and proactive scenarios. The implementation of the inference engine is described in detail by means of the illustration of typical scenarios within a meeting room that built the basis for the requirements for the inference engine's realisation. In contrast to other well-known inference mechanisms the inference engine described here is able to deal with some kind of uncertainty as well as individual needs of different users. By introducing the concept of event sequences a rule is not only evaluated to true or false but also to a kind of interstage. With this approach the analysis of event sequences due to user interactions - regardless of their implicit or explicit characteristics - is possible; also in combination with the evaluation of "classical" environment variables like context data or user data. The context and user variables are not predetermined and can be chosen freely according to the given application set resp. their individual semantic meaning. Realizing this kind of semantical expressiveness we believe to have an instrument powerful enough to describe a variety of possible smart environment scenarios. The focus of the work that is described here is not in the implementation of "real" artificial intelligence but in the realisation of configurable reliable environment behaviour. Here we tried to achieve the most appropriate agreement between configurability (and also readability by a human developer) and mightiness. We believe that our system is able to realise scenarios of sufficient complexity without danger to get lost by ambiguous and confusing rules resp. rule sets.

In the future some important aspects have to be evaluated: We have to ascertain whether unexperienced users are able to define and configure reasonable rules after a short training period, because this aspect was one of the major motivations for an own implementation of an inference engine. The larger a rule set becomes the higher is the probability of having contradicting rules. Till now there are no mechanisms for conflict resolution (or event conflict detection) of contradicting rules implemented. Future evaluations will give hints what possible conflicts may happen and will indicate possible precautions. Additionally - this point is very much correlated with the previous aspect - we have to evaluate whether a common rule set and a common set of user variables is definable that is able to disburden the user's life in our smart environments at a larger scale. An earlier application [23] that recommends movies considering the mood of the user (here we used: merry, excited, sad, and amused) showed encouraging results. During evaluations we measured the acceptance of different adaptive movie suggestions. Some rules, that combined the mood of the user with her general preferences, reached an acceptance rate of 90%. Thus a validated list of appropriate user variables describing his internal emotional state and inferring reasonable room behaviour might be possible to elaborate. Also the integration of user feedback is important. Once a rule set is defined it is used without any modification at run time. Here, techniques that allow the direct involvement of the user(s) have to be examined and implemented.

ACKNOWLEDGEMENTS

The author wants to thank Prof. Thomas Kirste from the University of Rostock, Germany, for his valuable ideas and his infinite patience while explaining the secrets of Yacc. Additionally the author wants to thank the members of the consortium of the DynAMITE project [24]. In DynAMITE the software infrastructure was implemented that is used as the underlying communication technology for the applications that are described in this paper.

REFERENCES

[1] J. Krumm, S. Shafer and A. Wilson, *How a Smart Environment can use Perception*, published on Workshop on Sensing and Perception for Ubiquitous Computing, September, 2001

[2] C.D. Kidd, R. Orr, G.D. Abowd, C.G. Atkeson, I.A. Essa, B. MacIntyre, E. Mynatt, T.E. Starner and W. Newstetter, *The Aware Home: A Living Laboratory for Ubiquitous Computing Research*, Proc. of the 2nd Intern. Workshop on Cooperative Buildings, Integrating Information, Organization, and Architecture (CoBuild'99), pp. 191–198, Springer Verlag, London, UK, 1999

[3] R.A. Brooks, *The Intelligent Room Project*, Proc. of the 2nd Int. Conf. on Cognitive Technology (CT 97), pp. 271–278, Aizu, Japan, August, 1997

[4] M. Coen, *Design Principles for Intelligent Environments*, Intelligent Environments, Papers from the 1998 AAAI Spring Symposium, Technical Report SS-98-92, pp. 37–43, AAAI Press, 1998

[5] M. Coen, *Building Brains for Rooms: Designing Distributed Software Agents*, Proc. of 9th Conference on Innovative Applications of Artificial Intelligence, pp. 971– 977, Providence, Rhode Island, 1997

[6] G.D. Abowd, E.D. Mynatt and T. Rodden, *The Human Experience*, IEEE Pervasive Computing, 1(1), pp. 48–57, 2002

[7] A. Dey, *Understanding and Using Context*, Personal and Ubiquitous Computing Journal, Volume 5(1), pp. 4–7, 2001

[8] M. Hellenschmidt, T. Kirste and T. Rieger, *An Agent based Approach to distributed User Profile Management within a multi-model Environment*, Proc. of the IMC Workshop 2003 for Assistance, Mobility and Applications, pp.54–61, Stuttgart, Fraunhofer IRB Verlag, 2003

[9] S.C. Johnson, *Yacc: Yet Another Compiler Compiler*, Holt, Rinehart, Winston Eds): UNIX Programmer's Manual, 2, pp. 353–387, New York, USA, 1979

[10] BYACC/J, YACC compiler for Java, http://byaccj.sourceforge.net/, last known update: January 2nd, 2007

[11] M. Hellenschmidt and T. Kirste, *SodaPop: A Software infrastructure supporting self-organization in Intelligent Environments*, Proc. of the 2nd IEEE Conference on Industrial Informatics INDIN'04, Berlin, Germany, June, 2004

[12] M. Hellenschmidt and T. Kirste, *A generic topology for Ambient Intelligence*, Proc. of the Second European Symposium on Ambient Intelligence EUSAI 2004, Eindhoven, The Nederlands, 2004

[13] T. Zhang and S. Covaci, *Adaptive Behaviours of Intelligent Agents based on Neural Semantic Knowledge*, Proceedings of the 2002 Symposium on Applications and the Internet (SAINT'02), 2002

[14] Ch. Forgy, *OPS5 User's Manual*, Technical Report CMU-CS-81-135, Carnegie Mellon University, 1981

[15] J. Hoffmann and B. Nebel, *The FF Planning System: Fast Plan Generation Through Heuristic Search*, Journal of Artificial Intelligence Research, Volume 14, pp. 253–302, 2001

[16] A. Kulkarni, *Design Principles of a Reactive Behavioral System for the Intelligent Room*, Bitstream: The MIT Journal of EECS Student Research, April, 2002

[17] Ch. Forgy, *Rete: A Fast Algorithm for the many pattern/many object pattern match problem*, Artificial Intelligence, 19, pp. 17–37, 1982

[18] B. Grosof, *Courteous Logic Programs: Prioritized Conflict Handling for Rules*, IBM Research Report RC 20836, 1997

[19] A. Kakas, P. Mancarella, F. Sadri, K. Stathis and F. Toni, *The KGP model of agency*, Proc. of the General European Conf. on Artificial Intelligence, pp. 33–37, Valencia, Spanien, August 23–27, 2004

[20] T. Gu, H.K. Pung and Q. Zhang, *A Bayesian Approach for Dealing with uncertain Contexts*, Proceedings of the 2nd Intern. Conf. on Pervasive Computing (Pervasive 2004), Wien, Osterreich, 2004

[21] A.K. Dey, T. Sohn, S. Streng and J. Kodama, *iCAP: Interactive Prototyping of context-aware Applications*, K.P. Fishkin et al. (Eds.): PERVASIVE 2006, LNCS 3968, pp.254–271, Springer-Verlag Berlin Heidelberg, 2006

[22] T. Heider and T. Kirste, *Multimodal Appliance Cooperation based on Explicit Goals: Concepts & Potentials*, Proceedings of sOc-EUSAI 2005, Grenoble, France, October 12–14, 2005

[23] J. Nitschke and M. Hellenschmidt, *Design and Evaluation of Adaptive Assistance for the Selection of Movies*, Proc. of the IMC Workshop 2003 for Assistance, Mobility and Applications, pp.129–135, Stuttgart, Fraunhofer IRB Verlag, 2003

[24] DynAMITE, Dynamic adaptive multimodal IT-Ensembles, http://www.dynamite-project.org, 2006

Agent-Based Group Modelling for Ambient Intelligence

Judith Masthoff[1], Wamberto W. Vasconcelos[1], Chris Aitken[1] and Flávio S. Correa da Silva[2]

Abstract. Ambient intelligence allows physical environments to become sensitive and responsive to the presence of people and objects. An environment endowed with ambient intelligence is able to analyse its contexts, adapt itself to the presence of people and objects residing in it, learn from their behaviour and recognise and express emotion. Ambient intelligence is realised via devices which blend into the background, while supporting social interaction and improving people's experience within the physical space (e.g., by increasing safety or comfort). Often physical spaces must be shared by various people: adapting devices' responses and behaviour to simultaneously suit a group of people is an important and not much explored issue. To complicate matters further, group membership may change continuously. In this paper, we propose an approach based on group adaptation and software agents, to manage shared devices in ambient intelligence solutions. We present a proof-of-concept implementation embedding our approach, which allows engineers to design and experiment with distinct ways of managing shared devices – software agents are associated with devices and people, and interact with each other to agree on how shared devices should change their behaviours in view of the people in their radius of action.

1 INTRODUCTION

The notion of responsive environments is broad, encompassing essentially every space capable of sensing and responding accordingly to entities that inhabit them (these entities can be people, animals, or any sort of identifiable objects).

In this work, we focus on a narrower class of responsive environments, namely those provided with ambient intelligence. Ambient intelligence was characterised by Gaggioli [1] as referring to physical environments that are sensitive and responsive to the presence of people. Their key features are intelligence and embedding. "Intelligence" here refers to the fact that the digital environment is able to analyse the context, adapt itself to the people and objects that reside in it, learn from their behaviour, and eventually recognise as well as express emotion. "Embedding" means that devices with computing power will blend into the background of peoples' activities, and that social interaction and functionality will move to the foreground. In this paper, we are particularly concerned with ambient intelligence aimed at adapting to *groups* of users, with the group membership continually changing.

To illustrate the class of problems we aim at, let us consider the case of information delivery to groups of users. Many interesting applications can be envisaged that fit into this setting:

- Large displays can be installed in public spaces (airports, train stations, shopping malls, etc.) for the purposes of advertisement, entertainment and specific information delivery. The consumers of these services can form a very heterogeneous group of individuals. For example, in a train station we can find three individuals sharing the same physical space: the first one is a tourist with plenty of available time and interested in shopping local goods, the second one is a regular passenger who must wait every day for two hours at the train station to commute, and the third person is a hurried passenger looking for the right platform to get on a train that is about to depart. Ideally, a display that is visible to these three individuals should be sensitive to their interests and needs and adapt its displayed information accordingly.
- Digital display windows are becoming ubiquitous in all sorts of shops. Ideally, these displays should be sensitive to the customers who approach them, avoiding products that may offend or annoy customers and presenting content that is capable of raising the desire to consume, keeping customers in the shop for as long as possible and making the overall experience of visiting the shop as pleasant and entertaining as possible.

In this article, for the purposes of illustrating our approach, we employ the second application above: a bookstore where sensors detect the presence of customers identified by some portable device (e.g. a Bluetooth-enabled mobile phone, or a fidelity card equipped with an active RFID tag). In this scenario there are various sensors distributed among the shelves and sections of the bookstore which are able to detect the presence of individual customers (Figure 1 illustrates this scenario). The bookstore can associate the identification of customers with their profiling information, such as preferences, buying patterns and so on.

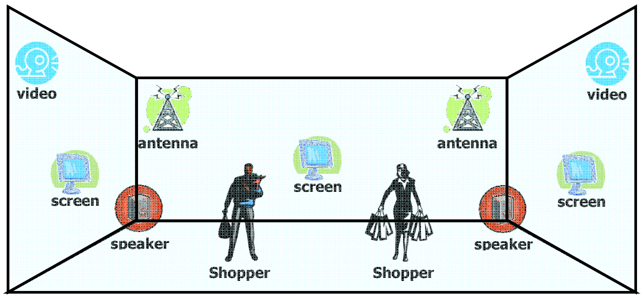


Figure 1. Example of Responsive Environment

[1] Dept. of Computing Science, Univ. of Aberdeen, AB24 3UE, UK. Email: {jmasthof, wvasconc}@csd.abdn.ac.uk; c.a.aitken@gmail.com

[2] Depto. de Ciência da Computação, IME, Univ. de São Paulo, Rua do Matão, 1010. Butantã, 05504090, SP, Brazil. Email: fcs@ime.usp.br.

With this infrastructure in place, the bookstore can provide customers with a responsive environment that would adapt to maximise their well-being with a view to increasing sales. For instance, the device playing the background music should take into account the preferences of the group of customers within hearing distance. Similarly, LCD displays scattered in the store show items based on the customers nearby, the lights on the shop's display window (showing new titles) can be rearranged to reflect the preferences and interests of the group of customers watching it, and so on.

The notion of a group in an ambient intelligence environment is slightly more complicated than e.g. a group of friends watching television together (an example used in our previous work in [4]). In a bookstore, a group can be a randomised collection of people who have never seen each other before, who happen to be together near a screen. However, families may also enter the store, so a group can also be more socially coherent or contain more coherent subgroups. In this paper, we will not take social coherence into account, but we will return to this issue in our conclusions.

This paper extends and combines our earlier work on managing responsive environments with software agents [2, 3] and group adaptation [4, 5]. We describe a scaleable and robust infra-structure implemented to support a team of software agents to manage devices shared by a number of people. Our approach naturally addresses a dynamic environment in which people and devices appear, disappear and move about in the physical space. Each component is associated with a software agent that represents the component's capabilities and needs. In order to manage the potentially conflicting interests of various agents sharing resources (e.g., a group of shoppers with different interests in front of a screen showing articles on offer) we have experimented with group adaptation techniques. With our infrastructure, different group adaptation techniques can be easily adapted and used to control shared devices.

The paper is organized as follows. First, we summarize our findings on group adaptation from [4, 5]. Next, we describe the agent architecture proposed in [2,5] for responsive environments. In Section 4, we integrate the work on group adaptation (described in Section 2) with the architectural work (described in Section 3). In Section 5, we discuss a proof-of-concept implementation. Section 6 contrasts this paper with related work. Section 7 presents our conclusions and provides directions for future work.

2 GROUP ADAPTATION

This section summarizes our findings on group adaptation from [4, 5] and relates them to ambient intelligence. Suppose the environment (e.g. a shop) contains three people, John, Adam and Mary. Suppose a device in this environment (e.g. a display) is aware that these three individuals are present and knows their interest in each of a set of items (e.g. music clips or advertisements). Table 1 gives example ratings on a scale of 1 (really hate) to 10 (really like). Which items should the display show, given time for four items?

Table 1. Example of Individual Ratings for Ten Items (A to J)

	A	B	C	D	E	F	G	H	I	J
John	10	4	3	6	10	9	6	8	10	8
Adam	1	9	8	9	7	9	6	9	3	8
Mary	10	5	2	7	9	8	5	6	7	6

Many different strategies exist for aggregating ratings of individuals into a rating of a group (e.g. used in elections, like when selecting the leader of a political party). Eleven of these (inspired by Social Choice Theory) are discussed in [4]. For instance, one could average the ratings of the individuals to obtain a group rating (making E and F the most preferred items by the group): the Average Strategy. One could take the minimum of the ratings, assuming that a group is as happy as its least happy member (giving a group rating of 1 for item A): the Least-Misery Strategy. We conducted a series of experiments to investigate which strategy is best (see [4] for details).

In experiment 1, we investigated how people would solve this problem, so given ratings for individuals (as in Table 1), which items they thought the group should watch, if there was time for say six items. We compared our subjects' decisions (and rationale) with those of the aggregation strategies. We found that humans care about fairness, and about preventing misery and starvation ("this one is for Mary, as she has had nothing she liked so far"). Subjects' behaviour reflected that of several of the strategies (e.g. Average and Least Misery were used), while other strategies were clearly not used. It should be noted that avoiding misery is not only desirable in socially cohesive groups (where members might want to avoid misery for each other). In a group of strangers, avoiding misery may well be even more important, as a stranger will have no reason to stay if they are miserable.

In experiment 2, we presented subjects with item sequences chosen by the aggregation strategies. Subjects rated how satisfied they thought the group members would be with those sequences, and explained their ratings. We found that the Multiplicative Strategy (which multiplies the individual ratings) performed best, in the sense that all subjects thought its sequence would keep all members of the group satisfied. Several strategies could be discarded as they clearly were judged to result in misery for group members. We also compared the subjects' judgements with predictions by simple satisfaction modelling functions. Amongst other, we found that more accurate predictions resulted from using quadratic ratings (which e.g. makes the difference between a rating of 9 and 10 bigger than that between a rating of 5 and 6) and from normalization.

In responsive environments, group membership changes continuously. Deciding on the next five items to show based on the current members does not seem to be a sensible strategy, as in the worse case, none of these members may be present anymore when the fifth item is shown. Additionally, overall satisfaction with a sequence may depend on the order of the items: for instance, it may be good for satisfaction to have mood consistency (not putting a depressing item in the middle of two happy ones), have a strong ending, and provide a good narrative flow. In experiment 3, we investigated how a previous item may influence the impact of the next item. Amongst others, we found that mood (resulting from the previous item) and topical relatedness can influence ratings for subsequent items. This means that in a responsive environment, aggregating individual profiles into a group profile should be done repeatedly, every time a decision needs to be made about the next item to display.

When adapting to a group of people, you cannot give everybody what they like all the time. However, you do not want anybody to get too dissatisfied. For instance, in a shop it would be bad if a customer were to leave and never come back, because they really cannot stand the background music. Many shops currently opt to play music that nobody really hates, but most people not love either. This may prevent losing customers, but would not result in increasing sales. An ideal shop would adapt the music to the customers in hearing range in such a way that they get songs they really like most of the time (increasing the likelihood of sales and returns to the shop). To achieve this, it is unavoidable that customers will occasionally get songs they hate, but this should happen at a moment when they can cope with it (e.g. when being in a good mood because they loved the previous songs). Therefore, it is important to monitor continuously how satisfied each group member is. Of course, it would put an unacceptable burden on the customers if they had to rate their satisfaction (on music, advertisements etc) all the time. Similarly, measuring this satisfaction via sensors (like heart rate monitors or facial expression recognizers) is not yet an option, as they tend to be too intrusive, inaccurate or expensive. So, we propose to model group members' satisfaction, predicting it based on what we know about their likes and dislikes.

In [5], we investigated four satisfaction functions to perform this modelling. We compared the predictions of these satisfaction functions with the predictions of real users. We also performed an experiment to compare the predictions with the real feelings of users[3]. The satisfaction function that performed best defines the satisfaction of a user with a new item *i* after having seen a sequence of items *items* as:

$$\text{Sat}(items + \langle i \rangle) = \frac{\delta \times \text{Sat}(items) + \text{Impact}(i, \delta \times \text{Sat}(items))}{1+\delta}$$

with the impact on satisfaction of new item *i* given existing satisfaction *s* defined as

$$\text{Impact}(i, s) = \text{Impact}(i) + (s - \text{Impact}(i)) \times \varepsilon$$

for $0 \leq \varepsilon \leq 1$ and $0 \leq \delta \leq 1$.

Parameter δ represents satisfaction decaying over time (with δ=0 past items have no influence, with δ=1 there is no decay).

Parameter ε represents the influence of the user's satisfaction based on previous items on the impact of a new item. This is based on the psychology and economics literature discussed in [5] which shows that mood impacts evaluative judgement. For instance, half the subjects answering a questionnaire about their TVs received a small present first to put them in a good mood. These subjects were found to have TVs that performed better.

Parameters δ and ε are user dependent (as confirmed in the experiment in [5]). We do not define Impact(*i*) in this paper, but refer readers to [5] for details: it involves quadratic ratings and normalization as found in the experiment discussed above.

The satisfaction function given does not take the satisfaction of other users into account, which may well influence a user's satisfaction. As argued in [5] (based on social psychology), two main processes can take place. Firstly, the satisfaction of other users nearby can lead to so-called emotional contagion: other users being satisfied may increase a user's satisfaction (e.g. if somebody smiles at you, you may automatically smile back and feel better as a result). An experiment in [5] shows that this emotional contagion depends on the relationship you have: you are more likely to be contaged by somebody you love or respect (like your child or boss) then by somebody you do not know.

Secondly, the opinion of other users nearby may influence your own expressed opinion, based on the so-called process of conformity. Two types of conformity exist: (1) normative influence, in which you want to be part of the group and express an opinion like the rest of the group even though you still believe differently, and (2) informational influence, in which your own opinion changes because you believe the group must be right.

More complicated satisfaction functions are presented in [5] to model emotional contagion and both types of conformity. However, the work presented in this paper uses the function given above (and its variants), postponing the incorporation of group influence to future work.

3 AGENT-BASED AMBIENT INTELLIGENCE

Software agents [6] have been used in responsive environments solutions (e.g., [2, 3, 7, 8, 9, 10]). The association of distributed threads of execution with physical components allows for arbitrary functionalities to be used in the management of resources and coordination of activities. These functionalities are combined with the desirable features of software agents such as proactiveness and social abilities (communication) [6]. For instance, a digital camera able to take pictures can be associated with a software agent that will manage any requests from other components for pictures, but the agent will also store the last n pictures taken. Even though the camera itself may not have provisions for storing more than one picture, by associating an independent thread of execution with it, we are able to extend its functionalities.

The same physical components can be associated with different software agents at different times, thus allowing for hassle-free versioning. In such case, engineers and programmers devise new versions of software agents to replace previous ones, fixing any bugs, improving on existing features or adding new functionalities to take advantage of new components. The new software agents can take over from their previous counterparts without the need to redesign the whole solution from scratch

We propose to assign a software agent to every device and person in the environment, following [3], to endow it with ambient intelligence. We illustrate this approach through figure 2: the rectangular box represents the physical environment and the "cloud" above it stands for the digital (logical) environment. For instance, customer 1 is represented by user agent c_1 and device d_1 by agent d_1. Each device has an action radius which may be determined, *e.g.*, by the range of its sensors or the visibility of its display. For instance, device d_2 is only reacting to two people in its action range (customers 4 and 5). Hence, c_4 and c_5 are the only user agents currently communicating with d_2.

The environment is dynamic: both devices and humans may enter, move around, and leave at any moment. So, their corresponding agents need to be created, updated and terminated automatically. Additionally, the connections between the agents cannot be static, since with whom the agents need to communicate can change continuously.

[3] This was done in another (educational) domain. See [5] for a discussion on why this was necessary.

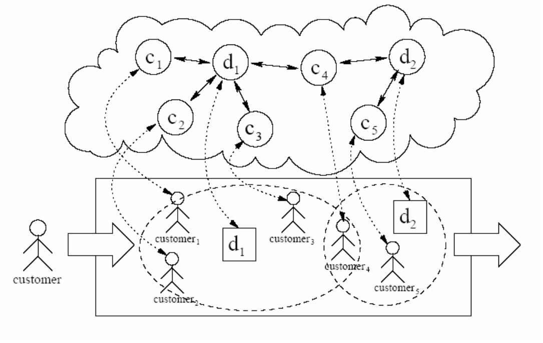


Figure 2. Agent-Based Responsive Environments

To allow for ad-hoc communication among various parties, we follow [2, 3] and use a blackboard architecture, implemented using JavaSpaces[4] [11]. So, agents do not communicate directly, but constantly monitor and post messages on the tuple space. Similarly, physical entities communicate with the tuple space rather than directly with their corresponding agents. Administrator agents manage the 'digital cloud', continuously monitoring the tuple space and creating, updating and terminating agents to reflect what happens in the physical world.

4 INTEGRATING GROUP ADAPTATION

To integrate group adaptation into the architecture, we made the device agents into aggregator agents: they decide what the device should do (e.g. which music to play) depending on the opinions of the user agents within their action radius.

The goal of a user agent (c_1 etc in Figure 2) is to increase the satisfaction of its physical counterpart by influencing the items that are played on the shared device[5]. It does so by viewing what is being displayed at a given time, updating its satisfaction and notifying the display's aggregator agent. To update satisfaction, the formula described in Section 2 is used (and its variants described in [5]), with predetermined values for its parameters[6].

The goal of an aggregator agent (d_1 etc in Figure 2) is to control the shared device and keep the users within its action radius as satisfied as possible. It does so by continually asking each user agent within its action radius for its satisfaction in relation to the items displayed so far, and its profile (which provides ratings for the possible items to display next).

The strategies described in [4] (like the Average Strategy and the Least Misery strategy) did not use information about the group members' satisfaction so far. Receiving this information from the user agents allows our aggregator agents to use more sophisticated strategies. It seems sensible for the aggregator agents to try to increase the satisfaction of the least satisfied user within their action radius, whilst trying still to take into account the opinions of all other group members.

Aitken [8] proposed that the aggregator agents determine (1) an aggregated profile (for the possible items to display next), using one of the standard algorithms described in [5], and (2) the least satisfied member so far. The aggregator agents then select the item with the highest (individual) rating for the least satisfied member. If multiple items with such highest rating exist -let us call these candidates- the aggregated profile is used, selecting the candidate that has the highest rating in the aggregated profile. For instance, in the example in Table 1, suppose John is the least satisfied member so far. Based on this, items A, E, and I are candidates to display next, as they have the highest rating for John. If the Average Strategy were used to determine the aggregated profile, item E would be displayed, as E has the highest average rating of the candidates for the group as a whole.

We will call this aggregation strategy the Strongly Support Grumpiest Strategy. Table 2 shows the items selected, and the resulting moods, when applying this strategy to the example data of Table 1. To calculate the moods, we have used δ=0.8 (a high value as this gave the best results in [5]) and ε=0.2 (a low value as this seemed best in [5]). We used the Average Strategy for determining the aggregate of the group. Table 2 shades the most miserable member at each moment in time, which was used by the strategy to select the next item.

Table 2. Strongly Support Grumpiest's decisions and mood

Step	Item selected	Mood		
		John	Adam	Mary
1	E	6.6	1	7.5
2	F	7.4	6.1	8.1
3	H	5.7	8.8	4.7
4	A	9.7	-4.3	14.3
5	D	5.2	3.3	9.4
6	B	1.4	7.3	5
7	I	7.4	1.2	4.4

The Strongly Support Grumpiest Strategy has some limitations. For instance, in the example of Table 1, suppose that Mary was the least satisfied member so far (based on items displayed not shown in Table 1). Strongly Support Grumpiest would result in item A being displayed next. Whilst this clearly would make Mary more satisfied, it has a bad effect on Adam's satisfaction. Displaying item E may well be a better option, as it has almost the same rating for Mary whilst being significantly better for Adam. This could be incorporated into the strategy by making higher-level groupings of ratings, e.g. treating 9 and 10 as "highly satisfying".

However, it is not clear how far one should go with this. In the example of Table 2, the choice to show A in step 4 makes Adam miserable, but, as item E has already been shown at this point, the next highest rated item for Mary is item D, with a rating of 7. This leads to the question whether we ought to consider all items with a "quite satisfactory" rating (7 or above) for the most miserable member instead of just the items with the highest ratings. We will call this alternative strategy the Weakly Support Grumpiest Strategy. Table 3 shows the items selected, and the resulting moods, when applying this strategy to the example data of Table 1. As before, we have used δ=0.8, ε=0.2, and the Average Strategy. Note, that misery still occurs (Adam is unhappy after step 6) though it happens later. Instead of using the Average Strategy, it may well be better to use on of the

[4] JavaSpaces is part of Sun's JINI Network Technology, available at `http://www.sun.com/software/jini/`

[5] In our case, no cheating takes place. User agents honestly report their mood, rather than e.g. always claiming to be very miserable in order to get a better next item.

[6] Both the variant to use and the parameter values are specified in configuration files.

strategies known to avoid misery (Least Misery, Average Without Misery and Multiplicative Strategies discussed in [4]).

Table 3. Weakly Support Grumpiest's decisions and mood

Step	Item selected	Mood		
		John	Adam	Mary
1	E	6.6	1	7.5
2	F	7.4	6.1	8.1
3	H	5.7	8.8	4.7
4	A	3.0	10.3	4.2
5	D	3.4	8.3	2.6
6	B	8.4	-4.5	13.2
7	I	3.1	3.1	7.0

An alternative to these two Support Grumpiest strategies would be to attach weights to users, in a manner similar to that described in [12]. Weights would depend on the user's satisfaction, with satisfied users having a lower weight than dissatisfied ones. Using weights works well with some aggregation strategies (like the Average Strategy and Multiplicative Strategy), but is impossible to do with others (like the Least Misery Strategy). The advantage of the two Support Grumpiest strategies is that they work for all aggregation strategies, allowing the designer of a responsive environment to experiment with different options. Such experimentation is important, as whilst results from [4, 5] show an advantage of using the Multiplicative Strategy, the best strategy to use may well be domain dependent.

Another alternative would be to use the aggregation strategies discussed in [4], but apply them only to all members of the group that are currently unsatisfied. More research will be needed to decide on the best way to use the users' satisfaction so-far, and to compare it with the use of the strategies discussed in [4].

5 IMPLEMENTATION

The ideas presented in this paper have been successfully implemented as a proof-of-concept prototype: a PC with an of-the-shelf Bluetooth™ USB adaptor (our sensor) detected Bluetooth™-enabled mobile phones within its range and delivered music and/or video clips via the PC. The owners of mobile phones had to previously register their profile with preferred genres and artists/groups, and any dislikes, this information being stored in a database to which software agents representing the humans had access. This implementation has been reported in [8]: we used JADE[7] to start up and manage our agents as lightweight threads, communicating via JavaSpaces [11], defining a computational environment [13] using freely available technologies.

We chose to use Bluetooth™ to detect users entering or leaving the environment, as it is a widely accepted open standard that is already integrated into many devices (like mobile phones). Infrared is only useful in line-of-sight, so was judged impractical for our purposes, as users would have to scan when entering or leaving the environment. We did not use RFID tags as these were not readily available, but we do not anticipate any significant problems if we were to use them instead of Bluetooth-enabled devices.

For evaluation purposes, we wanted to see graphs of the users' predicted moods. We implemented a graph writing component using Java2D[8]. Figure 3 shows a screenshot with a mood graph for two users in the environment. Hovering over the mood graph shows the details of the item being played at that time (the screen shot shows that a Bach MP3 file has been played as the third item). In this example, the Strongly Support Grumpiest Strategy has been used with the Average Strategy as its sub strategy, and satisfaction has been modelled with a delta value of 0.8 and an epsilon of 0. A history of graphs is kept, allowing the designer to compare algorithms easily.

In the example shown in Figure 3, Adam entered the environment first. Using user ratings as in Table 1, the aggregator agent decided to play item B (one of the items with highest rating for Adam). Whilst item B was playing, John entered the environment. John did not like item B, so his mood became negative, and the aggregator agent decided to play item E next. This resulted in Adam's mood becoming the lower of the two, and item F being played next.

For the shared device, we implemented a media player using the Java Media Framework[9]. This allowed us to control the device directly through the JavaSpaces (which would be impossible if using an existing media player such as Windows Media Player[10]).

To ensure a degree of robustness, each agent has a backup. The main agents keep a dialogue with their backups, notifying them that they are still alive. If the main agent does not respond, then the backup spawns a new main agent, thus allowing the application to stay active.

The software has been extensively tested to ensure it can deal with a reasonable number of users (10-50 seems appropriate for the kind of scenarios we are interested in) and keeps working when individual processes fail (so, the backup system works). While the implementation provides a proof-of-concept, its goal is, in fact, far more ambitious. It provides an important tool for further research into this area. The software has been kept as generic as possible and facilities have been provided to tailor it: it is easy to modify the group modelling (e.g. add other aggregation algorithms, modify the satisfaction function, change parameters), and model other responsive environments.

6 RELATED WORK

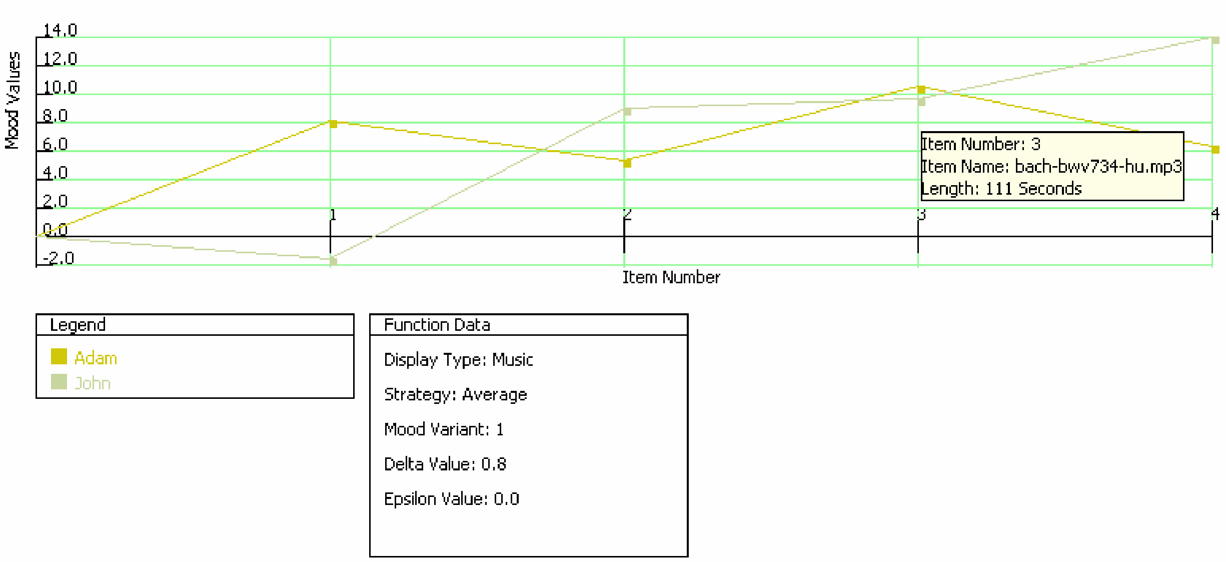
The Intelligent Inhabited Environments research group [10] at the University of Essex explicitly proposes, as we do, the construction of intelligent responsive environments through the coupling of the physical world and virtual worlds inhabited by software agents. However, their test bed, the iDorm experiment,

[7] Java Agent DEvelopment Framework, available at `http://jade.tilab.com`. Although JADE has its own communication facilities, we did not make use of them. Instead, we used JADE to facilitate the management and debugging of our agents. By using JavaSpaces, we confer openness on our solution, as any Java-enabled device can communicate with other components by posting and retrieving entries from the space.

[8] `http://java.sun.com/products/java-media/2D`

[9] `http://java.sun.com/products/java-media/jmf`

[10] `http://www.microsoft.com/windows/windowsmedia`



Item Number: 3
Item Name: bach-bwv734-hu.mp3
Length: 111 Seconds

Function Data

Display Type: Music

Strategy: Average

Mood Variant: 1

Delta Value: 0.8

Epsilon Value: 0.0

Figure 3. Screenshot of Mood Graph

which is a student dormitory facility to serve a single student, equipped with a host of sensors and effectors that can monitor the activities in it and respond accordingly, only allows for single-occupant scenarios.

In [2, 3] a negotiation protocol is proposed to allow user agents suggest settings for a shared device. The protocol is one-off in that it does not keep track of previous results. User agents communicate their best choices to the agent managing the shared device; these preferences define a space of possible configurations from which one final configuration must be drawn. Each of the preferences is also associated with the "power" of influence of that user agent: depending on how high this power is, the final configuration will be closer to that agent's choice. This is a primitive kind of group adaptation, in which the power of influence remains static.

Group rating naturally connects with the area of Collaborative Filtering [14], in which systems are built to predict a person's affinity for items or information by connecting that person's recorded profile with the profiles of a group of people and sharing ratings between similar persons. Location aware collaborative filtering and the filtering of interests of a group instead of a single person – which are specific proposals and contributions of our work – are two recent research topics not yet explored.

In [15], a system called MusicFX is described which is used in a company's fitness centre to select background music to suit a group of people working out at any given time. MusicFX selects radio stations, rather than individual songs (so, is less concerned with a sequence of items). It uses a version of the Average without Misery strategy (see [3]). It does not try to model user satisfaction.

A relevant project, with similar goals and methodology to ours, is found in [16]. In this work, the authors propose the application of aggregation strategies based on our overview in [4] to determine the contents of public displays. It is even argued, in agreement with our approach, that a distributed architecture is appropriate to leverage system performance. Our contribution with respect to that work lies in the explicit use of multi-agent technologies, which enables a more descriptive and yet concise presentation of our architecture. In addition, we describe how the modelling of user mood can be used as part of the aggregation process.

7 CONCLUSIONS & FUTURE WORK

This paper shows how existing work on agent architectures for ambient intelligence can be combined with work on group adaptation to obtain responsive environments that take the affective state (satisfaction) of their users into account. A proof-of-concept implementation was presented, which has been functionally tested and which will provide a test bed for further research into this area. We intend to perform a range of experiments to see how well the aggregator agents function and to explore the advantages and disadvantages of several approaches for incorporating user satisfaction into the decision making.

In the architecture and implementation presented so far, the satisfaction of a user only depends on the items displayed, not directly on the other users in the environment. So, it does not yet allow for contagion and conformity. We would like to extend our work to incorporate this. This would mean that user agents should be communicating with the agents of users nearby to express their satisfaction. The importance of modelling contagion and conformity will depend on the application domain. For example, when adapting music, contagion and conformity are likely to be higher in certain environments (like a pub) than in other environments (like a bookshop), as users are more aware of each other (looking at each other rather than at the books) and are more likely to know each other (as mentioned the relationship type influences contagion). To reduce communication, it may well be sufficient for agents to communicate their satisfaction only to agents representing users that their users have a good relationship with. So, for instance, the agents representing a mother and her child would exchange information, but the agents of two strangers not. This would be one way of dealing with socially coherent subgroups.

We have considered in our studies and experiments rather sophisticated rating strategies, derived from Social Choice Theory. The items considered to be rated, however, have been a little simplistic in our experiments so far. Considering for example the train station scenario devised in the beginning of this article, we can have a group of heterogeneous agents with diverse interests competing for the display. These different interests may not be comparable (e.g. the interest in learning about available products in nearby shops and the need to obtain information about train departures), and in this case more sophisticated decision procedures must be implemented in the aggregator agent, probably resorting to multi-attribute decision procedures. We have thus far also employed simplifying assumptions about the behaviour of the users, as well as of our designed aggregating agents. One of our major simplifying assumptions is that the goals of the users are unique and stable, and we extend this assumption to the aggregating agent. More refined implementations shall be considered in the future, in which we take into account that users can change their minds and interests dynamically and yet predictably, and in which we refine the behaviour of the aggregating agents so that they can change their goals and strategies depending on the group of agents that is sensed to be in the vicinity of a display.

In this paper, we have not discussed the user experience in detail. We have talked about the display contents and background music etc changing automatically, but not about how much information users will be given about *why* things are changing, and *how* the decisions are made. We expect these user interface issues to be domain dependent. For some application domains, like background music, it may be sufficient for users entering the store to be aware that the music will start adapting to them after the current song has finished. In such a scenario, users are unlikely to want detailed information displayed about the decision making process. The system we implemented can give very detailed information about why decisions were made, but this detail is currently intended for researchers only. It should also be noted that giving insight into the decision making may result in privacy issues (as discussed in [5]). Users may not always want their private tastes disclosed to those who happen to be standing near them. Even without giving explicit explanations for decisions made, it may be possible to deduce some information about individual profiles by observing the decisions the system makes. This is an issue that warrants more research. We made a starting point on this in [5], where we investigated which aggregation strategies were best for maintaining privacy.

REFERENCES

[1] A. Gaggioli. Optimal Experience in Ambient Intelligence. In: *Ambient Intelligence*. G. Riva, F. Vatalaro, F. Davide, M. Alcañiz (Eds.). IOS Press (2005).

[2] F. S. Correa da Silva and W. W. Vasconcelos. Managing Responsive Environments with Software Agents. *Journal of Appl. AI* (in press).

[3] F. S. Correa da Silva and W. W. Vasconcelos. Agent-Based Management of Responsive Environments. Procs. 9th AI*IA, Milan, Italy. LNAI Vol. 3673, Springer-Verlag, Berlin, Germany (2005).

[4] J. Masthoff. Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers. *UMUAI,* 14:37-85 (2004).

[5] J. Masthoff and A. Gatt. In Pursuit of Satisfaction and the Prevention of Embarrassment: Affective State in Group Recommender Systems. *UMUAI,* 16:281-319 (2006).

[6] M. Wooldridge, *An Introduction to Multi-Agent Systems*, John Wiley & Sons Ltd., England, U.K., (2002).

[7] J. M.V. Misker, C. J. Veenman, and L.J.M. Rothkrantz, Groups of Collaborating Users and Agents in Ambient Intelligent Environments. In: *Procs. 3rd Int'l Joint Conf. on Autonomous Agents & Multi-Agent Systems* (AAMAS'04), NY, USA. ACM Press. (2004)

[8] C. Aitken. *Designing Software Agents to Manage Responsive Environments*. B.Sc. Hons. Report. Dept. of Computing Science, University of Aberdeen, UK, `http://www.csd.abdn.ac.uk/~wvasconc/aitken_chris.pdf`. (2006).

[9] G. Vizzari, *Dynamic Interaction Spaces and Situated Multi-Agent Systems: from a Multi-Layered Model to a Distributed Architecture*, Ph.D. Dissertation, Universita degli Studi di Milano-Bicocca, Milan, Italy. (2004)

[10] H. Hagras, V. Callaghan, M. Colley, G. Clarke, A. Pounds-Cornish, and H. Duman. Creating an Ambient Intelligence Environment Using Embedded Agents. *IEEE Intelligent Systems*, 19:12-20, (2004).

[11] E. Freeman, S. Hupfer, and K. Arnold. *JavaSpaces: Principles, Patterns and Practice*. Addison-Wesley, USA. (1999)

[12] J. Masthoff. Selecting News to Suit a Group of Criteria: An Exploration. In: *Procs. 4th Workshop on Personalization in Future TV - Methods, Technologies, Applications for Personalized TV,* Eindhoven, Netherlands. Springer (2005).

[13] D. Weyns, M. Schumacher, A. Ricci, M. Viroli, and T. Holvoet. Environments in Multi-Agent Systems. *The Knowledge Engineering Review,* 20:127-141 (2005).

[14] J. L. Herlocker, J. A. Konstan and J.Riedl, Explaining Collaborative Filtering Recommendations. *Procs. ACM Conf. on CSCW*. Pennsylvania, USA (2000).

[15] J. McCarthy and T. Anagnost. MusicFX: An Arbiter of Group Preferences for Computer Supported Collaborative Workouts. In: *Procs. ACM Conf. on CSCW*, Seattle, USA, pp. 363-372 (1998).

[16] S. Pizzutilo, B. De Carolis, G. Cozzolongo, F. Ambruoso. Group Modeling in a Public Space: Methods, Techniques and Experiences. In: *Procs. of WSEAS AIC 05*, Malta, Sept. 2005. pp. 175-180 (2005).

Emotional Design for Public Displays

Daniel Schulz [1] and **Hans Jörg Müller** and **Antonio Krüger** [2]

Abstract. This paper proposes an emotional design process for Public Displays. Public Displays are becoming ubiquitous in public spaces, and they are used for advertisements as well as general information. As Public Displays are always owned by a specific organisation, the emotions evoked by the displays are projected onto their owner. Thus, the role that emotions play with Public Displays is not to be underestimated. We shed light on this role by analysing research from the emotional design of websites and develop a design process to design Public Displays to evoke specific emotions.

1 INTRODUCTION

Imagine walking through a city of the future to your workplace. Nearly all public space is covered by electronic displays, showing advertisements as well as general information. Upon leaving the subway, you pass a store with displays crying for your attention. The images blink and move, changing colors all the time. You are not interested in the shop, and the advertisements make you angry and agressive, you even start to hate the shop itself. As you head on, you pass a second shop where the displays present a beach atmosphere, with mild waves, palm trees moving softly in the wind. You imagine to feel a soft breeze on your skin, and see a virtual fire cracking in the fireplace inside the store. As you are lured in, you spend a few minutes in the shop, just to relax and feel well. As you afterwards enter your companys building, the crisp and cristal clear displays present you the newest company information. You see companies stock going up, and a light and open atmosphere fills the lobby. In the right mood to kick off some new projects, you start working. Obviously, the emotions evoked by Public Displays and therefore associated with the displays owners are increasingly important with electronic displays filling the public space. The goal of this paper is to evaluate design factors which influence the users' emotions evoked by Public Displays. The emotional impact is one of the aspects determining the users acceptance of a system. This paper is structured as follows: In the next section the role that emotions play in the use of this system is discussed in the contexts of HCI and organizational behaviour. Subsequently, an overview of empirical studies on the emotional impact of websites is given. As this domain is closely related, the results of these studies concerning design factors and emotional outcomes can be adapted. The last section proposes an emotional design process for Public Displays.

2 EMOTIONS AND THEIR ROLE FOR PUBLIC DISPLAYS

2.1 Definition and Models of Affect and Emotion

There is a variety of definitions and models for emotions and related concepts [27]. For the course of this paper, the following terminology and definitions are used:
Affect is a term encompassing all sentiments, emotions, feelings, and moods [20, 27, 28]. It is a two-dimensional concept, one dimension being the valence value (positive/negative), the other being the arousal value (degree of activation) [27]. Positive and negative affect may coexist [20]. Emotions distinguish themselves from other forms of affect in the way that they are related to a cause. Emotions are rather temporary compared to other affective states [2]. They are multifaceted phenomena, being expressed through behavioural expressions, expressive reactions, physiological reactions, and subjective feelings [3]. The emotional system is intertwined with the cognitive system [8]. Thus, emotions influence cognition, behaviour, decision processes, creativity, curiosity, and learning. Positive emotions are associated with a broader, more creative way of thinking, whereas negative emotions lead to a more focused way of processing with more emphasis on cognition [18]. Norman presents an integrated model of the three different levels of processing involved with the emotional and the cognitive system [8, 11, 18]. Every stimulus is processed by all levels. The visceral level is the lowest and fastest level of processing. It incorporates subconscious "hard-coded" reactions to stimuli. The evaluation on this level is simple in comparison to the higher levels. The visceral processing has direct connection to the motor and the sensory system, enabling quick reactions. It is the beginning of the affective processing; the results are passed on to the higher levels. No learning occurs and no cultural differences are to be found. The behavioural level incorporates the learned reactions to everyday behaviour. Processing on this level is still subconscious. The behavioural level is also connected to the sensory and the motor system. As behaviour is learned, the processing on this level is dependant on experience and culture. The behavioural level can control the visceral level and is itself under the control of the reflective level. The highest form of processing takes place on the reflective level. Input is received from the other levels and reflected upon consciously. Reasoning, cognition, and interpretation take place here. These processes can influence the lower levels, but no direct connection to the motor system exists.

2.2 Role of Emotions for Public Displays

Public Displays are electronic displays that are installed in public spaces, some of which can adapt to their environment [16]. HCI has traditionally concerned itself with instrumental aspects of interactive systems like usability [8]. Recently, non-instrumental aspects gained

[1] Institute for Information Systems, University of Münster, Germany
[2] Institute for Geoinformatics, University of Münster, Germany, email: joerg.mueller@uni-muenster.de

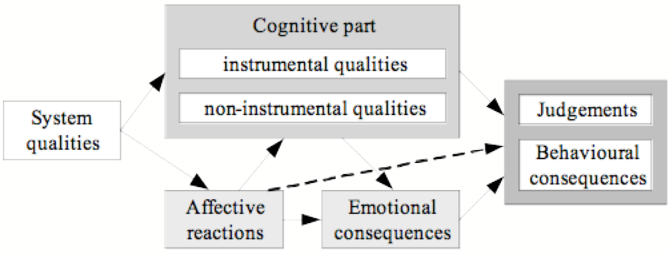


Figure 1. User experience process.

attention, like emotions e.g. [18], aesthetics e.g. [25], fun e.g. [2], or affect e.g. [11].

Mahlke presents a framework (see Figure 1) for the user experience of interaction [14]. It helps to interpret the way a system's qualities influence the user's behaviour. Mahlke differentiates between the cognitive and the affective processes but also shows their interconnections. The cognition of instrumental, task-related qualities like usability and non-instrumental qualities like aesthetics are integrated. Concerning affect, Mahlke distinguishes immediate affective reactions, in reference to Normans visceral level, and long-term emotional outcomes during and after the interaction. The immediate affective reactions influence the cognitive processes as well as the emotional consequences and can have an impact on the overall user impression. The emotional consequences are shaped by the result of the cognitive processing and both influence the final judgements and the user's behaviour. The impact of affect on the cognitive evaluation of a system depicted in this framework is supported by empirical studies for general user interfaces [24] and for websites in particular [27, 28].

Concerning the application of this framework to Public Displays, the ratio of instrumental to noninstrumental aspects evaluated by the user will be different than with information systems accessed via a desktop computer. Typical usage of Public Displays is passing-by displays and catching a glimpse of the information provided, sometimes combined with a longer reading time or interaction. Hence, the aesthetics of the information presentation are more important for the success of this system. A design with emotions in mind should target increased attention and better information reception, thus furthering the information exchange achieved with Public Displays.

In organizational behaviour the role of emotions has for a long time been neglected as the rationality of humans was the predominant assumption. Recently, emotions have found more attention in this discipline [6]. Emotions have an organizational impact on e.g. learning processes, decision making, job satisfaction, collaboration, change processes, motivation, and stress-coping [6]. As Rafaeli and Vilnai-Yavetz have shown, the emotions that result from the sense-making of an organizational artefact are projected onto the organization itself [21]. Thus, the emotions evoked by the Public Displays owned by organisations may on the one hand support information flow, learning processes and decision making. On the other hand, the use of such displays in organizational buildings may improve the corporate identity. This is especially important for displays installed in public places with customers passing by.

3 EMPIRICAL STUDIES ON WEBSITE EMOTIONS

A lot of design aspects of Public Displays concerning the information presentation are similar to the design of websites. The major difference is that while users actively retrieve websites on a desktop computer, the Public Displays are part of the user's environment and are observed while passing by. Nevertheless, a review of the literature on the emotional design of websites is reasonable in order to derive guidelines for Public Display design. Recently, there have been some studies in this context, investigating constructs like aesthetics, emotions or hedonic aspects of websites [9, 23].

As with Public Displays first impressions are very important, literature on the first impression of websites is of particular interest here. Tractinsky et al. conducted a study where fifty websites were presented to the participants [26]. In the first phase of the study the presentation time was 500 ms, in the second phase it was extended to 10 seconds. Every website was rated by the participants on a ten point scale for its visual attractiveness. Additionally, the response latency for this rating was used as an objective measurement. The authors' results show that the measurements for the short and long viewing time are highly correlated. Thus, the authors conclude, aesthetic impressions of websites are formed quickly. This result makes it more likely that guidelines for the emotional design of websites are also applicable for the design of Public Displays.

Lavie and Tractinsky address the measurement of perceived aesthetics of websites [12]. They found perceived aesthetics to be a two-dimensional construct. The first dimension, named "classical aesthetics", refers to traditional aesthetical concepts measured with items like "clean" or "aesthetic". The second dimension they found was named "expressive aesthetics". It refers to the designers' creativity, originality and the ability to break design conventions. To measure this dimension, items like "original" and "uses special effects" are used.

Ngo et al. developed mathematical measures that evaluate the aesthetics of graphical user interfaces concerning their structure [17]. These measures span factors like balance or complexity. The authors have shown empirically that these measures predict the rating of beauty for a given interface. Still missing is empirical data concerning the interplay of these factors and the relative impact of each factor. The influence on emotions is also yet to be analysed.

Schenkmann and Jönsson conducted a study with 13 websites to find out which subjective factors influence the viewers first impression [22]. Participants were not allowed to interact with the sample sites. The three factors having the greatest impact on the preference of websites were found to be beauty, mostly illustrations vs. mostly text and overview. Thus the authors advise to have more illustrations than text and make a page give a good overview.

The question of which aspects influence the perceived aesthetics of a company website is addressed by Thielsch et al. [23]. They conducted an online survey regarding heuristics for aesthetical website design and regarding the importance of website aspects for perceived aesthetics. The sample sites were clustered in three groups, "automobile industry", "financial services", and "other companies". Each participant was automatically assigned to one of these groups. The results show that there is nearly no difference in the aesthetical perception of company websites for different sectors. First impression, navigation, composition, and colour were found out to be the most important aesthetical aspects of a website. Significant acceptance was found for heuristics regarding simplicity, structure, and straightforwardness. Attention and memory as responses to e-commerce websites are both impacting buying behaviour. As they are also affected by a subject's emotions, they can be used as an indirect measurement of emotional states.

Lee and Benbasat manipulated three aspects of an e-commerce website (image size, fidelity, and motion) and measured attention and product recall in a laboratory experiment [13]. Memory was found to be influenced only by image size; larger images enhanced users' image recall performance. Attention was affected by visual fidelity and motion. The use of motion and high visual fidelity images on a website improved attention.

Kim et al. conducted several studies to identify the relationship between key design factors and dimensions of secondary emotions evoked by websites [10]. They used a multidimensional perspective on emotions, differentiating between primary and secondary emotions. They define primary emotions to be basic, generic emotions. Secondary emotions are understood as individual-dependent and domain-specific and are derived from primary emotions. As their study is restricted to websites, they chose secondary emotions because of the domain-specificity. They carried out three studies building on top of each other. The first was aimed at identifying the emotional dimensions of peoples' feelings when viewing websites. The authors identified 278 terms through a literature review. Homepage designers provided 48 sample pages which were chosen to be distinctively different from each other. These samples were used in a survey with 418 participants who were asked to rate a subset of the sites in a questionnaire containing a randomized selection of the original 278 dimensions. No interaction with the websites was performed. The results of this survey formed the basis a for cluster analysis. The results thereof are 13 secondary emotional dimensions evoked by websites like bright, tense, or adorable. The second study was performed to develop sample pages for each dimension and to identify the design factors contributing to the evocation of these emotions. Professional homepage designers provided four sample sites for each emotional dimension. Their starting point was a descriptive text which was the same for all sites except those parts relevant for the description of target emotions. The design factors applied in the samples were analysed by regarding the final results and by evaluating protocols and videotapes of the design sessions. Design factors were identified in a three fold structure, either targeting the foreground objects, the background, or the relationship of both. In the first two categories, shape, texture, and colour were the main design factors. The relationship category addressed the matching of the colours of the title, the menu, and the main images. The third study was conducted to identify the quantitative relationship of the prior results. 515 participants were asked to rate the sample pages of the second study based on 30 adjectives describing the 13 emotional dimensions. Again, no interaction with the sites was performed. The results show the validity of all dimensions and of the adjectives chosen to describe them. The quantitative relationship between design factors and resulting emotions is described in the form of regression equations, each showing a good statistical fit.

The results of this study allow the design of websites with desired emotional reactions in mind. As the authors point out, the results are specific for the Korean culture. Furthermore, the results are only valid for the website domain. An adaptation of the research method is an option to derive design rules for other domains.

Papachristos et al. [19] have build upon the results of Kim et al. and investigated the relationship of colour and emotional dimensions of a website. In their study the 46 participants had to rate one website layout presented in eight different colour schemes. Twelve emotional dimensions like pleasant, aggressive, or reliable, were chosen in a prestudy survey . The rating results together with the attributes of the different colour schemes were used as training data for a Bayesian Belief Network. The trained network allows designers to choose which emotional dimensions should be present with his website and to what extent. Additionally, the structure of the resulting network shows the order of colour attributes regarding their impact on emotions. The authors note that the resulting network is not applicable in general because of the small number of participants in the experiment. The confinement on one website layout might also limit the generalisability.

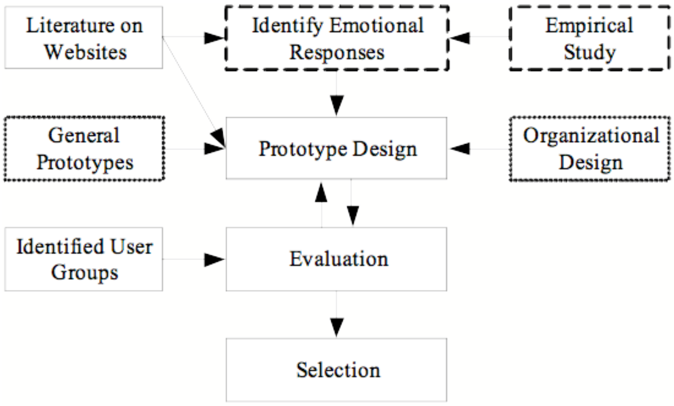


Figure 2. The Emotional Design Process.

4 AN EMOTIONAL DESIGN PROCESS FOR PUBLIC DISPLAYS

As Public Displays are installed as visible artefacts of organisations, they must be tailored to fit the organization's public image, among other aspects regarding the emotional design. The most important influences in this context are the corporate identity and design, the groups of people coming into touch with the system, and the information categories which are chosen to be communicated. In general, there are two approaches to designing Public Displays with emotions in mind. The first is to take emotions into account only when designing the prototype that is later to be customized. The other approach is to reconsider emotions when implementing the system in organizations. The latter is discussed in the following, as the first is just a simplification.

The emotional design of Public Displays is mainly focused on the aesthetics of the information presentation. Desmet et al. presented a methodology for designing consumer products with an optimal emotional experience [4] which can be adapted here with some adjustments. A proposal for the design process is depicted in Figure 2. The first step will be to identify possible and desired emotional responses to the system. These can be extracted from the works of Lavie and Tractinsky, Kim et al., and Papachristos et al. [10, 12, 19]. As attention and information recall are important for the success of this system, these aspects should be taken into consideration as well as proposed by Lee and Benbasat [13]. Additional emotional dimensions can be acquired via an empirical study with the already existing prototype of the system, as done by Kim et al. for websites [10]. The desired emotional dimensions can also be identified in a study.

Desmet et al. have done this for consumer products [4]. The next step of the process is the design of prototypes. In a first phase, general prototypes are designed which act as a starting point for adjusting the design to specific organisations. When preparing the system for an organisation, these prototypes are enriched with elements from the corporate design like colour schemes, logos, and typefaces. Especially for the design of the general prototypes, guidelines can be derived from the work on websites depicted above. These guidelines can be found in the works of Thielsch et al., Lee and Bensabat, Ngo et al., and Kim et al [10, 13, 17, 23]. These prototype designs then have to be evaluated in an experiment to ensure the conformance of the evoked emotions to the desired ones. The potential users can be broken down into several user groups. This is important due to the subjectivity of the emotional experience.

The measurement of emotions in these studies can be conducted with several instruments. These are pictorial tools, questionnaires, and physiological methods [3, 5]. Questionnaires have been used in the studies on websites depicted above; pictorial tools have been used in studies on consumer products [3, 4]; and physiological methods like facial electromyography and electrodermal activity have been used for graphical user interfaces [1, 15]. Mahlke et al. suggest that a combination of different methods is a promising approach [15].

The last step of the design process is the selection of appropriate designs for the implemented system. This may refer to one design which is chosen for the system in whole. Another way is to choose several designs, each emotionally fitting best for one or several user groups. If a Public Display is capable to adapt to the audience in front of the display, not only the information might be chosen adequately but also the design. For example businesspeople might prefer another design than families as tolerance of information technology is higher.

5 CONCLUSION

This paper presents an approach to the emotional design of Public Displays. The role of emotions for such systems is discussed, empirical studies in the related domain of websites are summarized, and a design process for Public Displays is proposed. Although the research on emotions and aesthetics of interactive systems has recently become popular, a lot of works in this field are focused on the constructs itself and on the relationship with other aspects such as usability [9, 23, 25]. Published results on guidelines for designing a particular system with desired emotional outcome are rare. There are also critical voices. Hassenzahl for example states that one cannot design the emotional experience directly; only the design for an experience is possible [7]. This is due to the fact that emotions are highly subjective and context-dependent. Hassenzahl also assumes that the fulfillment of peoples' needs by a system promotes positive emotions. Thus, although the aesthetics of Public Displays have to be taken into account, the success in the long run might most likely be more correlated with instrumental aspects like the fulfillment of peoples' information needs.

REFERENCES

[1] Benedek, J., Hazlett, R.L. Incorporating Facial EMG Emotion Measures as Feedback in the Software Design Process. In Proc. Human Computer Interaction Consortium, (2005).

[2] Carroll, J.M. Beyond fun. Interactions 11, 5 (2004), 38- 40.

[3] Desmet, P.M.A. Measuring emotion. In Blythe, M., Monk, A., Overbeeke, K., and Wright, P. eds. Funology: From Usability to Enjoyment, Kluwer Academic Press (2003).

[4] Desmet, P.M.A., Overbeeke, C.J., and Tax, S.J.E.T. Designing products with added emotional value; development and application of an approach for research through design. The Design Journal 4, 1 (2001), 32-47.

[5] Dormann, C. Affective experiences in the Home: measuring emotion. In Proc. Home Oriented Informatics and Telematics 2003, (2003).

[6] Hrtel, C., Zerbe, W.J., and Ashkanasy, N.M. Emotions in Organizational Behavior. Lawrence Erlbaum Associates, 2004.

[7] Hassenzahl, M. Emotions Can Be Quite Ephemeral. We Cannot Design Them. Interactions 11, 5 (2004).

[8] Hassenzahl, M. Hedonic, emotional, and experiential perspectives on product quality. In Ghaoui, C. ed. Encyclopedia of Human Computer Interaction, Idea Group (2006), 266-272.

[9] Hoffmann, R., Krauss, K., A Critical Evaluation of Literature on Visual Aesthetics for the Web. In Proc. 2004 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries, ACM (2004), 205-209.

[10] Kim, J., Lee, J., and Choi, D. Designing emotionally evocative homepages: an empirical study of the quantitative relations between design factors and emotional dimensions. International Journal of Human-Computer Studies 59, 6 (2003), 899-940.

[11] Landry, J. The Affective Component of Organizational Information Systems, Metropolitan State College of Denver, Computer Information Systems Department, (2005).

[12] Lavie, T., Tractinsky, N. Assessing dimensions of perceived visual aesthetics of web sites. International Journal of Human-Computer Studies 60, 3 (2004), 269-298.

[13] Lee, W., Benbasat, I. Designing an Electronic Commerce Interface: Attention and Product Memory as Elicited by Web Design. Electronic Commerce Research and Applications 2, 3 (2003).

[14] Mahlke, S. Understanding users' experience of interaction. In EACE '05, University of Athens (2005), 243 - 246.

[15] Mahlke, S., Minge, M., and Thring, M. Measuring Multiple Components of Emotions in Interactive Contexts. In Proc. CHI 2006, ACM Press (2006), 1061-1066.

[16] Müller, H.J. and Krüger, A. Towards Situated Public Displays as Multicast Systems. UbiqUM 2006 Workshop on Ubiquitous User Modeling, The 17th European Conference on Artificial Intelligence.

[17] Ngo, D.C.L., Teo, L.S., and Byrne, J.G. Evaluating Interface Esthetics. Knowledge and Information Systems 4 (2002), 46-79.

[18] Norman, D.A. Emotional design: why we love (or hate) everyday things. Basic Books, 2004.

[19] Papachristos, E., Tselios, N., and Avouris, N. Modeling perceived value of color in web sites. In Proc. SETN 2006, Springer (2006), 567-570.

[20] Perlusz, S. Emotions and technology acceptance.. Development and validation of a technology affect scale. In Proc. IEEE International Engineering Management Conference, (2004).

[21] Rafaeli, A., Vilnai-Yavetz, I. Emotion as a Connection of Physical Artifacts and Organizations. Organization Science 15, 6 (2004), 671686.

[22] Schenkman, B.N., Jönsson, F.U. Aesthetics and preferences of web pages. Behaviour & Information Technology 19, 5 (2000), 367-377.

[23] Thielsch, M.T., Schrameyer, M., and Ullmann, A. "Websthetik" - eine empirische Untersuchung zur wahrgenommenen sthetischen Qualitt von Webseiten. 7th General Online Research (GOR), (2005).

[24] Tractinsky, N., Katz, A.S., and Ikar, D. What is beautiful is usable. Interacting with Computers 13 (2000), 127-145.

[25] Tractinsky, N., Towards the Study of Aesthetics in Information Technology. In Proc. Twenty-Fifth International Conference on Information Systems, (2004), 771-780.

[26] Tractinsky, N., Cokhavi, A. and Kirschenbaum, M., Using Ratings and Response Latencies to Evaluate the Consistency of Immediate Aesthetic Perceptions of Web Pages. In Proc. Third Annual Workshop on HCI Research in MIS, (2004).

[27] Zhang, P. and Li, N., Love at first sight or sustained effect? The role of perceived affective quality on users' cognitive reactions to information technology. In Proc. Twenty-Fifth International Conference on Information Systems, (2004).

[28] Zhang, P. and Li, N. The importance of affective quality. Communications of the ACM 48, 9 (2005), 105-108.

Social Robots as Interface with Smart Environments

Giovanni Cozzolongo and Sebastiano Pizzutilo[1]

Abstract. Interaction with Smart Environment (SE) should be natural and easy in order to be effective. In this paper we propose the use of a social robot as interface between users and services of a SE. We focus, in particular, on the need for the robot to comprehend the user's intention in order to respond accordingly. Since the speech is considered one of the more natural and immediate input channel in human-robot interaction we discuss the importance of recognizing, besides the linguistic content of the spoken sentence, the valence of the user tone of voice in order to understand properly the user's communicative intention during the interaction**.**

1 INTRODUCTION

Following the Ambient Intelligence vision [35], a Smart Environment (SE) has the main aim of facilitating users in interacting with its services by making their fruition easy, natural and adapted to their needs. However, most of the times user interfaces for handling the functions and services of SE requires navigation through menu options just to switch off the lights [33] or a complex setting procedure in order to change the behavior of the environment in typical scenarios.

In this paper we presents the first results of a study aiming at using a Social Robot as an interface between users and SE services. This choice was made according to the results of a previous project in which we developed a MultiAgent System for controlling a smart home environment [11]. In this context, the interaction with home services was mainly triggered by sensor data obtained from the user and other observations in the world (e.g. location, noise, temperature, light conditions) combined with reasoning on explicit user action in certain contexts. To this aim, we developed as counterpart of the interaction an intelligent agent acting "behind the scene". This agent was able to infer user needs from his/her actions and, reasoning on the current context, to answer to the user in the appropriate way by changing the state of the environment. However, after the evaluation phase of the project, 80% of subjects declared to feel uncomfortable to interact with an invisible presence and without explicit control over the home services. They declared to prefer to have as a counterpart an explicit interface to the environment in order, to request services, clarify some potential misunderstanding about task execution, express their approval and disapproval and change the behavior of the environment accordingly [12].

Afterward, according to several research studies about the topic [17, 40], we decided to introduce the figure of a mediator between the user and the environment; in particular we decided to employ a social robot as interface between the two participants.

These robots can be thought, on one hand, as a mobile and intelligent interface to the environment system [7, 39]. On the other hand, they embody the role of friendly companions [10]. Sony AIBO and iCat have been created for this purpose [3,38].

In fact, according to the 'Computers As Social Actors' (CASA) paradigm, the interaction with technology is driven by rules that derive from social psychology and can be applied directly to human-technology interaction [34]. In our opinion these aspects become even more relevant when media are not boxed in a desktop computer but are integrated pervasively in everyday life environments. However, when the robot has the role of facilitating the interaction with smart environment services, besides being social it has to be useful. In this case, it is necessary to consider as requirements the awareness of the user and environment situation, the recognition of his/her intentions and the generation of strategies and plans for satisfying the recognized user goals.

Then, the long term goal of our research is to develop an EBDI (Emotion, Belief, Desire and Intention [20]) mind for this type of social robot and, after having designed the basic architecture of its mind [13] we started to develop its behavioral models. As a first important and challenging step, we faced the problem of the recognition of the user's intentions.

Since speech is considered one of the most natural ways of interacting with a robot [14, 28], we take into account two sources of knowledge deriving from the user spoken input: the **linguistic information content** and the **acoustic features** of the speech.

In fact, in order to express their intentions humans use words and transfer emotions and emphasis by modulating their voice tone. According to several studies [25, 28], the linguistic analysis is not enough to understand which is the real user's intention toward the robot and the environment.

While words still play an important role in intention recognition, taking into account which is the user attitude while speaking adds another source of knowledge that is important for disambiguating the interpretation of the real user's intention.

Here we present results of an empirical study demonstrating the feasibility of modeling user's intention recognition starting from the analysis of a corpus of human-robots spoken dialogues.

First of all, as shown in the first Section of the paper, we collected some evaluation evidences about the appropriateness of having a social robot as mediator.

Starting from these results we designed and implemented a prototype of the intention recognition module that is described in the next section. Then we focus on the description on how we built the intention recognition model and how we validated it showing the results of an evaluation study. Conclusions and future work directions are discussed in the last Section.

[1] Dipartimento di Informatica – Universita' di Bari

2 EVALUATING SOCIAL ROBOTS IN MEDIATING BETWEEN USERS AND THE ENVIRONMENT

In the initial phase of this research our aim was to understand whether using a social robot was more effective for people than interacting with an invisible agent representing the environment.

In order to collect some evaluation data, we performed an experiment based on a Wizard of Oz (WOZ) approach [6, 22, 30].

First of all we considered as a smart environment our research laboratory. We thought it was an appropriate place since it is well equipped for performing this type of experiment. Moreover, typical users of this environment belong to different categories and age and attend the lab with different goals: looking for people, papers, meeting, etc.

In this testing experiment we used two groups of 5 subjects each with an age between 20-28 years old equally distributed in gender and background. The Wizard was represented by the same person in all the experiments. We assigned to each group the same goal:

finding a paper that a professor left in the lab for him/her and complete the set of references within 10 minutes using a computer connected to Internet.

The subjects belonging to the first group were told they could use "the environment" for getting help and directions. In particular they could interact with an invisible agent controlling the environment through a speech based interface.

The second group of subjects was told that they could use AIBO for directions and help. In this way we wanted to collect data showing the positive impact of having a social robot as an interlocutor in interacting with environment services.

Table 1. Outline of the survey questionnaire

Questions
Q1: Did you think *the environment agent/Aibo* was intelligent?
Q2: Did you think *the environment agent/Aibo* was sociable?
Q3: Did you find effective the information provided to you by *the environment agent/Aibo*?
Q4: Did you feel comfortable during the interaction with *the environment agent/Aibo*?
Q5: Would have preferred to use another interaction mean? If yes, which one?

In both cases, at the end of the experiment, subjects had to fill a final questionnaire whose questions are listed in Table 1.

With the exception of Q5, the answer to each question was expressed in a 5-point Likert scale (from 1 – not at all to 5 – a lot).

This questionnaire was aiming at collecting a subjective evaluation of the interaction in both modalities.

Results in Table 2 seem to show that there was not a significative difference in terms of effectiveness in receiving information and of intelligence. However the social aspects and the sense of comfort during the interaction were higher when a social robot was used as mediator between the user and the environment. As far as question Q5 is concerned, 3 subjects in the A group answered "yes". Two of them indicated an embodied character as possible interface. The third subject proposed the use of a mobile phone.

Table 2. Results of the study for the two group of subjects. A = "invisible agent" group; B= AIBO group.

Question	**mean-A**	**mean-B AIBO**	**t-test**
Q1: intelligent	2,6	3,2	0,1
Q2 : sociable	1,8	3,4	0,01
Q3: effective	2	2,6	0,35
Q4: comfort	2,4	3,2	0,02

According to these findings we considered important to go on with our research and to adopt a social robot, AIBO in this case, as counterpart of the interaction.

3 UNDERSTANDING USER'S INTENTIONS

Speech is a natural way, for humans, to interact with robots. When the robot acts as mediator with the environment, it receives in some way a delegation from the user to perform some tasks [9]. Moreover, since humans most of the times establish a social relation with the robot, they expect, from it, an appropriate social response. Therefore, during the interaction the robot has to be able to understand which is the expectation of the user and, while their social relation evolves, the robot should learn from the user's feedback and correct its behavior accordingly.

Speech conveys two main types of information: it carries linguistic information according to the rules of the used language and paralinguistic information that are related to acoustic features such as variations in pitch, intensity and energy [5, 23, 27].

Usually the first component conveys information about the content of the communication and the second one about the user's attitude or affective state.

Starting from the work of Breazeal and Aryananda [4] and taking into account some theories that state the importance of the voice tone in interacting for educative and training purposes with preverbal infant [16] and domestic pets [36] we decided to start investigating how to recognize user's intentions from spoken input.

Research in emotional speech has shown that acoustic and prosodic features can be extracted from the speech signal and used to develop models for recognizing emotion. Much of this research has used acted corpus of speech as training data and their research did not take into account the semantic content of what being conveyed [24, 29].

According to Litman [25], in natural interactions users convey emotions by combining several signals. Therefore, acoustic-prosodic features should not be considered alone but combined with other information sources such as the linguistic content of the spoken sentence. Indeed, while acoustic-prosodic features address how something is said, lexical features represent what is said and, together, these features have shown to be useful for recognising intentions in human communication [28].

To this aim we have coupled the linguistic parser with an acoustic analyzer able to extract the prosodic features of the user spoken input. Then, using a probabilistic model based on a Bayesian Network (BN), our system infers the user's intention by combining these two knowledge sources.

For instance, if the user says to the robot "where are you going?" with a neutral or positive attitude the recognized intention should be: the user wants to know where the robot is going but he/she is not disapproving the action. Instead, if the same sentence would be pronounced with a negative/angry attitude this should be interpreted as: the user disapproves the robot action, the robot should stop and interpret the sentence as a negative feedback and revise its belief set accordingly.

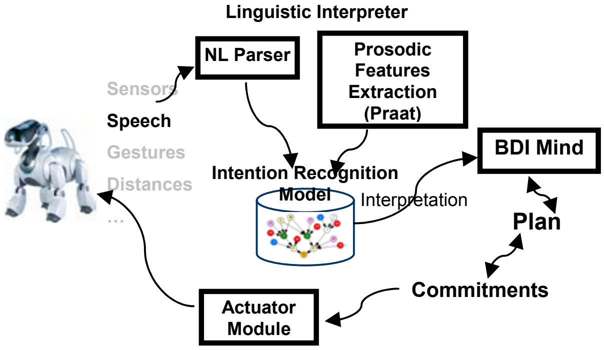


Figure 1. Intention Recognition Schema .

Figure 1 illustrates the schema architecture of the speech based intention recognition module.

The transcript of the spoken input is analyzed by the linguistic parser in order to recognize the type of user's move (i.e. inform, ask, etc.). At the same time the audio file relative to the user's move is analyzed using Praat functions [32].

Once we have extracted the parameters relative to pitch, intensity and energy of the sound, these, together with the interpreted category of move become the evidences or the initial distribution (in case of uncertainty in parsing the sentence) of some of the variables of the BN that allows inferring the attitude and intention relative to the user's move.

Since the main aim of this research does not concern the prosodic features analyser but the one of understanding how these can improve intention recognition, we will focus the rest of the description on model building and on its validation in the context of the interaction with a smart environment.

4 LEARNING THE INTENTION RECOGNITION MODEL STRUCTURE

There is a substantial body of literature on emotion recognition from speech [2]. These works show how certain emotional states are often correlated with particular acoustic features such as pitch (fundamental frequency F0), energy, timing and voice quality.

Obviously, the quantity and types of features to utilize depend on how many emotions or attitudes are relevant for the purpose of the system.

For instance in the work of [4], which is also aimed at recognizing affective intent in human-robot interaction, it was important to distinguish among four categories of affective intents: approval, prohibition, attention, comfort. These categories were selected since the main aim of the work was to "educate" the robot like parents do with their children [16]. According to these results, pitch mean and energy variation seem to work with a good accuracy in distinguishing between the mentioned categories. Therefore, as a first step in building our model, we identified the classes of variables that were important for achieving our goal:

- the typical *input moves* used to interact with the robot in a smart environment;
- the *communicative intention* to be recognized
- the user *attitudes* that could influence intention recognition.

4.1 Collecting data

Since our aim was to identify the basic set of user move, intention and attitudes and the relations among them, we performed another set of WOZ studies.

In this testing experiment we used two groups of 8 subjects each with an age between 20-28 years old equally distributed in gender and background. The Wizard was the same in all the experiments. The goal to achieve by all the subjects was the same as for the first study. However, this time the behavior of the robot was different for the two groups.

For the *first group* of subjects AIBO was behaving in a cooperative way: the robot was helping the user in achieving his/her goal. In particular AIBO was greeting the subject, showing him/her where the professor left the paper and which computer could be used to search for the missing references. During this subtask AIBO was near the subject observing what he/she was doing, ready to provide help on request without disturbing the subject during task execution.

For the *second group* AIBO was behaving not cooperatively. In particular, AIBO was not helping the user in achieving his/her goal by ignoring the subject requests for three times before answering. Moreover, when the subject was busy in completing the missing references of the paper AIBO was trying to distract him/her from task execution.

In our opinion, this difference would elicit two sets of users' behaviours that could emphasise interaction situations in which the subject approved or disapproved the robot behaviour.

Before starting with the experiment we administered to subjects a simple questionnaire aiming at collecting some personal data (age and gender) and at understanding their background (department, year of course, artificial intelligence background).

Following the described approach we collected a corpus of 592 moves (25 moves on average for subjects belonging to the first group, 49 for those belonging to the second one).

Each move was recorded using a wireless microphone whose output was sent to the speech processing system. We assigned to each utterance, a unique identifier, the correspondent transcript and the related 16-bit single channel, 8 kHz signal (in a .wav format).

The resulting dataset has been analyzed by 6 human annotators. Since we are interested in finding the relations between the type of user move, the corresponding attitude and intention, we used 2 annotators for each step of this process so that the labelling of each component was not influenced by the others.

Step 1. As far as the type of communication content is concerned, the possible categorical label to be used have been extracted mainly from the Speech Act Theory [36] and are listed in Table 3.

Table 3. Categories of user moves

Move Type	***Function***
Greet(U,AIBO)	the user U greets AIBO
Call(U,AIBO)	the user calls AIBO
Request(U,AIBO,*a*)	the user asks AIBO to perform an action *a*
Order(U,AIBO,*a*)	the user orders to AIBO to perform an action *a*
Inform(U,AIBO,*f*)	the user informs AIBO about a fact *f.*
Ask(U,AIBO,*f*)	the user makes a question to AIBO about a fact *f.*
Thank(U,AIBO)	the user thanks AIBO.
Reproach(U,AIBO)	the user reproaches AIBO's behaviour
Compliment(U,AIBO)	the user makes a compliment to AIBO.

We provided the annotators with the set of human written transcripts of all subjects moves collected during the WOZ study in both modalities and we asked them to use the labels in Table 3 to annotate them. They could also introduce new labels if they did not recognize in the move any of the speech acts listed.

Table 4. Inter-Annotator agreement.

Label	**% agreement**	**kappa**
Greet	1	1
Call	0.9	0.8
Request	0.7	0.4
Order	0.8	0.6
Inform	0.7	0.4
Comment	0	-1
Ask	0.9	0.8
Thank	1	1
Reproach	0.8	0.6
Compliment	0.9	0.8

In order to test the validity of our annotation we followed the method found in [9]. Then, to have a measure of the level of agreement between annotators, we calculated the percentage of cases that were labelled in the same way by them, we computed the percentage of agreement and then we calculated the *Kappa statistics*.

Kappa is widely accepted in the field of content analysis and allows different results to be compared.

Table 4 summarizes inter-annotator agreement. Apparently the annotators agreed in recognizing most of the moves from the communicative content.

However, the level of agreement about the "inform" and "request" speech acts was lower than we expected.

Step 2. Our main goal was to identify which were the user's attitudes that could change the linguistic interpretation of the intention and that in someway could provide a feedback to the robot and the environment.

For our purpose we did not need a very sophisticated distinction between all the possible user affective states. In our opinion, in this type of interaction it is important to understand which is the valence of the user attitude.

Therefore we considered only the negative, neutral and positive categories measured along a 5-point scale (from 1- very negative to 5- very positive).

The ***positive*** attitude was important to reinforce positive feedback toward the robot and the environment while the ***negative*** one was important to disapprove the robot or the environment actions.

The ***neutral*** attitude was considered important for interpreting the user intention only from the linguistic part of the user input.

Table 5. Categories of user moves.

	% agreement	**kappa**
very positive	0.9	0.8
positive	0.8	0.6
neutral	0.62	0.24
negative	0.8	0.6
very negative	0.98	0.96

The annotation was performed as for the previous step except for the fact that we asked the two annotators to label each move according to the 5-point scale mentioned above by listening to the audio file.

Results are shown in Table 5. Apparently the annotators agreed clearly in recognizing strong attitudes (positive and anger) while there was a strong disagreement for the neutral one.

Step 3. Then the last step was to understand which intention categories were relevant in our domain in order to provide the annotators with a set of labels to be used in their task. To this aim we identified the ones reported in the first column of Table 6. The annotation process was the same also for this step, the percentage of agreement and the results of the kappa statistics is shown also in Table 6.

Step 4. After the human annotation process, the dataset needs to be completed with the related acoustic features. We used Praat functions [32] in order to extract from the audio file of each move the following features:

- related to fundamental frequency (f0): pitch minimum, mean, maximum and standard deviation;
- 2 related to energy (RMS): max, standard deviation.

We did not consider the speed rate since our sentence where very short and this parameter seemed to be not relevant.

Table 6. Results of the intention annotation phase.

Intention Type	***Function***	***agreement***	***kappa***
WantToDo(U,AIBO,*a*),	with U denoting the user, AIBO the robot, *a* the action that the user wants the robot to perform in the environment; the action *a* may be performed by the robot directly or by the environment devices through the robot as interface;	0.7	0,4
WantToKnow(U,AIBO,*f*)	where *f* denotes a fact that the user U wants to know;	0.75	0,5
Approve(U,AIBO)	positive feedback corresponding to the intention to approve;	0.7	0,4
Disapprove(U,AIBO)	negative feedback corresponding to the intention to disapprove;	0.75	0,5
GetAttention(U,AIBO)	indicates the intention to get the attention of the robot;	0.8	0,6
GetComfort(U,AIBO)	indicates the intention to be soothed by the robot or by the environment;	0.6	0,2
GiveSocialCue(U,AIBO,*c*)	corresponds to the intention of performing a social communicative act such as greeting.	0.5	0

After adding these features to the dataset, it was necessary to transform the numeric values relatives to the pitch and energy into discrete values in order to handle these data in our model.

To this aim we used a three-value scale (low, normal, high). In order to assign each numerical value corresponding to the pitch and the energy values to one of these discrete values we calculated the 33% and 66% *percentile*. We divided in this way the numeric interval of each of the extracted features into three parts. Then, values falling into the first numeric set were considered as *low*, those falling in the second one as *normal* and the rest *high*.

For instance the final annotated entry for the question “what are you doing?” expressing the intention to disapprove the robot action with a very negative voice tone is the following:

gender	**move**	**intention**	**attitude**	**p_min**
f	ask	disapprove	vneg	high
p_mean	**p_max**	**p_var**	**e_var**	**e_max**
high	high	normal	high	high

4.2 Building the model

In order to learn the dependencies among acoustic features, linguistic content of the user move, attitude and intention, we used the NPC learning algorithm of Hugin 6.5[2] on the labelled database of cases in the corpus.

In particular, we randomly extracted 37 cases from the labelled database of selected cases (37 was the average of moves for each subject during the experiment). This subset was used in the testing phase, as we will show later on, while the rest of cases was used for learning the structure of the Bayesian Network.

In our opinion the use of a probabilistic model was appropriate in this context. In fact, understanding human attitude and intention from speech input involves capturing and making sense of imprecise and sometimes conflicting data. Moreover we expect that, as it happens between humans, it is not always possible to recognize the attitude and the intention without mistakes. For example, humans are able to understand the speaker’s emotion correctly from the voice tone in 65% of the cases [31].

Furthermore, recognizing the attitude and intention taking into account uncertainty allows planning the answer in a probabilistic way, setting different priorities for the different triggered goals according to the probability of these two variables. Then, these priorities can be updated dynamically according to the increasing/decreasing of the uncertainty of a state of some variables, as it usually happens in a smart environment.

The model resulting after some optimization steps is shown in Figure 2.

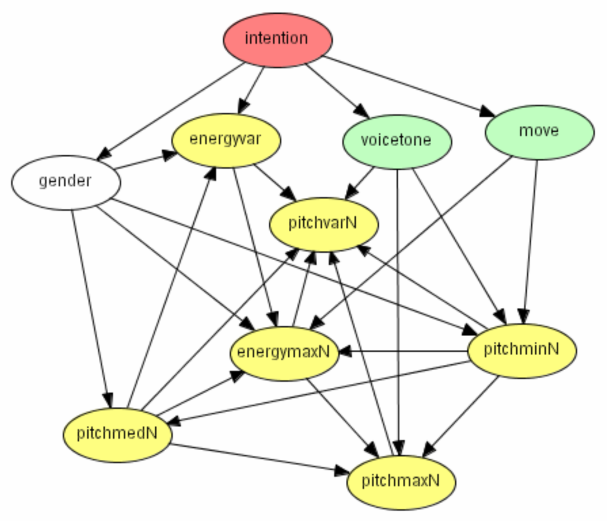


Figure 2. Intention Recognition Model.

In our case, variables introduced are mainly related to:

a. the recognized move: it is extracted by the NL parser and belongs to one of the category listed in Table3;

b. the extracted *acoustic features*: pitch features and energy variation and maximum were considered relevant to attitude and therefore intention recognition;

c. the *gender* of the user which was the only identification feature that we considered in our dataset; prior studies [23, 29] have shown that gender can play an important role in emotion recognition.

[2] www.huginexpert.com

d. the *voice tone*: this variable can assume values in the set very positive (5), positive (4), neutral (3), negative (2) and very negative (1);

e. the *intention* beyond the speech act: this is the variable that we want to monitor using the model and can assume values in the set described in Table 6.

5 INTENTION RECOGNITION: EVALUATING THE MODEL

In validating our model we performed two types of experiments. The first one was performed validating the model on a subset of data extracted randomly from the selected corpora. The second one was performed in order to validate the model against new user's data.

Experiment 1. In this first experiment we selected randomly 37 entries from our annotated corpus (37 was the average of moves for each subject). Then we used as input evidences of our model the following values: first only the speech features, then only the linguistic move and then both of them. Finally we compared the predicted result with the human annotation in all the three cases.

Table 7 shows the mean accuracy expressed as percentage of correct cases. Since we use a probabilistic model we considered as correct a prediction of the value of the intention variable when the probability of one of its seven states was above 0.50.

Table 7. Accuracy of intention recognition.

test	% correct
speech	50%
move	58%
speech+move	78%

As a general comment we can say that the model predicts better using **speech + move** vs. the other two conditions especially in those cases when the linguistic content and the voice tone contrast in identifying the type of intention.

For instance, the sentence "where are you going?" is parsed as "Ask(U,AIBO,where(is_going,AIBO))" gave as output that the most probable (47% on 8 possible choices) user's intention is "Want_to_Know(U,where(is_going,AIBO))" and the most probable valence of the user's voice tone is "neutral" (40% on 5 possible choices). While the probability of the disapprove intention is around 16%.

However, if we add evidences about acoustic features, then the most probable intention becomes the disapproving (51%) and the most probable valence of the voice tone becomes negative (89%).

Experiment 2. The aim of the second experiment was twofold: first of all we wanted to test the model against new voices and, secondly, to observe how incremental learning could help in improving intention recognition.

Instead of performing a new set of WOZ studies, in the second experiment we used an acted corpus. Indeed we wanted to be sure to test our model for particular situations in which the NL parser was recognizing a move type whose underline intention was different from the one expressed by tone of voice.

According to Abelson's functionalist model of emotion [1], the relation between goals, context and outcome is important in emotion appraisal. According to this theory Enos and Hirschberg [15] have proposed an approach to elicit acted emotional speech that is based on the idea of providing to the actor the goal and the context in which to play certain actions. Since our aim was to infer user intention from speech, this approach seems to be appropriate. Then, we asked to our actors (1 male and 1 female) to play the following sentence "What are you doing?" in order to express the following intentions: i) *want_to_know,* ii) *approve* and iii) *disapprove*.

The written sentence was given to the NL parser whose output was "Ask(U,AIBO,what(is_doing,AIBO))". Then we classified as low, medium, high the acoustic features according to the method explained previously.

Results in this case were worst than in the first experiment. As shown in the first column of Table 8 the model did not recognize the approval intention which had almost the same probability of the want_to_know case.

We suspected that this result was due to the fact that the features of the acted voice were different from the one in the training dataset. For this reason we performed the EM learning on this data (10 iterations) and we tested again the model. The second and third columns in Table 8 show the improvement in the percentage of intention recognition.

Table 8. Accuracy of the intention recognition model on new data.

Intention	% correct	% correct I iteration	% correct II iteration
disapprovement	59 %	67 %	72 %
approvement	28 %	39 %	48 %
want_to_know	51 %	60 %	69 %

6 CONCLUSIONS AND FUTURE WORK

As part of ambient intelligent research, there is a need to recognize the user's intention during the interaction in order to adapt the environment behavior accordingly.

One way to do it is to observe user's actions in the environment and infer his/her intention in a transparent way. Another way is to use a kind of "mediator" agent that is considered responsible for the success of interaction between the user and the environment.

In this paper we presented our first results in recognizing user's intention when this mediator is represented by a social robot. This choice was supported by the result of a preliminary evaluation study aiming at understanding whether using a social robot was better than interacting with an "invisible" agent.

Since spoken input is considered one of the more natural ways to interact with robots we focused our research on the analysis of the spoken sentence using two information sources: from the linguistic content and the valence of the tone of the voice. In particular, we performed one set of experiments based on the Wizard of Oz method whose results were annotated and analyzed in terms of linguistic communication content, valence of the voice tone and intention.

This collected corpus was used to learn the structure of the Bayesian network model to be used for recognizing the probability for a user to have a particular intention toward the robot and/or the environment.

In order to validate this model we performed two experiments. The first one was performed using as testing dataset a randomly extracted subset of the corpus. The second one was executed using acted sentences expressing situation that showed an evident contrast between the parsed sentence and the underlining intention.

From the performed experiment we collected two types of results, one showing that using both knowledge sources for recognizing the user's intention improves the prediction accuracy of the model and the other concerning the fact that incremental and continuous learning is important if we want to build speaker independent models.

In our future work we plan to improve the model by taking into account the dynamicity typical of the domain and by adding some contextual features that can influence the recognition of intention. In particular, we need to investigate on how previous intentions influence the current one so as to express this relation as a function to build a temporal link in a Dynamic Belief Network (DBN, [19]).

As far as the model of the user intention is concerned, we need to evaluate its effectiveness in real context of use. In fact, we plan to partially overwrite the AIBO mind in order to give it the capability to use this model and to plan for the most appropriate actions accordingly.

The overall response of the users to the experiments seems to confirm that the idea of using an affective robot as a mediator is a good way to overcome barriers that people may find in using smart environments.

ACKNOWLEDGMENTS

Research described in this paper is an extension of the work we performed in the scope of the National Project Prin03.

REFERENCES

[1] Abelson, R.P. What Ever Became of Consistency Theory? Pers. and Social Psych. Bull. 9(1), 1983, 37-54

[2] Banse, R. and Scherer,K.R.. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology. 1996.

[3] Bartneck, C., & Reichenbach, J. Subtle emotional expressions of synthetic characters. The International Journal of Human-Computer Studies (IJHCS), 62(2), pp. 179-192. 2005.

[4] Breazeal C. and Aryananda L. "Recognizing affective intent in robot directed speech", Autonomous Robots, 12:1, pp. 83-104. 2002.

[5] Bretan, I., Ereback, A.L., MacDermid, C., and Waern, A. Simulation-Based Dialogue Design for Speech-Controlled Telephone Services. Proceedings of CHI'95. 1995.

[6] Buisine, S. and Martin, J.C. Experimental Evaluation of Bi-Directional Multimodal Interaction with Conversational Agents. Proceedings of Interact'03. 2003.

[7] Buhmann, J., Burgard, W., Cremers, A. Fox, T. H. D., Schneider, F.,J. Strikos, and S. Thrun. The Mobile Robot Rhino. AI Magazine, 16(1):31–38, 1995.

[8] Carletta, J. C.. Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics, 22(2), 249-254. 1996.

[9] Castelfranchi, C. and R. Falcone, Towards a theory of delegation for agent-based systems, Robotics and Autonomous Systems 24 (1998), 141-157.

[10] Dautenhahn K. Robots as social actors: Aurora and the case of autism. In Proc. CT99, The Third International Cognitive Technology Conference, August, San Francisco, pages 359-374, 1999.

[11] De Carolis B., Cozzolongo G, Pizzutilo S., Plantamura V.L.: Agent-Based Home Simulation and Control. In Proceedings of ISMIS 2005, LNCS. Springer: 404-412.

[12] De Carolis B., Cozzolongo G, Pizzutilo. A Butler Agent for Personalized House Control. In Proceeding of ISMIS 2006, LNCS. Springer. (in press).

[13] De Carolis B., Cozzolongo, G. Social Robots for Improving Interaction in Smart Environments. In Proocedings of the Workshop Emotion in HCI. In conjunction with the 20th British HCI Group Annual Conference. (in press).

[14] Drygajlo A., Prodanov, P.J.,Ramel G., Meisser, M. and Siegwart, R.On developing a voice-enabled interface for interactive tour-guide robots. Journal of Advanced Robotics, vol.17, nr. 7, 2003,p.p. 599-616.

[15] Enos F. and Hirschberg J. A framework for eliciting emotional speech: Capitalizing on the actor's process. LREC 2006 Workshop on Corpora for Research on Emotion and Affect, Genova, 2006.

[16] Fernald, A.. Four-month-old infants prefer to listen to motherese. Infant Behavior & Development, 8, 181-195. 1985

[17] Gárate, A., Herrasti, N., and López, A. 2005. GENIO: an ambient intelligence application in home automation and entertainment environment. In Proceedings of the 2005 Joint Conference on Smart Objects and Ambient intelligence: innovative Context-Aware Services: Usages and Technologies (Grenoble, France, October 12 - 14, 2005). sOc-EUSAI '05, vol. 121. ACM Press, New York, NY, 241-245.

[18] Ishii, H., and B. Ullmer. Tangible bits: towards seamless interfaces between people, bits and atoms. In CHI'97, pages 234 -- 241, 1997.

[19] F.V. Jensen. Bayesian Networks and Decision Graphs. Statistics for engineering and information science. Springer, New York, Berlin, Heidelberg, 2001.

[20] Jiang, H. and Vidal J.M. From Rational to Emotional Agents. In Proceedings of the AAAI Workshop on Cognitive Modeling and Agent-based Social Simulation, 2006.

[21] D. Kulic and E. A. Croft. Estimating Intent for Human Robot Interaction. Proc of the Int. Conf. on Advanced Robotics. Coimbra, Portugal, June 29 - July 3, 2003.

[22] Whittaker, S. Walker, M. and Moore, J. Fish or Fowl: A Wizard of Oz evaluation of dialogue strategies in the restaurant domain. Language Resources and Evaluation Conference. 2002.

[23] C.M. Lee, S.S. Narayanan, R. Pieraccini, Combining acoustic and language information for emotion recognition. Proceedings of ICSLP, 2002.

[24] J. Liscombe, J. Venditti, and J.Hirschberg, "Classifying subject ratings of emotional speech using acoustic features," in Proc. of EuroSpeech, 2003.

[25] D. Litman, K. Forbes, S. Silliman, Towards emotion prediction in spoken tutoring dialogues. Proceedings of HLT/NAACL, 2003.

[26] Maes, P. 1994. Agents that reduce work and information overload. Commun. ACM 37, 7. (Jul. 1994), 30-40.

[27] Mozziconacci S. and Hermes, D. J."Role of intonation patterns in conveying emotion in speech", in Proceedings, International Conference of Phonetic Sciences, San Francisco, August 1999.

[28] NAKAMURA Satoshi, Toward Heart-to-Heart Speech Interface with Robots. ATR UptoDate. Summer 2003.

[29] P-Y. Oudeyer. The production and recognition of emotions in speech: Features and Algorithms. International Journal of Human Computer Studies, 59(1-2):157ñ183. 2002.

[30] Oviatt, S. and Adams, B.: Designing and Evaluating Conversational Interfaces With Animated Characters. In J Cassell, J Sullivan, S Prevost and E Churchill: Embodied Conversational Agents. The MIT Press, 2000.

[31] Petrushin, V. A. "Emotion in Speech: Recognition and Application to Call Centers", in Proc. ANNIE '99. pp: 7-10. 1999.

[32] PRAAT: www.fon.hum.uva.nl/praat.

[33] Randall, D. (2003) "Living Inside a Smart House: A Case Study," in Harper, R., (editor), *Inside the Smart Home*, Springer-Verlag, London. 227-246.

[34] Reeves, B. and Nass, C. The Media Equation. New York: Cambridge University Press. 1996

[35] Riva, G., Vatalaro, F., Davide, F., and Alcaniz, M. (Editors). Ambient Intelligence: The Evolution Of Technology, Communication And Cognition Towards The Future Of Human-Computer Interaction (Emerging Communication), 2005.

[36] Searle, J. Speech Acts: An Essay in the Philosophy of Language, Cambridge, Eng.: Cambridge University Press. 1969.

[37] Serpell, James. The Domestic Dog: Its evolution, behavior, and interactions with people. Cambridge University Press, New York, New York. 1995.

[38] Sony. (1999). Aibo. from http://www.aibo.com

[39] S. Thrun, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz. MINERVA: A Second-Generation Museum Tour Guide Robot. In Proc. of the International Conference on Robotics and Automation (ICRA), pages 1999–2005. IEEE, 1999.

[40] von Wichert, G. and Lawitzky G. Man-Machine Interaction for Robot Applications in Everyday Environments. In Proc of the IEEE Int. Workshop on Robot and Human Interaction 2001 (RO-MAN 2001), Paris, Bordeaux, Frankreich, 18.-21. Sept. 2001.

An Initial Approach to Modelling Cultural Variability in Conversational Agents

Asad Nazir, Ruth Aylett and Alison Cawsey[1]

Abstract. Cultural differences do occur in communication. In order to facilitate more realistic communication between agents we need to take into account cultural variability. This paper gives a review of major variability models in anthropology and inter-cultural communication and then proposes a preliminary agent model for intercultural communication between conversational agents.

1. Introduction

Culture is an important part of the expression and communication of human feelings [5]. It influences the way every event and object is viewed and the same objects or events can have different conceptions based on the cultural norms and variables. In order to develop affective smart environments which respond to the individual it is useful to understand and model their culture. To do this we are developing agent models which can embody different aspects of culturally influenced personality.

To define cultural differences in agents we have to define certain variables which describe the cultural personality in an agent. There have been definitions of cultural variability by anthropologists such as Hofstede and Hall [11, 12], and these form the basis for our model. We aim to model characters which are able to display cultural properties based on anthropological research.

Culture impacts on many aspects of human behaviour. However, this research focuses on intercultural communication - i.e., communication between characters belonging to different cultures

In order to make agents (representing different cultures) interact, it is necessary to define cultural parameters and formalise how the agent's actions should depend on these parameters. Particular parameter settings then provide a simple way of defining different cultural stereotypes (while recognising that an individual's parameters will often differ from the stereotype).

Many such parameters defining cultural variability have been described in the literature (e.g., [4, 9, 14, 16]), but five have emerged as of particular importance.

These parameters of cultural variability are described in the next section. We then describe the particular issues related to intercultural communication, and our preliminary agent model.

2. Cultural Variability:

Hofstede [12] created a four factor cultural model, which is perhaps the most cited in cross-cultural communication papers. It is most important to the design of the model of communication in the agent system.

2.1 Individualism / Collectivism

Individualism Collectivism is one of the most important cultural dimension which affects behaviour at both cultural and individual level.

Collectivistic cultures emphasize community, collaboration, shared interests, harmony, traditions and public good. This culture can suppress emotions according to the mood of the collection. [15] Body movements and other kinesics are more synchronised. Individualistic cultures emphasize personal rights and responsibilities, privacy voicing one's own opinions freedom, and self expression.

This parameter, and others, may impact on the individual agent's psychological needs [11]. In this case, for example, it will influence the need for affiliation. This provides a link from the general cultural stereotype to the intentions and behaviours of the individual agent.

2.2 Uncertainty Avoidance:

In some cultures freedom produces uncertainty, which leads to stress and anxiety. These cultures may seek to avoid uncertainty by increasing rules of behaviour. Berger et al. [3] suggests that many southern European countries, as well as Japan and Peru, tend towards uncertainty avoidance. Other countries (including many northern

[1] School of Mathematics and Computer Science, Heriot-Watt University Edinburgh.

European countries) are, it is argued, better able to tolerate freedom and diversity without excess stress and anxiety [7]. A culture's rigidity and dogmatism are a function of the uncertainty avoiding dimension. This dimension also influences communication between individuals - particular direct or indirect forms of communication can be used to reduce uncertainty.

The need for certainty will therefore influence actions, including communicative actions, in an individual agent.

2.3 Power Distance:

The members of high power distance cultures tend to see power as a basic fact in society e.g. South Asia, Caribbean, France etc [12] while members of low power distance cultures tend to view that power should be used only when it is legitimate. E.g. European countries which are normally middle class democracies located at high latitudes.

This influences the way people from different cultures communicate with other people with different power distance and standing in the society or organisation.

2.4 Gender.

Gender is a big factor in defining rigidity in cultural roles. Members of cultures high in masculinity tend to value performance, ambitions, power and assertiveness.

Members of cultures high in femininity tend to place more value on quality of life, service, caring for others.

2.5 Context:

The parameters described so far come from Hofstede [12]. However, another pioneer in this research (at least as applied to cross-cultural business communication) is Edward Hall [11].

Hall presents a four factor model, in which cultures are measured on:

1. High vs. Low speed of messages,
2. High vs. low context,
3. Spatial distance,
4. Polychromatic ("multi-tasking") vs. mono-chromatic ("single tasking") approach to time.

From this model context is regarded as particularly important, and it has recently been shown, for example, that the high-low context distinction influences website design [20]. Context refers to the situation or background related to an event or communication. High context communication makes more use of this environment or situation and is less explicit. Context therefore influences the amount of expression which is explicitly represented in a communication in a particular culture. Hall explains properties of high and low context cultures.

In many Western, independent cultures and the languages used (e.g., European-American cultures and the languages such as English), a greater proportion of information is conveyed by verbal content. Correspondingly, contextual and non-verbal cues such as vocal tone are likely to serve a relatively minor role. These cultures are called Low-context cultures.

In contrast, in many Asian, interdependent cultures and the languages used there (e.g., cultures such as Japan, the Philippines, Korea, and China and languages such as Japanese, Tagalong, Korean, and Chinese), the portion of information by verbal content is small and, correspondingly, contextual and nonverbal cues are likely to play a relatively larger role. These languages and cultures are called high-context.

3. Inter-cultural Communication:

Having identified some key parameters defining different cultural stereotypes we can now move on to consider how these impact on communication. This section examines general issues related to intercultural communication, and examines briefly how culture impacts on verbal and non-verbal communication.

Intercultural communication can be simply called the communication between people from different cultures [10]. Because the values and conventions differ in different cultures, their perceptions and interpretations of feelings and models in mind are different. With the growth of globalization, the contact and interaction between people from different cultures has increased manifold and consequently the need for a more fruitful intercultural communication has increased.

Difficulties in Intercultural communication may arise because of variation in the norms of communication in the different cultures. These may arise in both verbal and non-verbal communication. Verbal communication comprises language and the context involved. Non-verbal communication includes body movements and other gestures which certainly differ in different cultures.

3.1 Verbal Communication:

Verbal Communication concerns the use of words to communicate. In different cultures language may be used differently (from politeness conventions to rhetoric). Both the way something is expressed (e.g., verbose versus concise) and the underlying content may be culturally determined. There are, for example, ancient philosophical differences between Eastern and Western Culture on rhetoric contents in communication [10].

At roughly the same time when Confucius and Lao Tze preached the futility of verbalization in the east, Socrates, Plato and Aristotle taught the importance of reasoning and logical persuasion on the other side of the world.

Depending on their culture agents will express themselves differently through words. Recognising and undertstanding a speaker's culture will correspondingly help in properly understanding what they are trying to communicate.

3.2 Non-Verbal Communication:

Non Verbal factors include gesture, eye contact, clothing, and facial expression – the display properties in cultural communication. They differ in different cultures, and being easily recognisable (compared with verbal differences) they serve as the clues for the user to identify the culture (and possibly understand cultural differences). Non-verbal communication is an *immediate behaviour-* more instinctive, and less easy to mask and control.

Behaviour	**Explanation**
Proxemics	Spatial difference while communicating
Kinesics	Movement of body parts
Haptics	Reaction to touch
Physical appearance	Clothing, skin colour, etc.
Oculesics	Use of Eye in Communication
Vocalics	Voice animations
Olfactics	Sense of Smell

Table1: [10] Types of non-verbal behaviours.

Non-Verbal communication is also important for agent to express cultural behaviour and to perceive cultural behaviour as positive or negative. Some Non-verbal cues in some cultures may be positive and in others they might amount to embarrassment.

3.3 Investigating Inter-cultural Communication and Cultural Differences

In order to model and simulate cultural variability in communication, we are designing conversational agents which will interact with each other in some environment and express verbal and non-verbal cues in their communication. By designing conversational agents which embody different cultures and which can interact with each other we aim to:

- Understand better how cultural variability can be formalised, and how it affects communication.
- Have a basis for developing conversational agents which can interact with the user in a culturally sensitive manner.
- Produce a practical demonstrator that can be used to help users understand problems and difficulties in intercultural communication, through observing interacting agents.

To realise the last of the above objectives we will create scenarios that illustrate particular difficulties in intercultural communication. There can be number of situations where the communication becomes difficult because of cultural differences. One of the ways to point out cultural differences is to raise flags or alarms that a particular misunderstanding has happened. These are a concept drawn from experiential techniques in intercultural encounters by Ned Seelye **[17]** who uses these techniques on international participants in universities and also through dramatic performances to get people to understand what problems have they encountered during an intercultural communication session.

To design these agents we need architecture for their behaviour, communication and emotions. The next section proposes a preliminary model for such a conversation based agent based on PSI theory by Dorner

4. Model for Cultural agents:

Our agent model needs to represent how communication and expression depends on cultural parameters. It should be able to show, for example, how communication is influenced by the (varying) need for affiliation and for uncertainty avoidance, and how communication can vary depending on how much contextually derivable information is included.

To model cultural variability we take as starting point existing models supporting the modelling of emotions and other social behaviours such as OCC [16], PSI theory [5, 7, 8], and Oliviera and Sarmento [15]. Out of these models the OCC has been used the most. However, the PSI model is the preferred model here because of its realistic way of emotion generation.

In PSI, emotions are modelled as emerging from the information processing and not as separate constructs. Behaviour emerges on the basis of needs and perceptions from the environment, and emotions are modelled as the mode in which the actions are acted out. Hence, advantages compared to the OCC model are:

- Emotions need not to be modelled separately but emerge from the system.
- There is no need to define a number of relevant emotions.
- Emotions emerge as a consequence of need states instead of linking them directly to events or actions - psychological basis [1].

- This leads to believable dynamics of emotional states that do not rely on thresholds and decay rates (as in OCC) but on current need states that also determine action selection.
- Arousal is already part of PSI and does not need to be calculated from the general intensity of emotions; rather, arousal (or activation) determines the emotional state (from a psychological perspective arousal does not decrease in the absence of intense emotions, but when basic needs are satisfied).

PSI will therefore be used as the basis for our model. PSI theory, has three main parts: motivation, action regulation and emotion modulation. Cultural parameters may influence all these aspects.

Figure 1 gives an overview of our preliminary architecture. The rest of this section describes some of the key components.

4.1 Long Term Memory:

In order to simulate interaction between two agents who belong to different culture the long term memory needs to have cultural parameters describing the culture of the agent. Also if an agent is to be able to reason about culture-related behaviours of another agent (or user) the agent needs to have an explicit representation of other cultural stereotypes in its memory.

4.2 Motivation:

In the cultural agent there can be two main motivators:

- Affiliation. This can be modelled based on the cultural parameters mentioned earlier. For example, an agent belonging to an collectivistic, high power distance culture will be have a higher need for affiliation as compared to an agent belonging to an individualistic, low power culture.

- Uncertainty avoidance. This may relate to the environment to communication and can be another motivator for action selection as the techniques for uncertainty management differ in different cultures.

Other motivators may be considered as the model is refined further.

4.3 Intention Generation:

Intention generation and selection will occur on the basis of the level of need or motivation, the perception of the environment and goal of the agents. Intentions will be selected from a memory of intentions where different intentions are stored. According to Dörner [5] the intentions are calculated with the following formula:

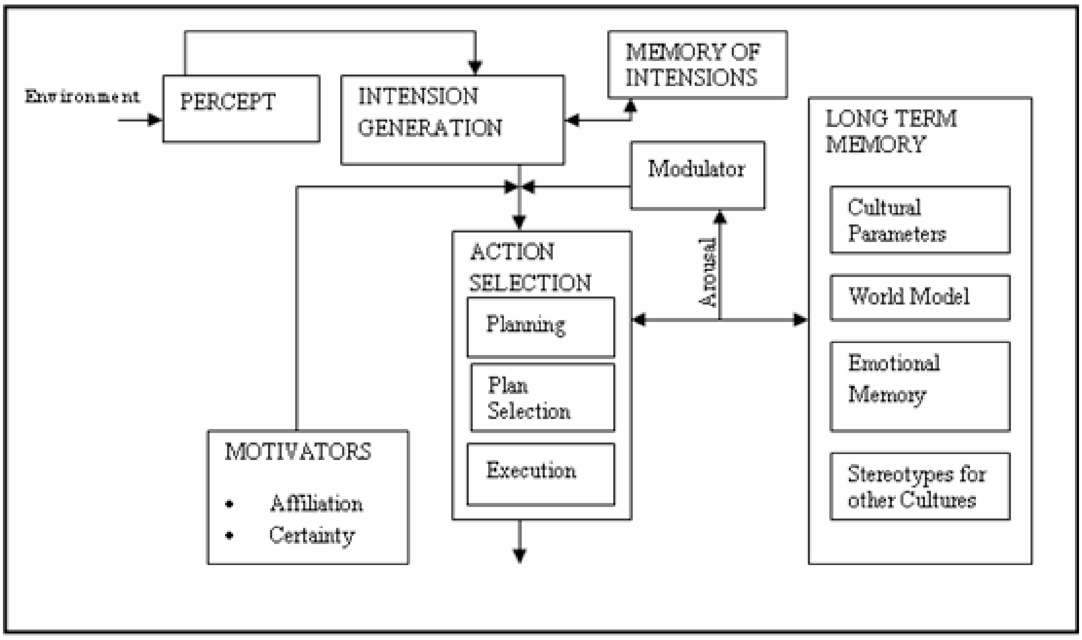


Figure 1: PSI based model for Cultural Agents

$$\mathbf{S_i = \sum (needs*satpot_{goal})* SP * Urgency}$$

In this case the relevant need will be the level of affiliation or certainty required. SP is the Success Probability of achieving a goal.

Calculation of Success and Urgency depend on:

- Perception of actual situation.
- Expectation of upcoming events:
- Experiences regarding goal-related action

4.4 Action Selection:

Once the intentions are selected then the action is to be selected and executed. There are two kinds of actions i.e. an automated action (ritualised) and a planned action based on a planning mechanism which depends on the cultural parameters, emotional modulation, and the stereotype associated with the other agent.

4.5 Emotional Modulation:

These modulations are realised by so called emotional parameters. Different combinations of parameter values result in the subjective experience of emotions. It involves three emotional parameters: Activation, resolution level and Selection Threshold.

- Activation, which is the preparedness for perception and reaction on side of the agent; this parameter increases because of the motivations and active intension values [4]. The concept of activation is similar to the psychological concept of “arousal”.

- Resolution level: [1] It decreases with an increase in activation it determines the accuracy of cognitive processes, e.g. perception, planning, action regulation

- Selection threshold: Prevents the currently active intention to be replaced by another, equally strong intention. It gives priority to current intention. Concentration of the agent depends on this parameter.

5. Conclusion and Further Work

This paper has presented a model of cultural variability and a preliminary agent architecture supporting the representation and reasoning about different cultural personalities. The work is at an early stage, but we hope to have both demonstrated the importance of representing cultural-related aspects in an agent model, and shown how standard anthropological models can begin to be mapped to an agent architecture.

Further work planned includes the refinement and extension of the model to cover other parameters of cultural variability, and implementation of the model. A proof of concept demonstration is planned illustrating difficulties in intercultural communication through simulated agents embodying different cultural personalities, and scenarios are being defined which we hope will illustrate communication difficulties and potential misunderstandings.

We argue that intelligent conversational agents should understand, and potential be able to embody different cultural personalities, and that understanding and responding to culturally influenced aspects of an agent’s personality should go hand-in-hand with understanding and responding to their emotions. We also recognise that any apparent cultural stereotyping in interaction must be done sensitively, recognising that the individual may not conform to the conventions of his or her culture. Communicating in a culturally sensitive way is something that is challenging for humans, and it is perhaps ambitious to hope to be able to provide conversational agents with this property. But at the same time ignoring the cultural dimension may result in misunderstandings and offence. Further work is needed in this area, and we hope that this preliminary model provides a useful starting point.

REFERENCES:

[1] Ayman Elkady, *The Simulation of Action Strategies of Different Personalities In Perspective of the Interaction between Emotions, Motivations and Cognition (An Experimental Study in the Field of Cognitive Psychology and Artificial Intelligence)* Inaugural Dissertation 2006.

[2] Bach Joscha, *Enhaancing Perception and Planning of Software Agents with Emotion and Acquired Hierarchical Categories.* Proceedings of Masho 02, German Conference on AI KI 2002, Karlsruhe, Germany 2002.

[3] Berger, C.R., & Calabrese, R. *Some explorations in initial interactions and beyond.* 1975.

[4] Brislin, R.W. (1993). *Understanding culture's influence on behaviour*. Fort Worth, TX: Harcourt Brace.

[5] Bartl, C., Dörner, D.: *Comparing the behavior of psi with human behaviour in the biolab game.* In Ritter, F.E., Young, R.M., eds.: Proceedings of the Second International Conference on Cognitive Modeling, Nottingham, Nottingham University Press (1998)

[6] Carl Ratner *A Cultural-Psychological Analysis of Emotions Culture & Psychology,* Vol. 6, No. 1, 5-39 (2000) SAGE Publications

[7] Dörner, D.: *The mathematics of emotions*. In Frank Detje, D.D., Schaub,H., eds.: Proceedings of the Fifth International Conference on Cognitive Modeling, Bamberg, Germany (2003) 75–79

[8] Dörner, D., Hille, K.: *Articial souls: Motivated emotional robots*. In: Proceedingsof the International Conference on Systems, Man and Cybernetics. (1995) 3828–3832

[9] Ekman, P. (1972). *Universals and cultural differences in facial expressions of emotion*. In J. Cole (Ed.), *Nebraska Symposium on Motivation 1971*, (Vol. 19, pp. 207-283). Lincoln, NE: University of Nebraska Press.

[10] Gudykunst, W.B. and Mody B. *Handbook of International and Intercultural Communication* SAGE 2002, 2nd Edition

[11] Hall, E. and Hall, M.R. *Understanding Cultural Differences*. Intercultural Press, Yarmouth, Maine, 1990.

[12] Hofstede G. *Cultures and Organisations* SAGE, 1991

[13] Khaslavsky, J., *Integrating culture into interface design*. in *CHI 98 conference summary on Human factors in computing systems*, (Los Angeles, California, 1998), ACM Press, 365-366.

[14] Mei Yii Lim, Ruth Aylett, and Christian Martyn Jones *Emergent Affective and Personality Model* IVA 2005, LNCS 3661, pp. 371–380, 2005. c_Springer-Verlag Berlin Heidelberg 2005

[15] Oliveira & Sarmento, *Emotional Advantage for Adaptability and Autonomy*, Proceeding of 2nd International join Conference on Autonomous Agents and Mul-tiagents Systems, AAMAS 03, Melbourne, Australia, ACM 2003, July 14-18, 2003

[16] A. Ortony, G. Clore, A. Collins, *The cognitive structure of emotions*, Cambridge University Press, Cambridge, England, 1998

[17] H. Ned Seelye (Editor) *Experiential Activities for Intercultural Learning* Intercultural Press (March 1996)

[18] Triandis, H.C. *Theoretical and methodological approaches to the study of collectivism and individualism* 1994

[19] Trompenaars, F. *Riding the Waves of Culture: Understanding the Cultural Diversity in Business*. Nicholas Brealey, London, 1993

[20] Würtz, E. A cross-cultural analysis of websites from high-context cultures and low-context cultures. *Journal of Computer-Mediated Communication, 11*(1), article 13. 2005.

The Reign of Catz & Dogz? The Role of Virtual Pets in a Computerised Society

A major concern for human computer interaction researchers is how to construct interfaces to future ambient and pervasive technologies which are naturalistic, unobtrusive and implicit. Perhaps in response to this there exists a good deal of well- established research which attempts to identify aspects of human-human communication (such as gesture, language and facial expression recognition) and implement these as modalities in human-computer interfaces. Such an approach is fraught with difficulty – frequently, reported work will ignore the complexities raised by context and culture, whilst recreation of interfaces which are ‘too-human’ can fall into the trap of the uncanny valley. One possible, and potentially very manageable, alternative to using aspects of human-human social cognition as inspiration and models for human-computer interaction is to consider human-animal interaction – or anthrozoology.

Sustained consumer interest in off-the-shelf robotic animals such as Aibo, i-Cybie, Robosapien and RoboPet, and the commercial success of computer- games such the Tamagochi, Catz and Dogz, and, in particular, Nintendogs, provide convincing evidence of the widespread appeal of interacting with artificial, albeit rather basic, representations of creatures. As the designers of such toys and applications are no doubt aware, an accepted consensus within anthrozoologic research is the quantifiable positive effects of human- animal relationships. Accordingly, the biologist E.O. Wilson coined the term biophilia as “the connections that human beings ... seek with the rest of life”, and argued that such cravings are determined by a biological need. However, to-date no link has been explored between such socio-biological theories and human interactions with artificial systems. The intention of The Reign of Catz & Dogz? symposium is to consider the future role that interactive artificial creatures will play in a society populated with pervasive computers, personal robots and ambient intelligence. A recent call for research in Europe advocated interfaces for robots which will be “present in everyday human environments” whilst the South Korean government is funding a strategy designed to put service-robots in every domestic household within ten to fifteen years. There are dissenting voices however which reiterate the position that computers and virtual agents can, fundamentally, never be truly social entities. Additionally, Sony recently signalled the end of their activity in personal and entertainment robotics.

The intention was to allow researchers in this area to table and discuss their views on the relevant contemporary issues prevailing and to crystallise the challenges facing us in the near future. The range of topics covered is broad. Lohse et al. examine how social robots should appear to humans and what they could/should be used for to slot into everyday life. Mival & Benyon cover similar ground examining screen based agents rather than embodied ones. Slater explores how designers and programmers might get humans to emote with the characters they create to sustain prolonged use. Casey & Rowland discuss one particular mobile phone game, whose characters were not intended as virtual pets but to which, nevertheless, players got attached. Ljungblad et al. take us back to designing robots for everyday use, drawing on analogies from the animal kingdom. Grant looks at the role of speech in devices and asks, what happens to imaginative play and art when 'things' speak? In addition there will be demonstrations of a range of virtual pets throughout the day.

Shaun Lawson and Thomas Chesney (Symposium Chairs)

Programme committee: Dave Hobbs (University of Bradford); Deborah Wells (Queen’s University, Belfast); Ehud Sharlin (University of Calgary); Richard Hetherington (Napier University); Trevor Jones (University of Lincoln); Vincente Matellan (Rey Juan Carlos University); Shaun Lawson (University of Lincoln); Thomas Chesney (Nottingham University).

Introducing the COMPANIONS project: Intelligent, persistent, personalised multimodal interfaces to the internet.

Oli Mival & David Benyon[1]

ABSTRACT

The paper introduces the COMPANIONS project, a 4 year, EU funded Framework Programme 6 project involving a consortium of 16 partners across 8 countries. It's aim is to develop a personalised conversational interface, one that knows and understands its owner, and can act as an alternative access point to resources on the Internet, all the while nuturing an emotional involvement from it's owner/user to invoke the shift from interaction to relationship. On a technical level it intends to push the state of the art in machine based natural language understanding, knowledge structures, speech recognition and text to speech. With these technical developments will come advanced interaction design elements, some of which were initiated on the SHEFC funded project, UTOPIA (Usable Technology for Older People: Inclusive & Appropriate), examining the potential for developing artificial companions for older people.

1. THE UTOPIA VIEW OF COMPANIONSHIP AND THE ELDERLY

Companionship is a concept that is familiar to all, yet defies simple explanation. Psychology considers it a central need, yet balks at a concise definition of what constitutes a companion beyond "a relationship…with mutual caring and trust" [2], p467. What is clear, is the importance of companions to emotional well being. Indeed the loss of companions is considered a primary cause of depression among older people [7]. It is therefore important to consider that the loss of human companions is a natural consequence of growing old. There is a diminishing of the supportive ties of family members, of friends and of other relationships from previous, concurrent, and following generations through death or distancing by migration or relocation. Furthermore, social roles and ties are lost through retirement and any parental function is reduced as children grow up and become independent. This substantial erosion of social networks inevitably leads to the loss of companions and is often accompanied by an experience of emotional impoverishment, not infrequently experienced by the elderly as a pervasive depression "without a reason" [3].

With consideration of this natural decline in human companionship, the potential value of developing artificial companionship become distinctly apparent.

On a simple level, older people have relationships with companions, be they pets, friends or care assistants. But what constitutes the difference between an interaction and a relationship? To form a relationship, the user needs to care about the interaction, to invest emotion in it. The artificial companions evoke the emotional investment through replicating recognizable real world behaviour. The movement of AIBO's head when he is stroked is remarkably realistic and is as endearing as the similar movement of an animal. Thus the user invokes affection and, crucially, attributes personality in much the same way as with a real pet. The importance of behaviour in the attribution of personality can be simply highlighted through the example of cats and dogs, the most common household pets. There is a strong cultural belief that cats have a higher intelligence than dogs, and that dogs are excitable and gullible compared to the cool, sophisticated elegance of cats. These personality attributes are derived from the behaviour and relationships humans have with each animal. In reality, cats have a much lower cerebral development than dogs and a relatively much smaller brain size. Yet intelligence is attributed through behaviour and human interpretation of that behaviour. Of course some of the products are more successful at evoking this personification behaviour within its user than others, though what factors are important to this process remains somewhat unclear. Realism of behaviour seems a likely contender as AIBO's movement and people's reaction to it suggests. However if technology as simple as the Tamagotchi can provoke such intense emotional responses as depression at its death, then the psychological impact must be as important as simple engineering issues. From this it may be suggested that the difference between a tool and a companion is a set of characteristics, a personality, which transforms an interaction into a relationship and evokes an emotional investment. Products which achieve this we call *personification technologies* [6]. To understand more clearly the potential factors at work in these relationships it is useful to examine the relationship between older people and their most basic companions, pets.

[1] Centre for Interaction Design
Napier University, Edinburgh
o.mival@napier.ac.uk - d.benyon@napier.ac.uk

2. THE USEFUL USELESSNESS OF PETS

The medical benefits of pet ownership are well documented [3]. Pet ownership can lead to an enhanced emotional status and provides significant support in reducing emotional trauma following bereavement. Indeed not only emotional health but also physiological health is enhanced through contact with animals, particularly in the elderly. Furthermore, studies have shown that when animals enter the lives of older patients afflicted with Alzheimer's disease or arteriosclerosis, the patients will laugh and smile more, are more socially communicative and less hostile to their care workers 16]. However, some older people live in accommodation which does not allow pets or may suffer from psychological or physiological deterioration that make pet ownership problematic and potentially unsafe for the animal. In situations such as these, the use of artificial pets may be an alternative.

It is interesting to note that pets are at the non-specific purpose end of the companionship spectrum. A cat serves no other function than to be a cat. Yet as discussed above, by simply being a cat it can affect the health and well-being of its owner. People take delight in its activities, and it is purely from its behaviour that the benefits are derived. Cats cannot read email to you, struggle as a webcam and do not react to guidance from a computer. They are autonomous objects driven by their own goals. Kaplan suggests that this autonomy, this non-functionality is an important design consideration when developing artificial pets, he suggests they should "be designed as free 'not functional' creatures" [4].

It is the intention to use these insights to drive the interaction design elements of the major new FP6 EU project COMPANIONS.

3. INTRODUCTION TO THE COMPANIONS PROJECT

The COMPANIONS project, a 4 year, EU funded Framework Programme 6 project involving a consortium of 16 partners across 8 countries. The project's vision is that of a personalised conversational, multimodal interface to the Internet, one that knows its owner, is implemented on a range of platforms, indoor and nomadic, and based on integrated high-quality research in multimodal human-computer interfaces, intelligent agents, and human language technology [8]. This project is an ECA (Embodied Conversational Agent) which differs from the ECA state of the art by having large-scale speech and language capacity; it also differs significantly from the standard "big engineering" approach to this area, by offering relatively simple architectures with substantial tested performance, based on extensive application of powerful machine learning methods.

Large groups of EU citizens will need new forms of interface to the Internet if they are to get benefit from it as it grows more complex: these include huge numbers of mainstream citizens, since the Internet simply does not serve the average non-technical person as well as it does the academics who invented it. Beyond a few simple purchases such as holidays half the EU population make no effective use of the Web at all, and a recent survey [5] shows one third of the UK population actively hostile to it. Use of the Internet to access and organise information is limited to current interaction mechanisms such as browsing. Such new forms of interface should be available across a wide range of platforms, from PCs and TV screens to mobile devices, including phones. The deluge of information on the Internet will increasingly be about individuals, and will include their own digital repositories (texts, videos, images) as well as information held about them. Most citizens will have little control over this, their own digital life, without some new form of assistive interface to the Internet. There is already an established need for individuals to organise their own life material, and to give a narrative structure to their lives, particularly when old, which now means shuffling old photographs on paper.

These needs have not been well met by current interfaces based on browsing, profiling and adaptation, and there is good evidence that a more directly personal interface will be more acceptable. A technological solution to this need, in part at least, is a persistent, personalised, companion agent, one that will "know" its owner, chat to the elderly to relieve their boredom, and become the multi-modal interface agent to the Internet for that owner, whatever their age or technical competence. The project calls these agents COMPANIONS.

COMPANIONS will learn about their owners: their habits, their needs and their life memories. This will allow them to assist with carrying out specific Internet tasks, which will be facilitated by having complex models of their owners, by which we mean whole-life-memories, or coherent autobiographies, built from texts, conversations, images and videos. Some of this will already be in digital form, but some will be information gleaned from conversations with the COMPANION, information relatives and friends will want later, after the owner's death, but might never have been able to ask, such as "where did you and your husband first meet?" The barrier to COMPANIONS so far, beyond very primitive forms, has been lack of progress in the adaptability of speech and language technology.

The objectives of the proposal are to develop autonomous, persistent, affective and personal interfaces, or COMPANIONS, embedded in the Internet environment, with intelligent response in terms of speech and language, integrated with the manipulation of visual images and their content. The aim is to have a higher level of performance in speaker independent speech recognition via robust dialogue management capacity. This needs to treat the content of communication from all modalities with regard to the mode in which they were originally expressed. Machine learning is a central part of the structure.

COMPANIONS must be believable, intuitive, and above all humane conversational interfaces, and must be proved acceptable to our sample target social groups; moreover those groups will be consulted before the platforms are integrated. COMPANIONS will be autonomous and have original aspects of persistent human personality to establish loyalty and trust between users and such agents. They will be sensitive to limited emotion in speech and to the content of images, and will be themselves capable of demonstrating emotional/affective behaviour through speech and visual appearance (e.g. an avatar on a PC screen or mobile). COMPANIONS will also, by communicating with each other enable and enhance communication between the human users, rather than only between humans and these machine artifacts.

An early implementation of a COMPANION is PhotoPal [9]. PhotoPal allows people to view their photos and talk about them with their COMPANION. Photos are automatically tagged with the relevant dialogue allowing PhotoPal to build up a rich representation of the person's activities and relationships. This allows PhotoPal to sort, style and send photos and for people to reminisce with their COMPANION. A second COMPANION is being developed in the domain of personal heath and fitness.

4. CONCLUSIONS

A significant aspect of the research in COMPANIONS is concerned with the form of embodiment of the COMPANION. It is not just that the embodiment might be in the form of a domestic animal. It could be something wholly new, but that demonstrates animal-like characteristics; particularly dependability, trust and affection. Companionship is certainly a characteristic that will arise from the interactions and relationship that results.

ACKNOWLEDGEMENTS

This work is funded by the European Commission under contract IST 034434.

REFERENCES

[1] T Garity,.L Stallones, M Marx, M. & P. Johnson, Pet Ownership and Attachment as Supportive Factors in the Health of the Elderly. *Anthrozoos*. Vol. 3, No.1 pps 35-44 (1989)

[2] H. Gleitman, *Psychology*. Oxford University Press (2000)

[3] K.L. Gory, & K. Fitzpatrick, The effects of environmental contexts on elderly depression. *Journal of Aging and Health*, 4(4):459-479 (1992)

[4] F. Kaplan (2001) Free creatures: The role of uselessness in the design of artificial pets. In *Proceedings of the CELE-Twente workshop on interacting agents 2001.*

[5] A. Light "EU Survey on Public Website Usage shows Potential" Usablility News, *9 February 2005*

[6] O. Mival S. Cringean, and D. Benyon Personification Technologies: Developing Artificial Companions for Older People, ACM Press, 1--8. (2004)

[7] C. Sluzki, *The extinction of the galaxy: Social networks in the elderly patient.* New York: Family Process (2000)

[8] Y. Wilks, Artificial Companions as a new kind of interface to the future internet. *Oxford Internet Institute, Research Report 13*, October 2006

[9] www.napier.ac.uk/companions

Studying reptile owners to avoid designing reptile-like agents

S. Ljungblad, M. Jacobsson and L.E. Holmquist
Future Applications Lab, Viktoria Institute, Hörselgången 4, 41756 Gothenburg, Sweden
(e-mail: saral@viktoria.se, majak@viktoria.se, leh@viktoria.se)

Abstract— **We are exploring strategies for designing novel robots, or more generally, personal embodied agents. The motivation is to open up the design space for robots in everyday environments, while at the same time ground new designs in existing human interests. We have investigated human interests in caring for snakes and spiders, not to design reptile-like pets, but to understand possible interests of future robot owners. From the resulting interview data we have derived a number of possible designs of agents. These are intended to meet human interests for low-level communicating autonomous artefacts. We here present one such design, which is ongoing work that investigates how to evolve visual patterns in an open-ended play between humans and embodied agents.**

WHAT different roles can robots have in everyday environments? Currently, robots intended for such environments are commonly considered as social companions [1], service or assistive robots, such as [2][3], entertainment robots [4] or therapy objects [5].

An underlying assumption of socially interactive robots is that the interaction should be similar to how human are interacting with each other [6]. However, robots with a notion of sociality, social skills and bonds with people are still more of a distant goal, than actual reality. In parallel to developing robots to become the future almost human “butlers” science fiction suggests, or anthropomorphic creatures, we want to look into alternative views of robots and interaction with them [7]. We are exploring much less sophisticated and narrower agent behaviours, taking inspiration from robots developed in basic research such as [8]. To investigate robotic behaviours that are interesting from a human perspective, we have interviewed people involved in specific human interests in autonomous and low-level communicating creatures. We have then investigated how to transform this data, as input in the design process of personal embodied agents.

INTERVIEWS

We held interviews with 10 people, six men and four women, who owned pets like snakes, lizards and spiders. We wanted to see beyond the actual artifact that their practice involved, and find underlying motivations for their interests despite the apparent limitations in interaction and communication possibilities. When analyzing the data, we found that different people had many different motivations, for example interests in building environments for the reptiles, simply watching them and to develop knowledge in how to care for them. Several such motivations and interests where then used to create four different designs of personal embodied agents [9]. Below we present ongoing work of one such resulting design case.

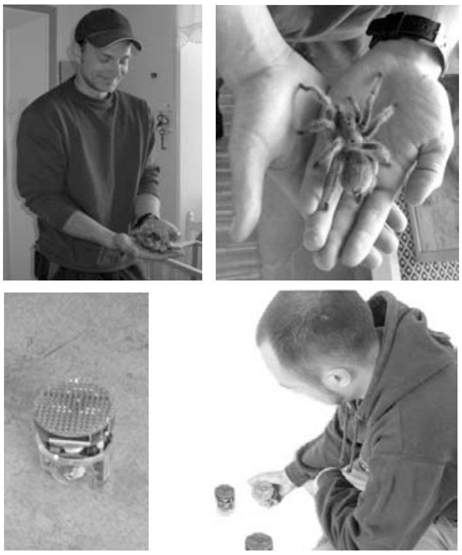


Fig. 1. (Top) Interviews were held with people owning reptiles and other exotic pets, such as spiders. (Bottom) One resulting design investigates an open-ended play with visual patterns between humans and embodied agents.

ROBOTS AS DYNAMIC HOBBY PIECES

This design case is presented as a brief scenario, intended to illustrate possible interest and interaction with agents intended for an open-ended play between humans and agent
.

Nadim has his robots as a hobby, rather than as pets. He is especially interested in robots that have visual patterns that evolve over time. He explores different ways to affect the visual outcome, and to do this he experiments with different lights, sounds and motions for his robots. Nadim also brings his robots to friends that have the same kind, so that the robots can affect each other’s patterns at different points in time. Nadim does not care if the robots evolve different personalities, nor is he interested in petting them. He simply wants to develop as interesting evolving patterns as possible, an interest he shares with his closest friends.

The robots we are developing are based on the E-puck platform [10]. We have extended the basic hardware platform with LED screens that can display dynamic and colorful patterns. We are currently investigating how visual patterns can be created and evolve between robots and humans, taking inspiration from basic research of communication between robots [8]. We aim to continue the design with possible users, evaluating the early concept and discuss challenges for future design.

REFERENCES

[1] F. Tanaka, J. R. Movellan, B. Fortenberry and K. Aisaka: "Daily HRI Evaluation at a Classroom Environment: Reports from Dance Interaction Experiments", *Proceedings of the 1st Annual Conference on Human-Robot Interaction (HRI 2006)*, Salt Lake City, U.S.A., March 2006

[2] Independence Enhancing Wheelchair, ActivMedia Robotics Availible: http://www.activrobots.com/ROBOTS/RoboChariot.html

[3] J. Forlozzi, and C. Disalvo "Service robots in the Domestic Environment: a Study of the Romba Vaccum in the Home" *Prooceedings of the 2006 ACM Conference on Human-Robot Interaction. (HRI 2006)*, Salt Lake City, U.S.A., March 2006

[4] Sony Aibo Robot Dog Available: http://www.sony.net/Products/aibo

[5] W., Taggart, S. Turkle, and C. D. Kidd, "An interactive robot in a nursing home: Preliminary remarks." *Toward Social Mechanisms of Android Science*, Stresa, Italy,. Cognitive Science Society July 2005

[6] T.Fong, I. Nourbakhsh and K. Dautenhahn "A survey of Socially Interactive Robots" *Robotics and Autonomous Systems 42* 143-166, 2003

[7] S. Ljungblad and L. E. Holmquist. "Designing Robot Applications For Everyday Environments" *Proceedings of sOc-EUSAI, Smart Objects & Ambient Intelligence Conference*, Grenoble, France October, 2005

[8] ECAgents project Available: http://ecagents.istc.cnr.it/

[9] Ljungblad S. Walter, K., Jacobsson, M. and Holmquist, L. E. Designing Personal Embodied Agents with Personas. In *Proc. Ro-Man'06*, IEEE Press (2006)

[10] E-puck Available: http://lsa.epfl.ch/~mondada/e-puck/

What can I do for you?
Appearance and Application of Robots

Manja Lohse[1] and **Frank Hegel**[1] and **Agnes Swadzba**[1]
and **Katharina Rohlfing**[1] and **Sven Wachsmuth**[1] and **Britta Wrede**[1]

Abstract. In recent years industrial robots have been successfully established because they fulfil meaningful tasks in production. In contrast the question of applications for social robots is still open. For quite some time they have only been used in research or at best as simple toys by real users in everyday life situations. However, we suggest that there are still unknown application fields that are suitable for existing robots. Therefore, our approach is to show short movies and descriptions of real robots to participants and ask whether there are any specific tasks these robots could perform in the naive users' everyday life. The systems' appearance and abilities strongly influence the user's expectations, that's why we suppose that we will find strong differences between zoomorphic robots like AIBO and iCat and other robots like BIRON (functional design) and BARTHOC (humanoid). We have conducted an online study with more than 100 participants to test this hypothesis.

1 Introduction

Developing useful applications for social robots seems to be a challenging task. At least today's scenarios are almost restricted to research and toys. Developers try to anticipate new applications but potential users are rarely included into this process. We argue that knowing consumers' opinions is important in order to design useful applications. This paper will introduce a first study with potential users. It focuses on robotic animals and compares them to a functional and a humanoid robot, respectively.

Beside market aspects, we argue that applications also offer attractive scientific aspects. First, many functional as well as socially relevant aspects are only observed when realistic applications are faced. Secondly, a thorough evaluation of the robot performance that includes social aspects of human-robot interaction gains significance from well defined application scenarios. Especially naive users need self-explaining robotic systems in order to get valid results in user studies. This can be supported by well motivated application scenarios.

In Section 2 we will introduce related work which gives a first impression of today's applications and the role of appearance in robotics. Section 3 describes the robot platforms AIBO, iCat, BIRON and BARTHOC which were shown to the participants during the trials. In Section 4 the method of the online study is explained in detail. Afterwards the results are presented in Section 5. The paper closes with a conclusion in Section 6.

2 Related Work

In this section we will introduce related work beginning in Section 2.1 with a description of applications from a research perspective. Section 2.2 will go deeper into everday applications and Section 2.3 presents the relationship between the appearance of a robot and its applications.

2.1 Applications from a Research Perspective

Social robots have been a focus of research for several years. Most of them have been developed for a dedicated scenario that is defined in order to demonstrate skills and features of the robot rather than in terms of applications. For example, the MIT robot Leonardo is learning the names of buttons from human demonstration [4], the robots Ripley and its successor Trisk learn to integrate different modalities [24]. The robots SIG [21] and Robita [27] focus on multiple speaker tracking and conversation participation. Others demonstrate fetch and carry tasks (e.g. Hermes [2]), object manipulations on a table [5, 20], or human guided spatial exploration [26]. In the same line, the AAAI conference 2002 defined robotic challenges that included social abilities. The robot had to start at the entrance to the conference centre, needed to find the registration desk, register for the conference, perform volunteer duties as required, then report at a prescribed time in a conference hall to give a talk [25].

As impressive as the demonstrated robotic skills are, these scenarios are still far from market relevant applications and miss certain evaluation aspects.

2.2 Everyday Applications

Robots are not part of everyday life yet. We have to think about what their place in our public and private life could be. Therefore, the first important step is to ask what would actually make social robots valuable as everyday objects. One possibility is to find out which qualities objects have in our everyday life as short-term or long-term applications [17]. Furthermore, the value of the objects has to meet the needs of the users.

[1] Bielefeld University, Applied Computer Science Group, Germany, email: {mlohse, fhegel, aswadzba, rohlfing, swachsmu, bwrede}@techfak.uni-bielefeld.de

Fong et al. [10] propose several application fields: social robots as test subjects for research on communication and human development theory, as short-term and long-term service assistants in public and private life, as toys and entertainment devices, for therapy, for research on anthropomorphism, and last but not least in the field of education.

In a workshop on designing robot applications for everyday environments, organizers and participants brainstormed on new application scenarios [18]. After refining their ideas they selected three application concepts: self-organizing robot plants, robots as travel companions, and amusement park guide robots. This was a first official workshop which tried to find new applications in social robotics. As can be seen, there are few approaches to find applications for social robots, but neither one considers the needs and opinions of potential users.

2.3 Appearance and Expression of Robots

The appearance of a robot influences what the interacting users expect and how they will judge a certain application. The appearance of a robot becomes especially important when assessing its performance and appropriateness for an application. Humanoid and animal robots convey anthropomorphic cues that get the user to make several attributions concerning the robot's abilities. Thus, because of the appearance, a user has expectations whether an application for a specific robot is appropriate or not [13]. The more human attributes a robot has, the more it will be perceived as a human [9] and the more the appearance is expressing human traits and values [8]. If a robot looks like an animal it will express the traits this specific animal has.

Most nonverbal cues are mediated through the face. A robot's physiognomy changes the perception of humanlikeness, knowledge, and sociability. People avoid negative robots and feel more common ground interacting with a positive expressive robot [12]. Furthermore, an expressive face indicating attention [6] and imitating the face of a user [16] makes a robot more compelling to interact with. Also, faces with large eyes and small chins in proportion to the rest of the face are so called baby faces. Men with baby faces are perceived more honest, kind, naive, and warm. The same happens with robots if they have a baby faced design [22].

We want to resume, that users rate applications of a robot more or less appropriate because of its appearance and expression. Therefore, the design of a robot is an indicator for its application.

3 Technical Description

This section gives a short technical overview of the robot platforms presented to the users during the survey. Each of the four robots shown is used for research.

3.1 iCat

The Philips iCat research platform shown in Figure 1 is a plug & play desktop user-interface robot which is capable of mechanically rendering facial expressions. It is developed by Philips Research (Eindhoven, the Netherlands).

The robot platform contains 13 RC servos controlling the eyebrows, the eyelids, the eyes, and the lips and two DC motors for moving the head and the body which enable the iCat to create facial expressions. Four multi-colour RGB-LEDs and capacitive touch sensors are located in the feet and the ears. The iCat can communicate its mode of operation (e.g. sleeping, awake, busy, or listening) with these LEDs. A USB webcam with a resolution of 640 × 480 pixels and 60 fps is placed in the nose. Therefore, the iCat can be used for different computer vision tasks, such as object and face recognition. Stereo microphones, a loudspeaker and a soundcard can be found in the feet of iCat and are used for playing sounds and speech. Thus, it is possible to record speech and to use it for speech recognition and understanding tasks. Finally, the robot is equipped with an IR proximity sensor.

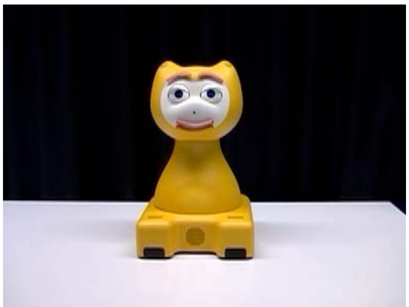


Figure 1. The Philips iCat research platform.

The iCat is a user-interface robot without an on-board processor. It can be controlled by a PC via USB. Researchers can use the Open Platform for Personal Robots (OPPR) software which provides a development environment for creating applications for user-interface robots. More details can be looked up in [23]. This website provides the infrastructure for supporting an online research community.

3.2 AIBO

The Sony AIBO Robot ERS-7 is presented in Figure 2. Its design is quite dog-like. AIBO has sensors on the head, the back, the chin, and the paws which allow the robot to examine itself and its environment.

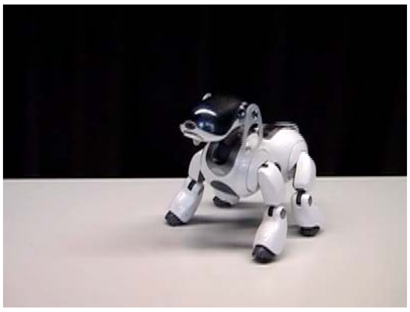


Figure 2. The Sony AIBO ERS–7.

Moreover, it can perceive sound using a pair of stereomicrophones. Therefore, it can react to voice. Because of the colour camera and distance sensors AIBO can recognise colours, faces, and obstacles. It is able to communicate its mood via sounds and a face display and via words with humans.

AIBO is using its four feet to act in its environment. With the acceleration sensors on-board it is able to balance its body. AIBO has – considering its feet, head, ears, and tail – altogether 20 joints (degrees of freedom) which give the robot the capability to perform dog-like moves. Consequently, one application for AIBO – Sony had in mind – is the so-called watchdog scenario. The robot can guard its home by taking photos of unusual things and informing its owner via email. More information about the AIBO platform can be found in [7].

3.3 BARTHOC

Figure 3 gives an impression of the humanoid robot BARTHOC (Bielefeld Antropomorphic RoboT for Human-Oriented Communication) [15]. This robot is designed by Bielefeld University in cooperation with Mabotic for research in human-like communication. The main focus of the design is to realise the expression and behaviour of the robot to be as human-like as possible. It can mimic human behaviour like speech, facial expressions, and gestures with his soft- and hardware.

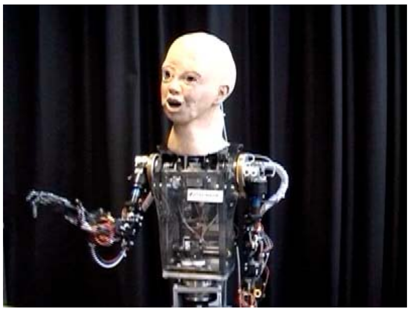


Figure 3. The humanoid robot platform BARTHOC (Bielefeld Antropomorphic RoboT for Human-Oriented Communication).

The anatomy of BARTHOC consists of a mechatronic head and two arms including hands. These components are mounted on a steel-frame backbone.

Each arm has three joints similar to the human ones. The given degrees of freedom (DOF's) allow BARTHOC to perform human-like gestures. The joints of hip, shoulders, upper and lower arms are driven by planetary geared DC motors with position feedback via precision potentionmeters. The hand is constructed as an external actuator type. Each finger is built with three spring pre-stressed joints driven by a flexible, high strain resistant nylon cable.

A complete mechatronic head has been developed with a more human-like appearance and human-like features. A camera is integrated in each eyeball for stereo vision and microphones are currently placed on the shoulders. Additionally, a removable latex mask is constructed to give the possibility to exchange characters. Actuators next to the upper lip and above the eyes simulate movements of lips and eyebrows. The movements drive the mask to express basic human facial expressions.

3.4 BIRON

The Bielefeld University developed a mobile robot platform called BIRON (BIelefeld RObot companioN) (see Figure 4). BIRON is based on an ActiveMedia Pioneer PeopleBot $^{\text{TM}}$.

A Sony EVI D–31 pan-tilt colour camera is mounted on top of the robot at a height of 141cm for acquiring images of the upper body part of humans interacting with the robot. A pair of AKG far-field microphones is located right below the touch screen display at a height of approximately 106cm. Therefore, BIRON has the capability to localise speakers and process speech. Finally, a SICK laser range finder mounted at a height of 30 cm facing front measures distances within a scene. Since BIRON has wheels, it is able to move and to follow a person.

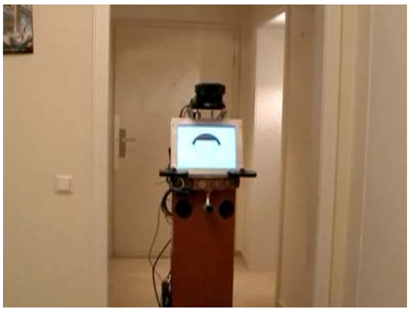


Figure 4. BIRON the BIelefeld RObot companioN.

The robot is equipped with two on-board computers. The first one is controlling the motors, on-board sensors, and performing the sound processing. The second one is used for image processing, especially skin-colour segmentation, face recognition, and face identification.

As BIRON can track humans and pay attention selectively to humans looking at it, a first application for the robot is the so-called home-tour scenario. In this scenario a human introduces all objects and places in a private home to the robot which may become relevant for later interaction. Additional information about the architecture of BIRON is given in [14].

4 Method

In our study we were mainly interested in the following questions:

- Which applications are proposed?
- Are there any differences between proposed applications for the robots according to their appearance?
- Which applications do people propose especially for zoomorphic robots? and
- What is people's attitude towards using zoomorphic robots?

We decided to conduct an internet survey for several reasons. First of all, interaction studies are very time-consuming. In contrast, online studies are very fast and provide a manifold sample. Therefore, they represent an alternative to traditional methods, especially when conducting highly exploratory studies [1]. Moreover, the internet survey – in which only short videos of each robot were shown to the participants – supports the general idea of the study. Subjects should only have a first impression of the robots. Thus, their assumptions are mainly based on the appearance of the robots and the information given about their functions. Their ideas were not restricted by technical problems which might have occurred in real settings.

We published the questionnaire on the website of an online laboratory and invited people via private and professional mailing lists to participate. Therefore, one part of participants was random users who visited the website of the laboratory and took part in the survey. Subjects who received an email are part of a deliberate sample because the mailing lists were chosen by the researchers. The sample is not representative because it only includes subjects that are using the internet frequently or have interest in psychology, surveys in general or robotics. Therefore, the results can not be generalised [3], which in any case is not the claim of the study.

The survey was conducted during one week in January 2007 with 113 participants (39,3% female, 60,7% male). Their age ranged between 9 and 65, with an average of 30,2 years. The majority is educated above average (highest degree: 34,5% high school graduates; 55,8% university graduates; 8,0% doctoral degree, 1,7% other). Nevertheless, we are of the opinion that the diversity of our sample is higher than in student samples which are often used in robotics research (Table 1). Subjects are naive in the sense that they are not working in the field of robotics. Most of them have German nationality (Table 2) which is due to the fact, that the questionnaire was published in German. Related to this, one more advantage of online-surveys is, that we will have the possibility to amplify the study by publishing the questionnaire on the web in different languages.

	n	percent
student	2	1,8%
university student	45	39,8%
employed	58	51,3%
unemployed	1	0,9%
others	7	6,2%
total	113	100%

Table 1. Participants' professional status.

	n	percent
Germany	102	91,1%
Switzerland	1	0,9%
Austria	3	2,7%
others	6	5,4%
total	112	100%

Table 2. Participants' nationality.

The study reported here is highly exploratory. Thus, the questionnaire contains several open questions. It is divided into introduction, sociodemographic questions and application questions for each of the four robots (e.g. Would you use this robot?; Which meaningful applications can you imagine for this robot?). For all of them a short video was shown, which illustrated their appearance. Figure **??** shows a screenshot of the questionnaire.

Especially the open questions concerning the applications are of interest for this paper. Participants were free to write down as many items as they could think of. Altogether, a fairly high number of 495 applications were mentioned (AIBO 148, BARTHOC 90, BIRON 120, iCat 137). Therefore, a content analysis in order to group the data was essential [11, 19]. The entries were analysed by three researchers. First, they were structured into more restrictive categories for each robot and then grouped into wider classes. The number of entries in each class was rechecked with the data. If people mentioned the same application twice for the same robot, only one entry was coded. Some people wrote remarks like ”just as first robot”. These answers were not coded because the order in which the robots were shown to the participants changed randomly. Subjects needed an average of 11min 32sec ($x_{\text{med}} = $ 7min 40sec) to complete the questionnaire.

5 Results

In this section we want to introduce the results of the study and outline some interesting discoveries concerning the distinct perception of AIBO and iCat. As described above we defined categories of applications which are listed in Table 3. Within the table, groups, which assimilate different categories, are specified. An example is given for each category.

Obviously, the categories have different levels of abstraction which is due to the varying complexity of the applications. Moreover, it is important to mention that these categories can only provide an insight into the applications for the four robots tested. Nevertheless, they give a first idea how naive people view robotics. We are aware of the possibility to further reduce and structure the categories introducing broader dimensions. However, this is not the goal of this paper because we want to give an overview of the diversity of applications mentioned by the participants. We decided not to include the two following categories in Table 3. Seven participants stated that BARTHOC could be used for a horror film or haunted house, because they thought that his appearance was very frightening. We think these comments are rather ironical than useful applications. Nevertheless, they are a hint that we have to keep working on the appearance of BARTHOC. Furthermore, three industrial applications were mentioned which are not subject of social robotics.

Moreover, Table 3 sums up the applications subjects mentioned for all the robots which were shown to them. The question which applications users ascribe to AIBO and iCat has still to be answered. We noticed that many tasks proposed for AIBO are typical for a dog (guard dog, guide dog, fetch and carry tasks). Even more people stated that they could imagine the robot as toy and pet, which is also proposed by the developers (Section 3.2). iCat was also seen as a toy but surprisingly only few people thought of the robot as a pet. A reason for this phenomenon might be seen in the appearance of the robot which is only a torso and not a complete cat. It also doesn't have the functionality of a real cat. This might as well explain why no cat-like tasks such as ”chasing mice” are attributed to iCat. Altogether, it is less similar to a cat than AIBO to a dog. Besides being used as a toy people uttered that iCat might be employed as a teacher (especially for languages) or for surveillance. One application which was brought up by six subjects exclusively for iCat was using the robot as an interface to control other technical devices. Since this is a scenario described by the developers (Section 3.1) one could think that the participants knew iCat. Surprisingly, they didn't say so in the questionnaire.

There's a huge gap between applications mentioned for AIBO or iCat and tasks people ascribe to the other robots. The most commonly mentioned applications for BIRON were Surveillance, Information

Category	**Specification**	**Example**	**(a)**	**(b)**	**(c)**	**(d)**	**total**
Security	Surveillance Military tasks Dangerous tasks Exploration Security	“the robot should watch my house”	4	14	19	27	64
Research	Research Robocup	“research in heuristics of movement”	4	1	8	2	15
Healthcare	Therapy Help for the sick and old	“help for people with disabilities”	5	14	13	5	37
Personal assistant, Interface	Personal assistant Butler Organizer Interface Household / Cleaning	“for programming VCR, TV, . . . ” “electronic butler”	3	13	4	27	47
Business	Sales Reception Representation	“the new generation of ticket machine” “the robot could welcome people”	28	10	1	10	49
Public assistant	Guide (e.g. museum) Informationterminal Translator	“Infoterminal where it is needed”	25	11	5	36	77
Toy		“to play soccer”	3	38	46	1	88
Pet		“replacement pet”	0	5	18	0	23
Entertainment		“to entertain and to kill time”	2	8	9	0	19
Teacher (e.g. language)		“conduct tutorials”, “language trainer”	4	16	3	3	26
Transport (fetch & carry)		“maybe it could fetch the newspaper”	2	0	11	5	18
Companionship for lonely people		“to keep company”	1	1	8	2	12
Caregiver for old/sick people or children		“to look after old people and children”	1	6	3	0	10

Table 3. Categories of applications, specification and examples; applications mentioned for the robots (a) BARTHOC, (b) iCat, (c) AIBO, (d) BIRON.

Terminal and Guide. BARTHOC was seen as Information Terminal, Sales Robot (e.g. ticket machine) and Receptionist. All these applications are rather ”serious” in nature. These two robots were not associated with toys at all.

In the following we want to point out some more interesting results connected to the applications mentioned by the subjects. One important insight is that the majority of participants refuse to use a robot no matter what it looks like (Table 4). There are only slight differences for the robots researched in this paper. One tendency to be found is that the highest number of subjects would use BIRON, the rather functional robot.

	n	yes	maybe	no
BARTHOC	108	11 (10,2%)	27 (25,0%)	70 (64,8%)
iCat	110	12 (10,9%)	28 (25,5%)	70 (63,6%)
AIBO	107	16 (15,0%)	24 (22,4%)	67 (62,6%)
BIRON	104	23 (22,1%)	26 (25,0%)	55 (52,9%)
Average		14,5%	24,5%	61,1%

Table 4. Willingness to use robots.

We also asked (a) which robot participants would like to own, (b) which they think is most enjoyable to interact with, and (c) which robot is most likeable (see Table 5).

	n	BARTHOC	iCat	AIBO	BIRON
a)	97	6 (6,2%)	18 (18,6%)	41 (42,3%)	32 (33,0%)
b)	100	9 (9,0%)	17 (17,0%)	57 (57,0%)	17 (17,0%)
c)	96	3 (3,1%)	38 (39,6%)	46 (47,4%)	9 (9,4%)

Table 5. Rating of the questions (a) Which robot would you like to own? (b) Which robot is most enjoyable to interact with? (c) Which robot is most likeable?

At first sight there seems to be a contradiction between subjects stating they wanted to use BIRON most and the preference for owning AIBO. Looking at the applications mentioned for BIRON it becomes obvious that participants ascribe tasks in the public like ”guide” to it which explains the difference. The majority of participants (57,0%) thinks that interaction with AIBO would be most enjoyable and rate AIBO (47,4%) and iCat (39,6%) as most likeable (Table 5). This might be due to appearance, social cues or familiarity. This question should be addressed by further research. Nevertheless, no matter why people prefer the robotic animals when asked for likeability, the results indicate, that zoomorphic design might be recommendable for future systems.

6 Conclusion

In this paper we tried to show in a first exploratory study that users should be included in the process of finding new applications for social robots. We propose that this is essential especially for the increasing number of off-the-shelf robotic platforms. A carefully designed application needs to consider a frequent tendency of users to reject all kinds of social robots. Furthermore, we found that the appearance of a robot strongly influences the user’s perception. Thus, it should - as well as the functionality - be in the focus of design decisions. In contrast to the humanoid robot BARTHOC and the functionally designed robot BIRON, AIBO and iCat are above all seen as toys. Future development will show whether their appearance is also suited for other applications like robotic interfaces, business, or security.

ACKNOWLEDGEMENTS

This work is partially funded by the German Research Foundation (DFG) within the SFB673 “Alignment in Communication” – Projects C2 and A4 – and the EU integrated Project COGNIRON.

REFERENCES

[1] W. Bandilla, *Web Surveys – An Appropriate Mode of Data Collection for the Social Sciences?*, Hogrefe & Huber Publishers, online social sciences edn., 2002.

[2] R. Bischoff and V. Graefe, ‘Demonstrating the humanoid robot hermes at an exhibition: A longterm dependability test’, in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems: Workshop on Robots at Exhibitions*, Lausanne, Switzerland, (2002).

[3] J. Bortz and N. Döring, *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*, Springer-Verlag, Berlin, Heidelberg, New York, 3 edn., 2003.

[4] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Chilongo, ‘Tutelage and collaboration for humaoid robots’, *International Journal of Humanoid Robots*, **1**(2), 315–348, (2004).

[5] M. Brenner, N. Hawes, J. Kelleher, and J. Wyatt, ‘Mediating between qualitative and quantitative representations for task-orientated human-robot interaction’, in *Proc. of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India, (2007).

[6] A. Bruce, I. Nourbakhsh, and R. Simmons, ‘The role of expressiveness and attention in human-robot interaction’, in *Proc. AAAI Fall Symp. Emotional and Intel. II: The Tangled Knot of Soc. Cognition*, (2001).

[7] Sony Corporation, 2007. http://www.sony.net/Products/aibo/index.html.

[8] C. DiSalvo and F. Gemperle, ‘From seduction to fulfilment: the use of anthropomorphic form in design’, in *DPPI ’03: Proceedings of the 2003 international conference on designing pleasurable products and interfaces*, pp. 67–72, New York, NY, USA, (2003). ACM Press.

[9] C. F. DiSalvo, F. Gemperle, J. Forlizzi, and S. Kiesler, ‘All robots are not created equal: the design and perception of humanoid robot heads’, in *DIS ’02: Proceedings of the conference on Designing interactive systems*, pp. 321–326, New York, NY, USA, (2002). ACM Press.

[10] T. W. Fong, I. Nourbakhsh, and K. Dautenhahn, ‘A Survey of Socially Interactive Robots: Concepts, Design, and Applications’, *Robotics and Autonomous Systems*, **42**(3 – 4), 142 – 166, (2002).

[11] W. Früh, *Inhaltsanalyse, Theorie und Praxis*, UVK Medien, Konstanz, 4 edn., 1998.

[12] R. Gockley, J. Forlizzi, and R. Simmons, ‘Interactions with a moody robot’, in *HRI ’06: Proceeding of the 1st ACM SIGCHI/SIGART conference on human-robot interaction*, pp. 186–193, New York, NY, USA, (2006). ACM Press.

[13] J. Goetz, S. Kiesler, and A. Powers, ‘Matching robot appearance and behavior to tasks to improve human-robot cooperation’, in *Proceedings of the 12th IEEE Workshop on Robot and Human Interactive Communication (ROMAN 2003)*, pp. 55–60, San Francisco, CA, (2003).

[14] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer, ‘BIRON – The Bielefeld Robot Companion’, in *Proc. Int. Workshop on Advances in Service Robotics*, eds., E. Prassler, G. Lawitzky, P. Fiorini, and M. Hägele, pp. 27–32, Stuttgart, Germany, (2004). Fraunhofer IRB Verlag.

[15] M. Hackel, S. Schwope, J. Fritsch, B. Wrede, and G. Sagerer, ‘A humanoid robot platform suitable for studying embodied interaction’, in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 56–61, Edmonton, Alberta, Canada, (2005). IEEE.

[16] F. Hegel, T. Spexard, T. Vogt, G. Horstmann, and B. Wrede, ‘Playing a different imitation game: Interaction with an Empathic Android Robot’, in *Proc. 2006 IEEE-RAS International Conference on Humanoid Robots (Humanoids06)*, pp. 56–61. IEEE, (2006).

[17] F. Kaplan, ‘Everyday robotics: robots as everyday objects’, in *Proceedings of Soc-Eusai 2005*, pp. 59 – 64, Grenoble, France, (2005).

[18] S. Ljungblad and L. E. Holmquist, ‘Designing robot applications for everyday environments’, in *sOc-EUSAI ’05: Proceedings of the 2005 joint conference on smart objects and ambient intelligence*, pp. 65–68, New York, NY, USA, (2005). ACM Press.

[19] P. Mayring, *Einführung in die Qualitative Sozialforschung*, Beltz Verlag, Weinheim, Basel, 5 edn., 2002.

[20] P. McGuire, J. Fritsch, J. J. Steil, F. Röthling, G. A. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter, ‘Multi-modal human-machine communication for instructing robot grasping tasks’, in *Proc. IROS 2002*, pp. 1082–1089. IEEE, (2002).

[21] H. G. Okuno, K. Nakadai, and H. Kitano, ‘Social interaction of humanoid robot based on audio-visual tracking’, in *Proc. Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE)*, pp. 725–734, (2002).

[22] A. Powers and S. Kiesler, ‘The advisor robot: tracing people’s mental model from a robot’s physical attributes’, in *HRI ’06: Proceeding of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pp. 218–225, New York, NY, USA, (2006). ACM Press.

[23] Philips Research. iCat research community, November 2006. http://www.hitech-projects.com/icat/.

[24] D. Roy, ‘Grounding words in perception and action: computational insights’, *Trends in Cognitive Science*, **9**(8), 389–396, (2005).

[25] R. Simmons, D. Goldberg, A. Goode, M. Montemerlo, N. Roy, B. Sellner, C. Urmson, A. Schultz, M. Abramson, W. Adams, A. Atrash, M. Bugajska, M. Coblenz, M. MacMahon, D. Perzanowski, I. Horswill, R. Zubek, D. Kortenkamp, B. Wolfe, T. Milam, and B. Maxwell, ‘Grace: An autonomous robot for the AAAI robot challenge’, *AAAI Magazine*, **42**(2), 51–72, (2003).

[26] T. Spexard, S. Li, B. Wrede, J. Fritsch, G. Sagerer, O. Booij, Z. Zivkovic, B. Terwijn, and B. Kröse, ‘BIRON, where are you? - Enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization’, in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 934–940, Beijing, P.R. China, (2006). IEEE.

[27] T. Tojo, Y. Matsusaka, and T. Ishii, ‘A conversational robot utilizing facial and body expressions’, in *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics*, pp. 858–863, Nashville, TN, (2000).

A Foundation of Emotion Research for Games & Simulations

Stuart Slater and **Kamal Bechkoum** and **Kevan Buckley**[1]

'Emotions are typically triggered by world events; they arise from experiences that thwart or stimulate our desires, and they establish coherent action plans for the organism that are supported by adaptive physiological changes. '[1]

ABSTRACT

This paper attempts to clarify much of the terminology involved with emotions, in order to avoid the ambiguous vernacular usage of terms. A classification of emotion terminology is provided that is aimed at programmers and developers working in games and simulations. Supporting research and theories are discussed to ensure that developers are guided by the substantial body of work done in the field of psychology such as the theory underlying basic emotions [2] and the work on facial emotions [3], that can be utilised when developing more human-like (anthromorphic) agents. A glossary of terms is provided at the end.

1 Introduction

Over the last 20 years, developers of simulations and computer games have made considerable progress in improving graphics, sound and artificial intelligence (AI) in successive generations of games. Games such as Prey[2] are so realistic in graphics and sound that they have been rated certificate 18 by the BBFC (British Board of Film Classification) on their content, to prevent inappropriate demographics from engaging with the content. In F.E.A.R.[3] the agents have been equipped with sophisticated artificial intelligence systems based on goal-orientated action planning (GOAP) [4], which results in agents with a semblance of emergent behaviour. These agents are capable of inter-agent communication and of novel behaviour such as hiding and 'flushing out players with grenades' [4].

Though the allocation of developer resources in areas such as sound and graphics has greatly increased, less development resources are generally attributed to agent AI. This is potentially because in many games such as first person shooters (FPS) the agents only live for a short time and with tight budgets and approaching deadlines, additional development in AI is not justifiable due to the minimal return in game-play. These agents do not need to exhibit behaviour beyond running into the players' viewing area, but they do need to look as realistic as possible and therefore the visual appearance is a high priority. The relevance of this rationale begins to diminish in other game genres such as simulation-style games, for example the Sims 2[4] and World of Warcraft[5] both involve the players' avatar interacting with non-player characters (NPC's) over extended periods. In these games improved AI is advantageous because it increases believability; therefore greater resources are invested accordingly. This investment supports evolving agent behaviours that at present are only seemingly devoid of one aspect, agent emotions [5].

2 Where are Emotions in Games?

Games such as World of Warcraft demonstrate that some developers have begun to tackle the issue of integrating limited emotions into agents. This has mainly been achieved with facial animations that involve the display of rage and fear, supported with scripted behaviours such as running away when the agent has taken considerable damage. This incorporation of limited agent emotions does add to the believability of agents, but does not alter the agent's behaviour in anything other than a scripted and repetitive way.

There are several possible reasons why developers are reluctant to incorporate more (if any) agent emotions in their games including:

1. *Difficulties constraining agents to the game architecture*. If agents exhibit unexpected behaviour (as a consequence of emotions) then potentially they could break games or adversely affect game-play.
2. *Developer concerns with increased processing and memory storage requirements*. Additional resources allocated to emotions will need to come from another area of the game such as graphics, which is always going to be difficult to justify.
3. *Developer knowledge*. Game developers have spent considerably more time developing graphics architectures than AI and therefore knowledge in areas such as emotion modelling is lacking.
4. *Differing of scientific ideas*. In 2006, 134 years since Darwin first published 'The Expression of Emotion in Man and Animals' [6] there is still much controversy in the scientific community surrounding emotions. Therefore which theories and research can be modelled and simulated?

For the purpose of this paper *developer knowledge* and the *differing of scientific ideas* will form the main focus of the discussion in order to provide a common foundation for subsequent work.

3 Why Did Humans Evolve Emotions?

A widely supported theory of 'purpose of emotions' is that emotions evolved to both allow primitive man to automatically engage in survival orientated behaviour, when confronted with dangerous situations [7][8] and to motivate behaviour towards supporting homeostasis by providing hedonic qualia. As a consequence of this motivation, emotions help direct attention and cognition towards the emotional stimuli [9]. As social skills were developed, emotions changed to allow humans to register and react to events without automatically committing them to a course of action [8], which may not be in reflection the best options in social situations. This change

[1] University of Wolverhampton, UK. e-mail: s.i.slater@wlv.ac.uk

[2] http://www.prey.com/ (Take-Two Interactive Software 2006)

[3] http://whatisfear.com/uk/ (Vivendi Universal Games 2005)

[4] http://thesims2.ea.com/ (Electronic Arts 2006)

[5] http://www.worldofwarcraft.com (Blizzard Software 2004)

was a clear example of an organism adapting to its changing environment i.e. evolving.

4 Emotions Terminology

Because emotions are an integral part of the human identity it is possible for individuals to identify through introspection their own perceptions of emotions, and to observe and recognise what they define as emotions in others. People can, and do actively discuss many terms related to emotions such as feelings, fear and moods. The difficulty is that these observations are very subjective and ambiguous by nature, therefore it is extremely difficult to achieve consensus on general emotion related terms. A further difficulty is that according to extensive research, emotion awareness is only achieved once the unconscious emotion-activation causes a physiological change such as an increased heart rate [10]. This phenomenon is classically called feelings and often leads to the labeling of the experience as a particular emotion. Unfortunately we cannot experience other people's feelings and therefore emotion classification based on reported feelings remains subjective.

Terminology for describing specific emotions (even the word emotion) is made more difficult because people tend to apply different definitions to emotion-related events and experiences [11] throughout their lives. These definitions are always based on retrospective experiences and by its very nature memory recall can be misleading (as will be shown later). The emotion terminology problem has been highlighted in previous research such as studies conducted on common descriptions of emotions which show variations in terminology understanding when groups are selected from sociology and psychology students [12].

Much of the debate that has encompassed emotion terminology has led some researchers to publish 'emotional dictionaries' [13] [14] that involve common words and their emotional significance.

5 Are Emotions Innate or Learned?

Do human beings develop emotions throughout their life or are they innate? [6] The question of whether emotions are with us from birth depends on our understanding of the term emotions and which aspect of emotion is being discussed. There is extensive agreement in the field that the physiological appearance of emotions such as fear can be observed at an early age [6] [1], and across different cultures [3], supporting the theory that people do not need to learn to appear to be afraid or sad. Other research clearly shows that through methods such as Pavlovian[7] (classical) conditioning that people can add triggers to their 'emotional alert database' [3] throughout their lives. In some cases this can have life changing effects such as in traumatic events when individuals can become conditioned to unconsciously react to subtle stimulus such as noises and odours.

There is general agreement across a range of researchers that:

1. People do relate new stimulus triggers to emotions, thus some *learning* is associated with emotions but not the development of new emotions.
2. Some emotions have visible indicators which will be termed *physiological effects* [2].
3. Emotion activation occurs as a consequence of a stimulus which is termed *appraisal.*
4. That an emotional response can be automatic which relies on unconscious processing; this is often labelled *autonomic arousal [15].*
5. *Feelings* are the realisation of bodily changes brought about by the activation of an emotion.
6. That there is an *impulse to action [15]* response to emotional stimuli e.g. the 'fight of flight response' evident from the emotion of fear.

If the previous points are drawn together then:

Emotions are an unconscious reaction to an appraisal of a stimuli deemed to require action, that often cause physiological changes and a motivation to act. The conscious realisation of the changes is commonly called feelings and at this point the person might become aware that they are experiencing an emotion.

6 How Many Emotions Are There?

There is much scientific debate as to how many emotions there actually are, which in some cases is two up to infinity [2]. This wide range of emotions can present a serious problem for the developer made a little easier because there are two distinct approaches related to the number of emotions - categorical and dimensional.

The first approach specifies a number of emotions that can be *categorised* by terms such as basic, secondary and universal. Supporters of this approach generally attribute a finite number of emotions to each category [14] [2], though there is a variation in the number of categories from three [7] up to twenty two [2].

The second approach is that emotions are *dimensional* i.e. that there are two or three [16] dimensions that cover the range of all emotions. Common labels for these dimensions include valence, arousal and dominance.

- *Valence* range would encompass happy to sad.
- *Arousal* range would encompass calm to excitement.
- *Dominance* range would encompass control to out-of-control

Support for the dimensional approach includes research using emotional stimuli presented via TV, radio and computers mapped to valence and arousal axes which allowed researchers to consistently predict emotional responses using the approach [16].

7 Categories of Emotions

7.1 Category One

Based on a categorical approach to emotions there is much agreement to the existence of six emotions in one category. The name of this category has, and is the subject of much debate, and includes basic emotions [2] and universal emotions [7]. It is a commonly supported view that category one emotions begin manifesting themselves soon after birth and are visible facially [17] [7] such as the curled upper lip and nose wrinkling common with disgust [7], or the raising of the eyebrows common with surprise thought to allow more sensory information to enter the visual field [17].

These category one emotions are often labelled anger, disgust, fear, happiness, sadness and surprise.

7.2 Category Two

Subsequent categories of emotions are labelled in many ways and it is a common belief that one of these categories contains emotions that allow the developing person to engage and integrate in a social context such as *contempt*, an emotion defined as 'feeling morally superior to another' [18]. This category is often referred to as social

[6] Not to be mistaken for the 'is behaviour nature or nurture' debate originally defined by Francis Galton (1822-1911)

[7] Originally defined by Ivan Pavlov (1849-1936)

[9] or secondary emotions [7] and includes contempt, embarrassment, guilt, jealousy, pride, remorse and shame.

Some theories relating to social emotions base this category on a blend of category one emotions such as *jealousy* which is defined as a blend of *anger*, *sadness*, *fear* and *disgust* [18] i.e.

- *Anger* at a person i.e. x has something I do not have.
- *Sadness* at not having something that x has.
- *Fear* in anticipation that x might get more and I might never have.
- *Disgust* at self for feeling jealous of x.

This category of emotions contains emotions that enhance survival by equipping the subject with the ability to both recognise category one emotions in others and to blend category one emotions together to deal with socially fluid environments through the use of both cognitive and auto appraisal mechanisms.

7.3 Further Categories

There is widespread debate concerning other categories including background emotions such as well-being, calm, tension, relaxation, fatigue and energy [7] but due to the widespread variance these will not be discussed further.

8 Emotion Triggers

Emotions are activated by something that the person perceives; this is often referred to as an emotional trigger. The identification and possible grouping of these triggers is a subject of much debate. Some theories include hierarchical approaches to triggers such as the OCC (Ortony Clore Collins) model that features 22 specific triggers grouped into three broad categories [19] others use a more general approach such as the nine trigger model proposed by Paul Ekman [17].

8.1 OCC Model [19]

The OCC proposes that the three broad categories that all emotion triggers can be grouped by are:

- *Consequences of events* – i.e. I am pleased/displeased that something happened to me or I am pleased/displeased that something happened to someone else.
- *Actions of Agents* – The triggers are thought to relate to standards and take the form of approving or disapproving of something that someone has done, maybe to me, maybe to someone else.
- *Aspects of Objects* – I like or dislike something

The triggers and corresponding emotions are shown in Figure 1 and allow a progressively granular view of triggers based on the hierarchical three main triggers.

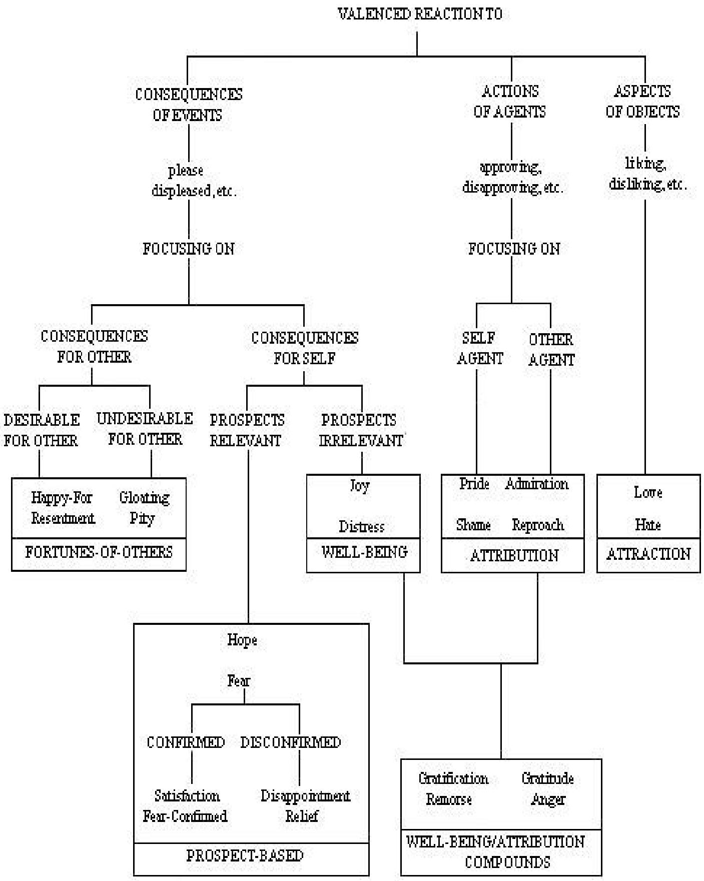


Figure 1. - The OCC Model of Emotion.

8.2 Nine-Trigger Ekman Model [17]

This model features nine broad categories for all emotional triggers:

1) *Automatic Appraisal* - i.e. interference with goals that causes an automatic emotion such as anger.
2) *Reflective Appraisals* – regretting decisions or entering an emotional state when reflecting on something that happened recently.
3) *Memory of a past emotional appearance*- remembering emotional events triggers an emotional state.
4) *Imagination* – thinking about something emotion related can incite an emotional state.
5) *Talking about past emotional experiences* – that the discussion of emotions brings forth emotions that can result in the manifestation of subsequent emotions.
6) *Empathy* – the triggering of an emotion due to someone else being in an emotional state i.e. pity or anger for someone who is sad.
7) *Others instructing us* to be emotional about something.
8) *Violation of social norms* – could result in anger.
9) *Voluntarily assuming the appearance of emotions* -there is evidence to support the theory that by making facial expressions related to certain emotions can result in a change of emotional state, such as smiling making some one happy [3].

9 Intensity of Emotion

The category one emotions such as anger and fear are not simply single state emotions such as "I am experiencing fear" but have intensities attached to each emotion [17] such as:

- Anger - ranges from slight irritation to rage or fury.
- Fear - ranges from apprehension to terror.
- Happiness –ranges from contentment to ecstasy.
- Surprise – ranges from startle to an extreme emotion of surprise.

Some researchers propose that emotional intensities can be based on analogue ranges rather than discrete labels where conceivably it is possible to have any degree of intensity; others have tried to have discrete labels related to emotions [20] such as anger having discrete states such as resentful anger, sullen anger and cold anger.

10 Emotional Control & Memory

An emerging area of emotion application has been a drive to empower individuals with skills to help them understand their own emotions (and consequences of) and how to deal with the emotions of others in a social and work environment. These developments suggest that being able to predict behaviour based on observable emotions is an important skill for the individual to attain, a skill previously thought to only belong to clinical practitioners such as psychiatrists and psychoanalysts. This growing area of research has led to new terminology usage for its description including "Emotional Intelligence" [21].

Though labels such as Emotional Intelligence are fairly recent, the research area related to emotional observation and control has been an active part of psychotherapy for over half a century. During this time some conclusions have been drawn regarding how emotions can be managed and controlled by the individual. One aspect of this research involves developing techniques to change responses to emotional triggers, [20] these changes are thought to involve: –

1) Closeness to the avoided trigger.
2) Resemblance to original situation.
3) How early the trigger was learned.
4) Initial emotional charge.
5) Density of experience
6) Frequency of activation of emotion recently.
7) Affective style.
8) Faster stronger emotional responses may have a harder time cooling off.

This control is made more difficult because when an individual enters an emotional state their memory recall skills become focused on information related to the emotion being experienced. This easier recall of emotionally significant memories coupled with the dubious nature of memory recall, such as the tendency to alter memories in an "emotionally gratifying and self-enhancing direction" [22] can reduce the ability to manage the emotion or process other information which could help the emotion to subside [20]. This fundamentally means that memories with an emotional significance may not be true reflections of events that occurred, but are manipulated during recall based on the individual's current state of emotional mind. This may add to conditions such as depression where individuals consistently remember negative memories and feelings, and could certainly be a useful indicator to the presence or imminent onset of such conditions [22].

11 What is a Mood?

Moods are often confused with being an emotion but the difference is that emotions only last for a short time, long enough for the individual to react to a stimulus and take some kind of correctative action. Moods are thought to last much longer, possibly a few days, and are usually linked to the slight background presence of a particular emotion. When an individual is in a mood then whatever slight emotion is present will cause easier activation of related emotions, i.e. if someone is in a 'bad mood' then they more easily enter a negative emotional state such as anger. This easier inducement may be due to triggers not normally associated with the automatic appraisal of the emotion and can be more difficult to observe due to the lack of facial expressions [17]. A consequence of moods is the effect on decision making and judgments that can both become biased towards the emotion underlying the mood [8].

Moods could be a result of unresolved or persistent emotions that have not been overcome fully, and as such are suppressed by the individual to allow 'normal' cognitive functioning. While in a particular mood there is a tendency to respond emotionally to certain stimuli, which ordinarily would not elicit a response. This 'easier' activation could be a consequence of the loss of normal cognitive control of certain emotion triggers, related to these same stimuli. Controls that have been developed over time to avoid automatic engagement of an emotion by a stimulus that has become 'emotionally neutral', such as triggers for annoyance and irritation. This control of emotional trigger stimulus is commonly known as *passive avoidance learning* [23].

12 Individual Emotions

12. 1 Fear

The most researched emotion is fear and is defined as a reaction to appraisals of threat [8]. The main reason for the focus on this emotion is that fear is the easiest emotion to re-create in a controlled environment, and the somatic markers (of the body) are easily identifiable including heart rate, sweat and skin contractions. Because fear is the most studied emotion there is general consensus on many aspects of its neurobiological basis such as the role of the amygdala [10] and subsequent behavioural effects such as the 'fight or flight' response.

The general fear response is thought to consist of [23]:

- *Defensive Behaviour* - such as involuntary freezing, believed to have evolved because many predators respond to movement [10]
- *Autonomic Arousal* – Automatic excitation of several body systems such as blood redirection to the muscles in hands and feet to begin fight or flight.
- *Hypoalgesia* - Reduction of pain from ordinarily painful stimuli brought about from the release of opiates into the system.
- *Reflex Potentiation* – Tendency to face target with eyes wide so to fully focus on stimuli. Reflexes increased through an increase in adrenaline and focus of attention.
- *Stress Hormones* released to engage body systems to run or fight.

Central to the role of fear is the role of the amygdala a small brain region that is thought to 'hijack' many brain systems when a fear stimulus is present. This hijacking occurs to protect the subject by mobilising many body systems rapidly (autonomic arousal) without the need for conscious processing. Conscious realisation does occur later but this is a consequence of bodily changes that are detected by the individual such as 'freezing', hair raising accompanying skin contractions and reallocation of blood to the hands and feet. The amygdala's effect on organs to produce hormones and the hijacking of many systems is called the fight or flight response of fear.

12.2 Aggression

Aggression is another emotion (category two or social emotion) that has been extensively researched and is normally directed outwards at others. Aggression is often classified in one of two ways [24], according to its underlying motivation:

1. *Hostile aggression* is the category of offensive aggression that is directed towards another. This form of aggression is motivated cognitively without an immediate threat, and is often as a consequence of impulsive anger such as anger resulting from frustration [25].
2. *Instrumental aggression* is a form of self-preservation aggression that is directed at immediate threats i.e. self defence. This form of aggression may require more calculated actions in order to remove an imminent threat and as such is often classified as a controlled form of aggression.

Both forms of aggression can push social boundaries to involve physical harm against others and such extreme actions result in violence.

13 Pathological Emotions

A persistent view of emotions is that much of what we know about emotions including how the brain functions is based on clinical studies of patients suffering from pathological conditions such as emotional disorders and brain damage. Much of the ongoing diagnosis and treatment of these conditions owes much to pioneering techniques such as psychoanalysis [26]. These techniques emerged to help deal with the growing range of conditions being identified in the field of mental dysfunction. In recent years, these identifications and related treatments have culminated in two key publications the DSM-IV-TR [27] and ICD-10[28]. The DSM publication is a publication based on an American approach to diagnosis and treatment of mental disorders whereas the ICD-10 publication is produced by the World Health Organisation. These two publications are extremely similar in nature and both centralise much of the current thinking related to mental disorders in order to help diagnostic practitioners deal with patients suffering from a range of mental disorders including:

- *Phobias*- Phobias are a fear related emotional disorder where the fear of a specific stimuli or situation is in excess of the threat posed. Examples are Arachnophobia a fear of spiders and Ophidiophobia a fear of snakes. There is evidence to suggest that extended exposure to the stimuli leads from a fear response to a state of anxiety [10].
- *Panic Disorder* – Panic disorders are diagnosed as a misinterpretation of body sensations such as increased breathing and heart rate brought about by experiences such as slight exertion or excitement. The subject experiencing panic disorders often misinterprets these bodily sensations as indicators that they are in danger or something is wrong and ultimately links the bodily sensations with negative thoughts and feelings which can ultimately develop into further conditions including anxiety and Agoraphobia which is a fear of being afraid [29]
- *Depression* – Depression is a common condition that develops as a consequence of experiencing sadness based on a loss in life. Because loss is an integral part of everyday life not every person who suffers a deep loss will develop depression, some research indicates that around 10% of loss results in depression [29].

13.1 Anxiety

Anxiety is an emotion related to fear. The difference is that anxiety is a reaction to a perceived threat whereas fear is a reaction to a present threat [10]. As already stated fear of a present threat involves many automatic responses, until the immediate threat is overcome and thus the individual gains cognitive awareness reasonably quickly, and can take further steps to deal with subsequent threats or reflect on the actions taken. With anxiety the individual cannot resolve the threat and thus the emotion system interferers with the cognitive system to create a mental state of continual fear that the individual's autonomic and cognitive system cannot resolve, thus:

1) *Anxiety causes worry* – The individual perceives the situation as difficult or impossible to deal with and therefore feels continually threatened.
2) *Anxiety induces negative thoughts* - The individual reflects on negative thoughts such as failing in similar situations previously, and can enter other negative emotional states such as sadness or depression. The individual focuses on failure and foresees failure at overcoming the obstacle or stimulus and thus feelings of self worth are questioned.
3) *Anxiety is obsessive* because the individual cannot focus on anything but the perceived threat, their senses become focused on locating threats in their environment a condition called 'Eysenck's hyper-vigilance theory' [30]. Because anxiety motivates individuals to scan for threats, focusing on other tasks becomes difficult and multi-tasking is almost impossible while in the elevated state of vigilance [29].

There is research evidence to support the theory that a lack of skills to handle many situations might be the root cause of many forms of anxiety such as threats in a social situation causing anxiety due to a lack of social skills, or the fear of failing exams being due to a lack of study/test skills [29]. This lack of skills can present the individual with a growing fear of the oncoming situation, resulting in severe problems when finally faced with the perceived threat because they lack the skills required to deal with the anxiety provoking situation.

14 Emotions & Personality

The final discussion area involves *personality* and is intended to differentiate personality and emotion in order to avoid the common interchangeable terminology usage by the lay-person, such as aggressive, anxious and moody, which is often used to describe a variety of moods, emotions and personality types.

It is not uncommon that personality types such as those used in models of personality devised by Allport [33], Eysenck [31] and Cattell [32] all include categorisation methods, featuring descriptions of personalities that encompass emotions, i.e. aggressive, passive and anxious that are featured in Eysenck's "dimensions of personality" [31]. This is because there is research evidence to support the view that, personality labels are used to describe the common emotional characteristics of a person [14]. This concept has led Plutchik to formulate a list of 67 common personality types, along with the corresponding emotions associated with each personality type as shown in figure 2.

Personality Description	**Emotion 1**	**Emotion 2**
Gloomy	Sad	Annoyed
Hateful	Angry	Disgusted

Sarcastic	Annoyed	Disgusted
Withdrawn	Afraid	Angry

Figure 2. An extract from Plutchik's emotion/personality descriptions [14].

Though it is feasible to identify emotions present with certain personality types, the common usage of personality models is to understand and predict behaviour.

Where theories differ in their approaches are:

- Whether personality can be applied to groups of people, commonly referred to as the *Nomothelic* approach, the models proposed by Eysenck and Cattell fall into this category. The alternative approach involves unique personalities for individual's, referred to as the *Idiographic* approach, Allport's model falls into this category.
- What the personality types are called.
- Which emotions make up the personality types (and what they choose to call the emotions involved.)
- Intensity of emotions that are involved.

15 Conclusions

This paper began with a review of agent emotions in commercial computer games, and highlighted the limitations of the emotions demonstrated by these agents. These characters usually feature only facial animations and limited actions, based on pre-defined scripting, which adds very little to interactivity with players. The review indicated that any pursuance of evolving interactive characters will require an increased depth of agent emotions. As a consequence, four areas have been identified that need to be investigated in order to encourage developers, to create agent's with more sophisticated emotions:

1) Whether agents with emotions will become unpredictable within game worlds, causing problems for developers i.e. how to constrain agents to game architectures.
2) How processing and memory usage will be affected by emotion architectures, and as a consequence will game-play actually improve or should resources be used elsewhere, such as in graphics or sound.
3) How can over 75 years of research in the psychology of emotions be presented, so that it can be modelled by developers?
4) Can conflicting theories regarding emotion be resolved so that developers do not implement a model of emotion that may not be supported by the wider psychology community.

Points 1 and 2 have not been discussed in this paper because they would require an emotion architecture to have already been defined, and any such architecture would require an understanding of the fundamental psychology of emotions. Points 3 and 4 highlight a need for this theoretical underpinning, which has formed the subject of the remaining areas of discussion in this paper.

The research presented, can be clearly divided between terminology usage and theories/research related to emotions (supported by suitable alternatives). Terminology usage has been included to allow a clear distinction between topics involved with the study and implementation of emotions. This distinction is required to remove the widespread variations, in emotion related terminology and therefore, some of the more widely used terms related to emotion have been summarised into a glossary of terms available in section 17.

Though the research presented here has been as exhaustive as possible, some areas of emotion still require clarification in order to answer questions that will ultimately arise during any kind of implementation; several of these have been included in section 16.

16 Further Work

There are many aspects to emotions that require further investigation to aid in the developing of emotions for agents. These aspects include:

- Information will be required, that relates to the timing of feelings based on emotion activation i.e. how long is the delay between emotional activation and conscious realisation.
- Details on the duration of various emotions i.e. how long do they last.
- Some work is required to map a range of secondary emotion names to personality models, to ensure consistent terminology usage.
- The number of secondary emotions will need to be identified, including consistent labels.
- Details of how sophisticated agent-emotions need to be, to enhance immersion and game play.
- A formal architecture will need to be developed, that encompasses the research presented here. This architecture should be suitable for subsequent implementation by developers of commercial games and AI middleware.

17 Glossary of Emotion Terms

17.1 General Terms

Anthromorphic – human like.

Emotions - Emotions are an unconscious reaction, to an appraisal of a stimulus, deemed to require action. This often causes physiological changes and a motivation to act. Emotions originally evolved to aid survival and continued to evolve to deal with ever changing situations in a social context.

Emotion Categories - Two main categories of emotions:

- Category 1 - Six basic emotions commonly labelled anger, fear, disgust, happiness, sadness and surprise. These emotions appear soon after birth and are visible facially.
- Category 2 – Social or secondary emotions including contempt, embarrassment, guilt, jealousy, pride, remorse and shame. There is a wide variation in the number of category 2+ emotions by different researchers. These can be classified as a mix of basic emotions such as:

  Jealousy = Anger + Sadness + Fear +Disgust

More Information:

Facial models of basic emotions see [17] [20].

Secondary emotion combinations see [14].

Emotion Intensity – Each emotion can have an intensity related to it, this can either be a discrete quantity such as furious for an extreme anger or a more fuzzy range relating to intensity such as I am very angry.

Emotion Learning - Nature/Nurture (Hereditary/Environment [32]) – It is commonly believed that human beings are born with a number of basic emotions that emerge soon after birth. Other emotions commonly called social emotions emerge later to allow the individual to interact with other people in a social context. We do not as such have the ability to learn new emotions, only new triggers (See Emotion Triggers).

Emotion Malfunction – A research field commonly called the pathology of emotions includes emotions that are to an extreme, or emotions that seemingly malfunction. Sometimes, depending on the researcher, these are referred to as mental disorders, and may include:

- Anxiety – Reaction to perceived threat, that leaves the individual in a mental and physical condition similar to fear. This causes worry, induces negative thoughts, and focuses attention and memory on negative feelings. Underlying cause is thought to be a lack of skills to deal with seemingly challenging situations.
- Depression – Consequence of sadness, usually from a loss in life, typically a death of someone who is close to them. Approx 10% who suffer a loss manifest depression, resulting in a protracted sad and/or negative mood.
- Phobias – fears of stimulus in excess of threat. Examples include arachnophobia a fear of spiders. The phobia triggers a defensive fear response, without any motivated behaviour to deal with the stimulus. If phobias are not dealt with they can develop into anxiety.
- Panic Disorder – Mis-interpreting of body sensations so that the individual thinks that some thing is wrong.

More Information:

DSM IV 2000 for a range of Mental Disorders including related pathological components.

Emotion Models – Two models of emotions were presented, the OCC model [19] a model that is based on a cognitive view of emotions, and the Ekman model [17] which is based on extensive research in emotions especially observable emotions such as facial emotions.

Emotion Triggers - Emotions are innate, i.e. we are born with them, but the actual triggers for emotions can be changed or additional triggers added to our emotional alert database through techniques such as classical conditioning. It is feasible to reduce the effect of triggers on activation of emotions, through the same conditioning process, which can sometimes lead to extinction of the trigger i.e. lack of activation of the emotion when confronted with the stimuli.

Emotion Stages - Five key stages occur during emotion activation:

1. *Cognitive* – Unconsciously an emotion is triggered by a stimulus.
2. *Motivated Behaviour* – Unconscious (autonomic) action is taken to deal with the immediate stimuli.
3. *Somatic Activity*- Body changes occur as a consequence of changes required to deal with the stimuli i.e. sweating.
4. *Subjective Experience* – Realisation of bodily changes i.e. I am experiencing an emotion
5. *Post reflective period* – Full cognitive control returned, as to be able to take post emotion action if necessary.

Feelings – The conscious realisation, that you are experiencing an emotion based on somatic feedback.

Hedonic – Refers to pleasure or pain.

Homeostasis - self regulation of body systems.

Memory – When in an emotional state, memory recall becomes focused on information related to the emotion. Memory recall has also been shown to be unreliable based on memories sometimes being remembered in an emotionally gratifying way.

Mood – Emotions typically last only up to a few minutes, should the emotion persist for an extended period; then it is classified as a mood. While in a mood (as well as experiencing an emotion), easier activation of related emotions and memories can occur. It is also been suggested that while in a mood, suppression of emotional trigger activations might be compromised, allowing easier emotion activation.

Motivation – Emotions motivate behaviour towards resolving the cause of the activation, this typically involves, focusing the individual's senses and attention on the stimuli.

Personality – a description of the common emotional characteristics of a person. A range of suitable personality models includes [33] [31] [14].

More Information

[31] Includes extensive derivation of personality types via a method called factor analysis. This method involves vast surveys and observations that are then factored to arrive at conclusions on personality types.

Qualia – human experience.

Somatic – is used when describing the bodies systems.

17.2 Specific Emotions

Aggression – (Category 2 – Social emotions) – Includes offensive and defensive aggression. When taken to an extreme, results in violence, i.e. physical harm to others.

Fear – Considered the most primal emotion, it is a commonly supported belief that fear-stimuli activate the fight or flight response when confronted with danger. This occurs in a part of the brain called the amygdala. Cognitive systems can become hijacked to allow the individual to automatically begin dealing with the threat. This prevents a delay that could occur if the individual had time to consider the options.

Autonomic systems activated during fear include:

- Defensive behaviour such as freezing.
- Pain reduction.
- Reflexes increased.
- Blood flow increased to hands and feet to enable fight or flight.

ACKNOWLEDGEMENTS

We would like to thank the referees for their comments which helped improve this paper.

Reference

[1] J. Panksepp. 'Affective Neuroscience'. *Oxford University Press (1998).*

[2] A. Ortony, T. J. Turner. 'What's Basic about Basic Emotions?' *Psychological Review, Vol97, No.3, 315-331, (1990).*

[3] P. Ekman, P. 'Emotions Revealed.' *Phoenix Books.UK. (2004).*

[4] J. Orkin. 'Three States and a Plan: The A.I. of F.E.A.R.'. *Game Developers Conference. San Jose. California. 2006. http://www.jorkin.com/gdc2006_orkin_jeff_fear.doc*

[5] S. Slater. 'Body Mind and Emotion, an Overview of Agent Implementation in Mainstream Computer Games'. *ICAPS 2006. English Lake District. June 2006.*

[6] C. Darwin. 'The Expression of Emotions in Man and Animals' (1872). *3rd Edition. Fontana Press. (1999).*

[7] A. Damasio 'The Feeling of What Happens'. *Published by Vintage, (2000).*

[8] G. L. Clore, Ortony, A. 'Cognition in emotion: Always, sometimes, or never?' *In L. Nadel, R. Lane & G. L. Ahern (Eds.).* The Cognitive neuroscience of emotion. *New York: Oxford University Press,(2000). htttp://www.cs.northwestern.edu/~ortony/papers/Cognition_Emotion.pdf.*

[9] L.E. Crawford, B. Luka, J.T. Cacioppo. 'Social Behaviour.' *Chapter 18. Stevens Handbook of Experimental Psychology 3rd Edition. Learning Motivation and Emotion. Volume 3. Wiley Press, (2002).*

[10] J. Le Doux. 'The Emotional Brain'. *Orion Books Ltd .Phoenix, (1998).*

[11] N.H. Frijda. 'Emotion experience'. *Cognition and Emotion. Psychology Press ltd., (2005).*

[12] R. Morgan, D. Heise. Structure of Emotions. *Social Psychology Quarterly, Vol. 51, No.1, 19-31, (1988).*

[13] J.A. Russell, A. Mehrabian. 'Evidence for a three-factory theory of emotions.' *Journal of Research in Personality, 11, 273-294 (1977).*

[14] R. Plutchik. "A psycho-evolutionary synthesis". *Longman higher education, (1980).*

[15] R. Plutchik. 'A general psycho-evolutionary theory of emotion.' *In R. Plutchik & H. Kellerman (Eds.). Emotion, Theory, research and experience, Vol. 1. Theories of emotion (pp. 3-31). New York: Academic Press, (1980).*

[16] R.B. Dietz, A. Lang. 'Affective Agents: Effects of Agent Affect on Arousal, Attention, Liking and Learning.' *Cogtech, (1999). http://rick.oacea.com/misc/polara/pubs/cogtech1999.html*

[17] P. Ekman, W. V. Friesen. 'Unmasking the Face'. *Malor Books. Cambridge MA,(2003).*

[18] P. Ekman. 'After word.' – *Charles Darwin – The Expression of the emotions in man and animals. (3rd Edition 1999) Fontana press.*

[19] A. Ortony, G. Clore, A. Collins. 'The Cognitive Structure of Emotions'. *Cambridge University Press, (1988).*

[20] P. Ekman, 'Emotions Revealed'. *Phoenix Paperback, 2003.*

[21] D. Goleman. 'Emotional Intelligence'. *Bloomsbury Publishing PLC, (1996).*

[22] R. Gelman, J. Lucariello. 'Role of Leaning in Cognitive development'. *Chapter 10. Stevens Handbook of Experimental Psychology 3rd Edition. Learning Motivation and Emotion. Volume 3, (2002). Wiley Press*

[23] G.E. Schafe, J. Le Doux. 'Emotional Plasticity'. *Chapter 13. Stevens Handbook of Experimental Psychology 3rd Edition. Learning Motivation and Emotion. Volume 3, (2002), Wiley Press.*

[24] K.E. Moyer. 'The Psychobiology of Aggression'. *NY. Wiley, 1976.*

[25] J. Dollard, L. W. Doob, O.H. Mowre, R.R. Sears. 'Frustration and aggression'. *New Haven, CT: Harvard University press,(1939).*

[26] S. Freud 'The Origin and Development of Psychoanalysis". *1st Published in American Journal of Psychology, 21,181-218, (1910).*

[27] 'DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders.' *American Psychiatric Publishing Inc. (2000).*

[28] 'ICD-10: The ICD-10 Classification of Mental and Behavioral Disorders: Clinical Descriptions and Diagnostic Guidelines'. *Published by the World Health Organisation, (1992).*

[29] A. Wells, G. Matthews. 'Attention & Emotion'. *Psychology Press. (1994). Reprinted (1999).*

[30] M.W. Eysenck. 'Anxiety the cognitive perspective'. *Hillsdale.NJ: Lawrence Erlbaum Associates.Inc, (1992).*

[31] H.J. Eysenck, 'Fact and Fiction in Psychology'. *Harmondsworth. Penguin, (1965).*

[32] R.B. Cattell, The Scientific Analysis of Personality. *Harmondsworth. Penguin Books,(1965).*

[33] G.W. Allport. 'Pattern and Growth in Personality'. *NY. Holt Rinehart Winston, (1961).*

Biographies

Stuart Slater is a senior lecturer in IT & Computing at the University of Wolverhampton. His current research involves the development of an 'agent emotion architecture' for use with commercial AI solutions.

Kamal Bechkoum is an Associate Dean in the school of IT & Computing at the University of Wolverhampton. His research interests are in the area of applied AI particularly the use of multi-agent systems in distributed applications

Kevan Buckley is a principal lecturer in the school of IT & Computing at the University of Wolverhampton. His research interests lie in the application of AI techniques to robotics, games, machine vision and forensic sciences.

Digital Puppetry and Talking Toys
Ten emerging theses involving talking toys and technology

Ian Grant[1]

Abstract. Digital artist and lecturer Ian Grant will outline developing trends and scenarios where talking and listening to the speech of humanoid and non humanoid objects, toys, robots is a part of play and other imaginative work. Working from an expertise in puppetry, automata and the emerging new field of 'digital puppetry', the author will take concepts from many disciplines, embedded computing, voice synthesis, performance studies and educational drama (metaxis, role-play, absorption, projection, performativity) and apply them to the world of interactive toys and 'will see what happens'.

To organise the discussion, the author discusses ten emerging ideas in relation to speaking puppets, digital technology and talking toys.

1 INTRODUCTION

"Do dolls have souls? All children talk to their toys, says Charles Baudelaire, and 'the overriding desire of most children is to get at and 'see the soul' of their toys'. Rainer Maria Rilke would agree, with the difference that his child expects the doll to answer back and is disgusted when it doesn't." (Parry, 1994, v) [9]

When artificial objects speak, the presence of breath is intimated and an illusion of life should be more or less guaranteed. However, from our everyday experience of talking technology, we experience alienation, self-conciousness, strangeness, a sense of artificiality or spookiness ('uncanny valley' like feelings) or an acute sense of dysfunctionality.

"In so far as a sound is recognised as a voice, rather than a sound, it is assumed to be coming from a person or conscious agency." (Conner 2000:24) [4]

I will refer to key moments in the history of mechanical and digital talking toys with special reference to the emergence of talking virtual 'creatures', not only as surrogate pets, but as kinds of digital puppets. I will draw on insights from the field of performance studies, particularly: puppetry and digital puppetry, and relate emerging ideas to the human computer interaction of talking digital toys and play objects. I intend to explore how the current technologies of computer based speech synthesis and recognition are being applied in the creation of talking toys and games.

I am interested in what happens to the perception we hold about objects when they 'talk' and are seen to be talking, either through human agency or through embedded technology, like voice chips and facial animatronics. For example: I will discuss the relationship between non-animate talking dolls and different kinds of animated puppets like the 'mouth' puppet and 'character' toys. Other examples of digital talking toys range from objects like talking greetings cards, Texas Instruments 'Speak and Spell' to humanoid or anthropomorphised talking toys like the MP3 playing storytelling bears, like *Teddy Ruxpin* or the *iTeddy*. What seductive, alluring pleasure and magic is lost when our play-things talk back at us? Or gained [2]?

"[The doll] remained silent, not because it felt superior, but silent because this was the established form of evasion and because it was made of useless and absolutely unresponsive material. It was silent, and the idea did not even occur to it that this silence must confer considerable importance on it in a world where destiny and indeed God himself have become famous mainly by not speaking to us." (Rilke 1913/1914 in Parry 1994,33) [11]

1.1 Key Questions

Some key questions and issues:

- What happens to imaginative play and art when 'things' speak?
- What anthropomorphic role does voice, or en-voicing, have when we create virtual creatures or digital pets?
- What are the emerging models of HCI in speech activated and speaking objects?
- How have embedded speech recognition and learning systems been used to devise 'dialogical' digital toys?
- How is digitised speech used in contemporary toys?
- How are talking toys used to create narrative based experiences?
- What is the current state of voice activation and command recognition in toys and how is it being used to stimulate or constrain imaginative play?
- What happens when talking toys automatically move or have facial animation?
- Do talking digital toys enhance 'imaginative' play, story-making and children's role play? What happens to improvisational play when toys both speak and move and are, in effect, programmed automatons?
- What are the trends in the interaction design of 'innovative' toys that embed digital speech technologies?

In order to discuss these questions effectively and include historical and contemporary references to key toys, companies, critical voices, theory, technology, design trends and other ideas, I propose ten emerging theses in section 2 as starting points for discussion.

[1] Thames Valley Univeristy, London, UK., email: ian.grant@tvu.ac.uk

[2] See: http://www.romanceher.com/talkingteddybear.htm - *"A digital message player inside the bear lets you record a 10 second message in your own voice. Each time the bear is hugged, the message plays back. This cute teddy bear is 12 inches tall. The white teddy bear is holding a bouquet of roses and the hidden pocket in the back which holds the recorder can also hold a note, ring, small gift etc. Batteries are included with it. Out of Stock Indefinitely."*

1.1.1 SCOPE

In researching this paper, I have uncovered a quantity of talking technology issues and experiments, both historical and contemporary, that cannot be fully discussed or included here due to space. These include the role of talking technology in deaf and special needs education, patents and animatronics in the techno-entertainment complex, linguistic and semantic modelling in a-life simulations and screen-based games, experiments in analogue and digital voice synthesis in computer games and chat-bots and the like. Emerging trends like the use of embedded audio recording devices to included player voices in games and tangible objects (in "Nintendogs" and "Talking Teddy" from "romanceher.com") This paper is very much an early exploration that, hopefully, will establish and discuss useful questions around the complex of multi-disciplinary issues involved: these areas include developmental and social psychology, learning and play theory, computer science and AI, HCI and design, and creative writing.

For reference purposes, here is a list of the talking toys and objects I've considered for this paper:

- Hasbro's "T.J. Bearytales$^{\text{TM}}$", "Talking Furby$^{\text{TM}}$", Talking ' Aloha Stich$^{\text{TM}}$', "FurReal$^{\text{TM}}$" (they are as the blurb states "for real")
- Mattel's "Teen Talk Barbie$^{\text{TM}}$"
- Backpack Toys' "Teddy Ruxpin$^{\text{TM}}$"
- Thinkway Toys' "Interactive buddies" (their slogan "I'm a thinking toy ®", e.g. "the Buzz LightYear room guard"
- Microsoft's Actimates$^{\text{TM}}$, particularly "Barney the Dinosaur".
- Crying baby simulators (used in UK health trusts to dissuade teenage pregnancy). "Baby Think it Over" in the USA (Stanford University).
- Various Greetings Cards with embedded speech (some recordable).
- Mattel's "Diva Petz$^{\text{TM}}$" with Voice Signal's Speaker-Independent Speech Recognition Technology.
- "Talking Nano$^{\text{TM}}$" (non-representational 'Tamagotchi')
- "Yo, Dude$^{\text{TM}}$", from DSI Toys, Inc.
- "Rock Buddies$^{\text{TM}}$", from MGA Entertainment.
- ""Amazing Amanda", "P.O.D.Z.$^{\text{TM}}$", from Playmates Toys Inc.
- "Hey, Man$^{\text{TM}}$, from Wow Wee
- romanceher.com's "Romeo" Talking Teddy
- Dr Allison Druin's "P.E.T.S".
- Justine Cassell's "StoryMat".

2 TEN EMERGING THESES INVOLVING TALKING TOYS AND TECHNOLOGY

1. Embedding talking technology in toys is more about control than play or exploratory learning.
2. Talking toys are first technological experiments, second, playthings.
3. Talking toys are monologic rather than dialogic.
4. Talking character-based toys, that are dependent on other media, are derivative and closed narrative systems.
5. Talking Toys represent an adult intervention into child-play.
6. What talking toys say is more important than how they say it.
7. Talking toys are of greater value when they are programmable and configurable by children.
8. Animated facial mechanisms attempt to re-embody disembodied voice and, in turn, over-concretise and limit imaginative play.
9. Talking toys are extraordinary simulators of intelligence and presence.
10. Talking toys, traditional and digital puppets and animated media forms are more inter-connected than we may first think.

2.1 Embedding talking technology in toys is more about control than play or exploratory learning

Puppetry, the emerging forms of digital puppetry and puppet like talking toys are all about 'control' in two important senses. There's the good old fashioned sense that such toys are 'cybernetic' systems where there is a feedback loop with the 'movement - action' chain and in the sense that such toys embed pedagogical rules and structures.

"The thing about playing is always the precariousness of the interplay of personal psychic reality and control of actual objects" (Winnicot 1971 cited in Cassell and Ryokai, 1, 2001 [2])

Certain talking toys, the interactive series of 'Actimates' from Microsoft, have been criticised for the empty way they encode, like passive vessels, content from other media channels. 'Barney the Dinosaur', for example is controlled by signals from PCs, TV broadcasts and video tape. The dolls mouth syncs and sings with the representation of Barney on screen. *"Most commercial applications in the domain of tangible personal technologies for children are variants on dolls, with increasingly sophisticated repertoires of behaviours. Microsoft Actimates' "Barney" and Mattel's "Talk with Me Barbie" have embedded quite sophisticated technology into familiar stuffed animals and dolls. These toys, however, deliver adult-scripted content with thin layers of personalization, and do not engage children in their own fantasy play. In both cases the toy is the speaker and the child is firmly in the position of listener."* (Cassell and Ryokai, 2001) [2]

Microsofts learning toy theorist and actimate guru, Erik Strommen positions Barney as a mate, a learning pal, a friend.

But another aspect of 'control', is the propensity of talking toys to lie:

An extended extract from an interactive toy conference review, *"Interactive Barney: Good or evil?"* : *"When I hear Barney say, 'You're my special friend' – that's a disingenuous statement," said Allen Cypher, a founder of Stagecast Software, which designs children's programs. "It's a fraudulent claim. It deceives kids into believing that Barney has some emotional attachment to them, and that's not true." Other panelists worried about Barney's "authoritarian tone," or that he discouraged imaginative play. And some said that, while Barney himself was basically harmless, he may be a harbinger of worse to come: an interactive Cartman from "South Park," perhaps, spewing expletives and insulting his owner"*[3]

And one member of the audience asked if a child could take Barney apart and "reprogram him to say, 'Please slap me.' " *"These products are designed to prevent that," Strommen said.' [7]*

When discussing 'control' it is important to note that it is not meant in a purely sinister, ideologically manipulative, way. Play, and particularly play where children animate and give voice to objects that surround them, is about children asserting control and (dis)order over facets of their environment: *"One essential aspect of childrens' spontaneous storytelling play is that it is child-driven. And this is important since children feel a sense of achievement and empowerment when they know that they can create and control the content of their play objects. So, if technology is to encourage childrens' creativity and, in particular, play a role in childrens' storytelling play, it must not dampen that child-driven aspect of their play."* [2]

[3] See Hasbro's 'Aloha Stitch Doll' for an example of such a moody toy.

2.2 Talking toys are first technological experiments, second, play-things.

The history of talking toys and automata is clearly a story of technical innovation and development for the purposes of celebration, entertainment and play. According to Jasia Reichardt talking statues have been known since 2500 BC: *"Some incorporated concealed speaking trumpets through which someone hidden could address a gathering. The idea was that gods communicateed through the statues which represented them"* (Reichardt, 1978, 9)[10] Of interest here, Jacques de Vaucanson created a number of mechanical automata including a 'flute player and defecating duck' (circa 1737-1738). On the 'flute player': *" This automaton 'breathed'. Even though the art of mechanics was sophisticated enough by then to make the machine perform many other movements, and even though Vaucanson unveiled the fact that this breath was created by bellows, the very act of breathing, seen in an inanimate figure, continued to cause a stir well into the following century."* (Wood, 2002, 21-22) [19]

The first talking doll was patented by Johann Nepomuk Maelzel in 1824. According to Gaby Wood *"He designed a pair of bellows that, when attached to a tube, a widening oral cavity and a set of valves, could say 'papa' or 'maman'"*. (Wood 2002, 118-119) [19]

Thomas Edison's 'Talking Doll' (1891) - conceived as an advertisement for his sound recording device - embedded a miniaturised phonograph mechanism that played wax cylinder recordings of nursery rhymes, prayers and stories: *"[The phonograph] began by speaking the words of a child, and it was not long before a child was invented to give it shape, or to give it life. So the capturing and reproduction of speech were accompanied by a casing for it in human form"* (Wood, 2002, 18) [19]

The context around Edison's toy development has shaped the industrialised processes surrounding technical innovation and toys ever since. There is little perceived difference between Edison grafting a mechanical phonograph into a toy and *iTeddy*'s implanted mp3 / mp4 playback device. Yet the former was an exercise in creating perfect representational forms of human (female) life, and the other a toy to placate media hungry children.

"Edison's colleague, W.K.L. Dickson, wrote that it was 'perhaps the daintiest and most suggestive of all the multiform uses to which the phonograph has been put.' He described 'roseate lips' which would 'lisp out the oft-conned syllables of nursery rhymes, pipe the familiar of Mother Goose's ballads, and give forth the cooing and wailing sounds of baby life Under such auspices into what enchanted realm will our ordinary toys be transformed." (Dickson cited in Wood, 2002, 114) [19]

Duncan Bannatyne, on a recent broadcast of BBC TV's venture capital reality show [4], said of *iTeddy*: *"I'm so sad. Reading bedtime stories is a father's [sic] job. I don't want to be replaced by a teddy bear."*

Talking toys that emerge from University research labs and university start ups are philosophically worlds apart from corporate toys from the likes of Microsoft, Disney franchises and the enormous toy companies like Hasbro and Mattel. The work of Justine Cassell at MIT with *"StoryMat"*, Dr Alison Druin with the *"PETS"* projects [5] (from the *Human Computer Interaction Lab* at the University of Maryland) are distinct in pedagogy and interactive strategy from most commercially available toys. The toys have a clear philosophy of use as 'learning technologies' rather than simply embedding the latest speech recognition and synthesis chips in order to maximise rich play or to aim for 'realism', or to service a franchise.

It should be noted that sponsorship relationships exist between the toy companies and innovative research groups in universities. An extended quotation from David Shenk's article *Behold the Toys of Tomorrow* illustrates the connections between technological innovation, the toy corporations and the University researcher. It also connects:

"The computerisation of toys also dovetails nicely with the ambitions of computer evangelists, those whose life's mission it is to deliver the power of computation into every aspect of every person's life. Nicholas Negroponte, the director of MIT's famously innovative Media Laboratory (the Vatican of techno-evangelism), noted last year in his Wired column that toys are the "fastest evolving vehicles on the infobahn," meaning that because of their astonishing turnover rate (each year, 75 percent of the toys on shelves are newly designed), they're the only class of objects that can truly keep up with the rapid pace of hardware and software innovation. That, combined with the tantalizing prospect of winning young, impressionable children over to the virtues of computers, has catapulted toy technology into high-priority status for the Media Lab. While researchers there have been exploring the issue for decades, they substantially upped the ante earlier this year with the formation of an industry-research consortium called "Toys of Tomorrow." A dozen or so companies, including Mattel, Tomy, Intel, and Bandai (makers of the infamous Tamagotchi "virtual pets"), have signed up, committing to at least three years of the $250,000 annual sponsorship fee. In return for the funding (a modest R&D investment for any sizable company), sponsors get first crack at the new technology and ideas – a head start that seems bound ultimately to be worth many times that sum.

The Media Lab is a Willy Wonka factory for technophiles, where the only limitations are in the creators' imaginations. Intoxicated by the MIT fumes, one thinks: How could this not be a boon to society?" (Shenk, 1999) [12]

It should be noted that talking toys and animated toys are often adult orientated, rather than for children. This may be because such devices express the extraordinary fascination with what contemporary technologies can do. Jacques de Vaucanson's 'defecating duck' was not a toy - but a remarkable exploration of what clockwork and air power could do. Likewise, Edison's talking doll:

"One can only conclude that the [Edison's] dolls were not for children, and adults like [Albert Hopkins (Editor of 'scientific american' c1890)] were not alone in picking up on their aggressive horror. Formanek-Brunell quotes a survey taken at the time Edison's dolls were manufactured, in which a four year old girl, fusing the animate with the inanimate in a way that recalls Vaucanson's duck, said she didn't like talking dolls, because 'the fixings in the stomach are not good for digestion'." (Wood, 2002, 118) [19]

2.3 Talking toys are monologic rather than dialogic

Talking toys are rarely conversational agents, and interaction is heavily pre-determined. Randomising responses is one strategy that provides an illusion of knowing, active conversation. Such illusions are broken when the pattern or repetition is noticed.

Arguably, talking toys fix patterns and structure play and are, in nature, didactic and instructional. However, structured or programmatic experiences are crucial to learning, play and language development. I am not simply dismissing such toys dogmatically from some valorised over-emphasis on the value of 'free play'. Lev Vygotsky:

"Let us turn now to the role of play and its influence on a child's

[4] See http://www.bbc.co.uk/dragonsden/

[5] P.E.T.S. - "Personal Electronic Teller of Stories" robotic pets that support children in the storytelling process

development. I think it is enormous. I think that play with an imaginary situation is something essentially new, impossible for a child under three; it is a novel form of behavior in which the child is liberated from situational constraints through his activity in an imaginary situation." (Vygotsky, 1933) [18]

I would argue that more 'situational constraints' are imposed by over-structured, adult led, media influenced play, than child centred play. It may be a useful moment to introduce the term 'metaxis' used in education drama contexts for describing the 'dualness' of perception during role-play and 'as-if' contexts. Metaxis has been defined as *"the state of belonging completely and simultaneously to two different autonomous worlds"* (Boal 1995, 43) [1]. This definition has interesting resonance when considering the virtual realities created by digital talking toys.

2.4 Talking character-based toys, that are dependent on other media, are derivative and closed narrative systems

Often talking toys become extensions of pre-existing media that project ideas outwards, from toy to child, rather than being empty vessels that facilitate projection from child to toy. This is particularly acute in any toy that represents known characters from other media productions, the huge so-called 'character toy' market.

What difference does it make if the imaginary topic of make-believe style play with talking toys is sourced from existing media, rather created from within than the child herself?

Justine Cassell creates story environments and interactive objects to research the quality of technology assisted spontaneous play and story creation. She [2] carefully documents and quantifies the generative, creative effects of certain (I would call 'dialogical') interactive technologies. Using quantitative and visualisation methods, she carefully annotates the original spontaneous vocal contributions offered by children while playing with interactive toys. She also transcribes and qualitatively analyses the text of stories children create using her 'environment'.

2.5 Talking Toys represent an adult intervention into child-play

First, adults buy toys and design toys, and their associated pedagogy, for children. Children of a certain age exert pressure and express desires for certain toys, stoked by the marketing messages of the larger toy companies. Most talking toys enshrine messages and pedagogy from adults to children and on occasion seek to replace or act as surrogates for social parental contact: A selling point of a recent the UK designed toy, *iTeddy*, a strange 'Tellie-Tubbies' and iPod hybrid, a bear with a media player embedded in it's stomach, was the 'comforting' effect the toy had to placate children during the absence of a peer, buddie, parent or supervisor. The surrogate suckling / child rearing function of childrens media and TV are transferred into the toy itself. Like many toys of this ilk, *iTeddy*, refreshes its onboard content of nursery rhymes and stories using networked connectivity to a custom web-site. Interactive toy guru, Erik Strommen is having a fascinating career that takes in companies as diverse as "The Childrens Television Workshop" creators of educational puppet-fest *"Sesame Street"*, several game companies and Microsoft, where he worked on and promoted the Actimate series of interactive toys. In his 1999 paper *Learning from Television With Interactive Toy Characters As Viewing Companions"* [16], and in later work [6], builds an extended theory of how talking interactive toys can act as 'scaffolding' for learning interactions, act as 'buddies' and simulated co-learners. In more recent work, Strommen clearly delineates between interactive toys as surrogates for adult interventions, toys as establishing shared contexts for extended social interactions (i.e. such toys need parental supervision and interaction) and pure play without any pedagogical intent: *"Whos in charge? If the children are the ones setting things up not just physically but conceptually, if they are showing each other what to do, collaborating, its play. If the children are being told what to do, led, directed, or tested, its not play"* (Strommen, 2004,) The more interesting counter-examples of toys not obviously promoting language acquisition are speaking toys that babble and create their own non-human languages that parody child language. Such toys do not induct nor reinforce the adult designed language structures of nursery rhymes, alphabet led rote learning and traditional language learning games Such toys and characters, eg. the Norns in the Creatures series 'language' [3] and the talking Furbies native language 'Furbish', communicate through prosody and gesture and are not limited by the need to process real language structures. Bizarrely (and wonderfully - in terms of creativity and useless play) such toys have lead to the players acquiring and learning nonsense languages.

On developing "Nornish":

"We decided to look for a way of converting anything the Norns said into sounds, in such a way that a) the words sounded like speech, and b) a word would sound the same every time it was spoken, and c) different words should have different pronunciations. Just to make life difficult for myself, I also added d) similar words should sound similar. The first step was to record some speech. Luckily, one of the artists working on the game had something of a gift for making bizarre noises, so we gave him a script full of gibberish and recorded him babbling away. This was then chopped up into individual syllables, and electronically treated to give male and female voices. Having been presented with a large collection of syllables, I went through them and decided whether they sounded like the start of a sentence, the middle of a sentence or the end of a sentence. Having established these groups, I let "nature" do the tricky bit for me. I came up with a way of using a random set of numbers to convert any group of three letters into one of these syllables. I then let these random numbers "breed" until I had a vocabulary that fitted all my requirements - all groups of letters had a corresponding sound, similar groups sounded similar, and I could recognise the starts and ends of sentences. Norns have a very small vocabulary, so in principle, it should be possible to learn to understand "Nornish" - I confess I've never had the patience, though I'd love to hear if anyone has." (Peter Chilvers, no-date) [3] People have learnt "Nornish".

2.6 What talking toys say is more important than how they say it

Talking toys enshrine a pedagogy that has, in effect, remained unchanged since Edison's 'talking doll' of the late 19th Century'. The pedagogy is built on cautionary tales and stories, wrote learning of nursery songs and instruction led play.

Elsewhere in this paper, I have mentioned Microsoft's Actimate 'Barney the Dinosaur' being accused of lying - mainly because he doesn't have a consciousness yet talks freely of love. The ideological

[6] *When the Interface is a Talking Dinosaur: Learning Across Media with ActiMates Barney* (1998) [13] *Learning from Television With Interactive Toy Characters As Viewing Companions* (1999) [14] *Interactive Toy Characters as Interfaces For Children* (2000) [15]

function of what talking dolls say is critical when evaluating talking toys.

The New York based Barbie Liberation Organisation (BLO), a group of feminist hactivists formed circa December 1993, undertook a remarkable operation to swop the voice boxes of 300 Barbie and G.I. Joe dolls.

"When Barbie speaks, little girls listen, which is why controversy erupted in 1992 when Teen Talk Barbie exclaimed, "I love shopping," "Meet me at the mall," and "Math class is tough." This last phrase struck an especially sour note, given the under-representation of women in the sciences" (Dery 1994) [5]

"The Simpsons" episode *Lisa vs. Malibu Stacey* is a wonderful parody of the 'Teen Talk Barbie' controversy. The episode explores issues of what dolls are programmed to say and activist / feminist responses to such things. [7]. There is even a subtle reference to the underground work of the *Barbie Liberation Organisation* when the girl doll "Malibu Stacey" is heard to say *"My Spidey-sense is tingling. Anyone call for a webslinger?"*.

Although the technologies of embedded speech are fascinating, whether using the latest embedded microprocessers for speech recognition and synthesis processors [8], bellows and air, miniature phonographs, it seems that whenever toys talk the meaning of the iteration outweighs the possibilities inherent in technical act of production.

2.7 Talking toys are of greater value when they are programmable and configurable by children

2.8 Animated facial mechanisms attempt to re-embody dis-embodied voice and, in turn, over-concretise and limit imaginative play

"Phenomenologically, there is a close relationship between the voice and the face; both the voice and the face are parts of us that are turned outwards and by which the world knows us, but which we can ourselves only see or hear partially. They signify intimacy and vulnerability. We are our faces and we are our voices . " (Conner, 2000, 401) [4]

It is a trend in recent talking toys, to have complex animated faces. This is in part due to a desire to a create an illusion of an 'embodied' voice. Disembodied voices tend towards the 'uncanny'.

In the history of puppet theatre, there is a pronounced difference between forms that attempt to locate the voice 'within' an object by rhythmically mapping gesture and movement to voice and articulated face parts, and those forms that rely on the play of light on static sculpted forms of faces to suggest expression and facial motion. The misbegotten primary aim of 'more' articulation is 'more' realism - greater verisimilitude in the imitation of living forms.

This is echoed in the aims of makers Hasbro and Voice Signals recent technology expressed in a press release. The seek to offer a more interactive, 'richer' play experience through speech activation and recognition:

"We are extremely pleased that Hasbro has selected Voice Signal's MicroREC speech recognition software for this product [Aloha Stitch],' commented Stewart Sims, Voice Signal's executive vice president of marketing. 'Hasbro has a well-deserved reputation for creating fun, innovative, quality products, and we are delighted that they have chosen Voice Signal to supply the speech recognition software that increases the interactivity of their toys and brings 'Aloha Stitch' to life." [17]

In Hasbro's 'Aloah Stich', the Voice Signal voice chip creates a 'bi-polar' toy, that varies its responses to a set number questions according to one of two moods. I quote an online review of the toy by a parent at length as it is hard to access information about the performance and interactive sequences of most of the toys under consideration in the present paper:

"So just what kind of smack does Stitch talk? Here's a rundown of cues and replies

You say - "What's your badness level" He says - "Mostly Good Today" or a sheepish, "I'm having a pretty good day" when he's nice and Naughty! Blarrrtghll! when he's rotten.

You say, "Are you hungry?" He says, "Coconut cake and coffee, please." when he's nice and "Not anymore, I ate your dinner" when he's rotten.

You say, "Sing a Song." He sings Aloha Oe when he's happy and burps it when he's rotten (way too funny).

You say, "I know you can talk." He says, "Okay Okay" when he's nice and "Doggies cannot talk" or "Bark! Bark!" when he's rotten.

You say, "Got to Sleep" He says, "Very Sleepy. . Snore" when he's nice and barks, "No, make me a sandwich!" when he's rotten.

You say, "Where are you?" He says, "I'm with my family" or a very sad, "I'm lost" when he's nice and "Stinky Planet Earth" (which comes out yarth) when he's rotten.

You say, "Will you play?" He replies, "Surf's Up! Cowabunga!" or "I can't I have nothing to wear." when he's happy and "I'm busy get lost!" or "I'm busy you go away, okay?" when's he's rotten."

(anon, 2004) [8]

Inanimate children's toys, which the child enlivens with movement and either unspoken or spoken voice are, to me, more enduring playful simulations - as the mutability, the changeability, the fluidity of roles are emphasised through imaginative projections on a static face, rather than the repetitive 'fixed movements' of most automated articulated movement.

In an article on 'Mike the Talking Head' (an extraordinary head mounted facial performance capture system and CG digital puppet documented circa 1988), Valarie Hall comments: *"When it all comes together, the quest for realism in character animation is but a test for the animator to find out how good he/she is at using the tools they have at their disposal."* (Hall, nodate) [6]

Any aesthetic assessment about 'realism' in talking toys and how it effects play needs to balance manufactures promotional excitements with an assessment of the whole interactive system.

3 CONCLUSION

By means of conclusion, I will simply state the two remaining theses and leave them hanging for further cogitation. Also, this study has uncovered a topic richer and more varied than I had initially speculated and it is a pregnant ground for future research.

[7] First broadcast: in "The Simpsons" Season 5, September 30, 1993 May 19, 1994

[8] See the wonderfully named E.L.V.I.S. "the Embedded Large Vocabulary Interface System platform" from VoiceSignals Technologies. [17]

3.1 Talking toys are extraordinary simulators of intelligence and presence

3.2 Talking toys, traditional and digital puppets and animated media forms are more inter-connected than we may first think

ACKNOWLEDGEMENTS

I would like to thank the referees for their comments which helped improve this paper and the symposium conveners for their patient support.

REFERENCES

[1] Augusto Boal, *The Rainbow of Desire: the Boal Method of Theatre and*, Routledge, London, 1995.

[2] Justine Cassell and K Ryokai, *Making Space for Voice: Technologies to Support Children's Fantasy and Storytelling*, http://web.media.mit.edu/ kimiko/publications/PersonalTech.pdf, Last Modified: 2001. Date Accessed: 01/03/2007.

[3] Peter Chilvers, *Creature Labs - Norn Babblings*, http://www.gamewaredevelopment.co.uk/creatures_more.php, Date Created: nodate. Date Accessed: 01/02/2007.

[4] Steven Conner, *Dumbstruck: A Cultural History of Ventriloquism*, Oxford University Press, New York, 2000.

[5] Mark Dery, *Hacking Barbie's Voice Box: 'Vengeance is Mine!'*, http://www.levity.com/markdery/barbie.html. New Media magazine, "Technoculture" column., Date Created: 05/1994. Date Accessed: 1/2/2007.

[6] Valarie Hall, *Mike (the talking head)*, http://mambo.ucsc.edu/psl/mike.html, Date Created: nodate. Date Accessed: 01/02/2007.

[7] Michael Newman, *Interactive Barney: Good or evil? Conferees worry about where computerized 'character' toys are going next*, http://www.post-gazette.com/businessnews/19990521barney1.asp, Date Modified: 21/05/1999. Date Accessed: 01/03/2007.

[8] Anonymous Parent, *Read Reviews of Hasbro Aloha Stitch Doll 3570 at eOpinions*, http://www.epinions.com/content_163285929604?linkin_id=8003929, Date Created: 28/11/2004. Date Accessed: 01/02/2007.

[9] Idris Parry, *Essays on Dolls*, Syrens, London, 1994.

[10] Jasia Reichardt, *Robots : fact, fiction and prediction*, Thames and Hudson, London, 1978. (by) Jasia Reichardt. ill(some col), plans, ports ; 28cm (Pbk).

[11] Rainer Maria Rilke, 'Dolls: On the wax dolls of lotte pritzel.', in *Essays on Dolls*, ed., Idris Parry, Syrens, London, (1994).

[12] David Shenk, *Behold the Toys of Tomorrow (The Atlantic Online - Digital Culture)*, http://davidshenk.com/webimages/atlantic1.htm, Date Created: 07/01/1999. Date Accessed: 01/02/2007.

[13] Erik F. Strommen, *When the Interface is a Talking Dinosaur: Learning Across Media with ActiMates Barney*, http://www.playfulefforts.com/archives/papers/CHI-1998.pdf, Online PDF of published work. Date Written: 1998. Date Accessed: 01/03/2007.

[14] Erik F. Strommen, *Learning from Television With Interactive Toy Characters As Viewing Companions*, http://www.playfulefforts.com/archives/papers/SRCD-1999.pdf, Online PDF of published work. Date Written: 1999. Date Accessed: 01/03/2007.

[15] Erik F. Strommen, *Interactive Toy Characters as Interfaces For Children*, http://www.playfulefforts.com/archives/papers/IA-2000.pdf, Online PDF of published work. Date Written: 2000. Date Accessed: 01/03/2007.

[16] Erik F. Strommen, *Play? Learning? Both...or neither?*, http://www.playfulefforts.com/archives/papers/AERA-2004.pdf, Online PDF of unpublished work. Date Written: 2004. Date Accessed: 01/03/2007.

[17] VoiceSignals Technology, *VoiceSignals Technology Press Release*, http://www.voicesignal.com/news/press/release_02_19_02.html, Date Created: 19/02/2002. Date Accessed: 1/2/2007.

[18] Lev Vygotsky, *Play and its role in the Mental Development of the Child*, http://www.marxists.org/archive/vygotsky/works/1933/play.htm, First Published: 1933. Date Created: 2002 . Date Accessed: 01/02/2007.

[19] Gaby Wood, *Living dolls : a magical history of the quest for mechanical life*, Faber, London, 2002.

Socially Promiscuous Mobile Phone Pets

Sean Casey and **Duncan Rowland**[1]

Abstract. Virtual pets offer an abstracted version of animal ownership. Currently, most simulated creatures tend to be sequestered from the real world and so have little or no knowledge as to their whereabouts. This limits the scope of the possible interactions between the player and their simulated pet. Gophers are artificial pets designed to be carried by users on their mobile phone. They are socially promiscuous (like cats) and enjoy visiting the phones of other players without the permission of their 'owner'. During their lives, these artificial animals collect topological information regarding their movements, and semantic data related to their location via interactions with the humans they encounter. Future versions of the game will utilise this information to create a more contextually aware experience.

1 INTRODUCTION

Traditional computer game platforms (PC and Consoles) tend to know little about their location and so offer the player a contextually ignorant experience. Ignoring for the moment the robotic variety, the majority of virtual pets run on standard gaming platforms such as these. The physical capabilities and affordances offered by these devices limit the depth of interaction offered by the pet.

The use of pervasive computing technologies (e.g. mobile phones and PDAs) to create socially rich, location sensitive entertainment experiences has been explored by recent investigations into pervasive gaming [1][6]. This paper suggests that by utilising such technologies, virtual pet experiences could be enhanced to create experiences sympathetic to the player's local environment. The paper discusses the implementation of an experimental virtual pet game called 'Gophers' [3]. It investigates some of the potential for releasing virtual pets into the real world, so they are able to visit other players and collect information regarding to their surroundings.

In the following sections, the design of the game is discussed, along with technological and development issues. A short synopsis of the Gophers gameplay is given and finally future research directions are discussed, along with possible enhancements that could be made to virtual pets through use of pervasive computing technology.

1.1 Social Butterflies

A key design goal of Gophers is to move the virtual pet experience away from the computer screen and into the real world. The system comprises of a pervasive gaming platform in which virtual creatures inhabit the physical world. This platform is based upon the Hitchers framework [5], developed at the University of Nottingham. 'Gophers', are highly customisable beings, able to represent any type of virtual pet the player conceives (and so are not just limited to representing small furry creatures!) Gophers are spatially located, given a real world location, just as real pets would be. As players move around the world, they encounter new gophers and these can be picked up and transferred onto their phone. Players create their own personalised creatures, giving them a unique look, name and assigning them a real-world task to complete. These new creatures are then released into the wild where other players can pick them up to help them with their missions.

Gophers are generally personable creatures that assimilate content from players through their interactions with them. They build personal narratives (stories about their travels), as they move around and these are presented in the form of blogs that players can access these through the Gophers website. This serves two purposes; firstly it enables players to keep track of their pets and secondly, it allows other players to decide if a gopher's mission has been accomplished.

1.2 Mobile Gaming Platforms

Mobile phones currently present a useful platform for the investigation of virtual pets. The recent mass adoption of mobiles phones has taken society to point where mobile computing technology is almost fully ubiquitous. These phones are mobile, highly personal general purpose computing devices, with users customising their handsets through downloaded ring tones and wallpapers. Additionally, the benefits of mobile data communications and location awareness can be utilised for the purposes of the study; communications, to allow the creature state and player interactions to be transmitted to a central server and location of the handset used as a first class element of the experience. The benefits of these platforms are already becoming revealed, through mobile social games, such as You-Who [18]. Clearly they also offer an excellent foundation for investigating pervasive virtual pet experiences.

2 RELATED WORK

Through augmenting the concept of virtual pets with mobile and pervasive computing technologies, there are a number of ways in which the user-pet experience can be enhanced:

2.1 Virtual Pets in the Real World

Mobile pets were some of the earliest virtual pets conceived, for example Tamagotchi [15]. Their recent move to mobile phone devices shows their continued popularity (with two virtual pet games featuring in the top wireless games of 2006 [14]). The

[1] University of Lincoln, U.K., email: scasey | drowland@lincoln.ac.uk

appeal of the mobile pet is their ability to remain with the user during their daily routine. Most of these could not claim to be aware of their surroundings or context and hence, are not truly pervasive. Despite this, there are two commercial examples which do contain pervasive features:

One pet which could be described as contextually aware (although non-mobile) is MOPy fish [12]. The pet monitors the 'Multiple Original Printouts' feature on HP printers and awards points for its use. These points can later be used to purchase items for Mopy's aquarium. However, the positive reward for use of printers is not particularly synonymous to looking after pets in reality.

Pets such as Vmigo [17] offer a much more realistic interpretation of this, with the pet's (dog) wellbeing linked directly to the players physical actions. When attached to a television, it offers a static game experience, but additionally includes a handheld pod with inbuilt pedometer, which can be detached from the device. A player must take this pod walking with them in order to exercise their dog.

Although Vmigo sees the virtual pet world being influenced by a player's actions in the physical world, the two worlds represent distinct and somewhat segregated experiences of the game. To create a true pervasive experience requires the combining of the virtual pet worlds with the physical. Furthermore, the game should be designed to run on familiar mobile devices, rather than the proprietary hardware favoured by Vmigo.

2.2 Pets with Original Narratives

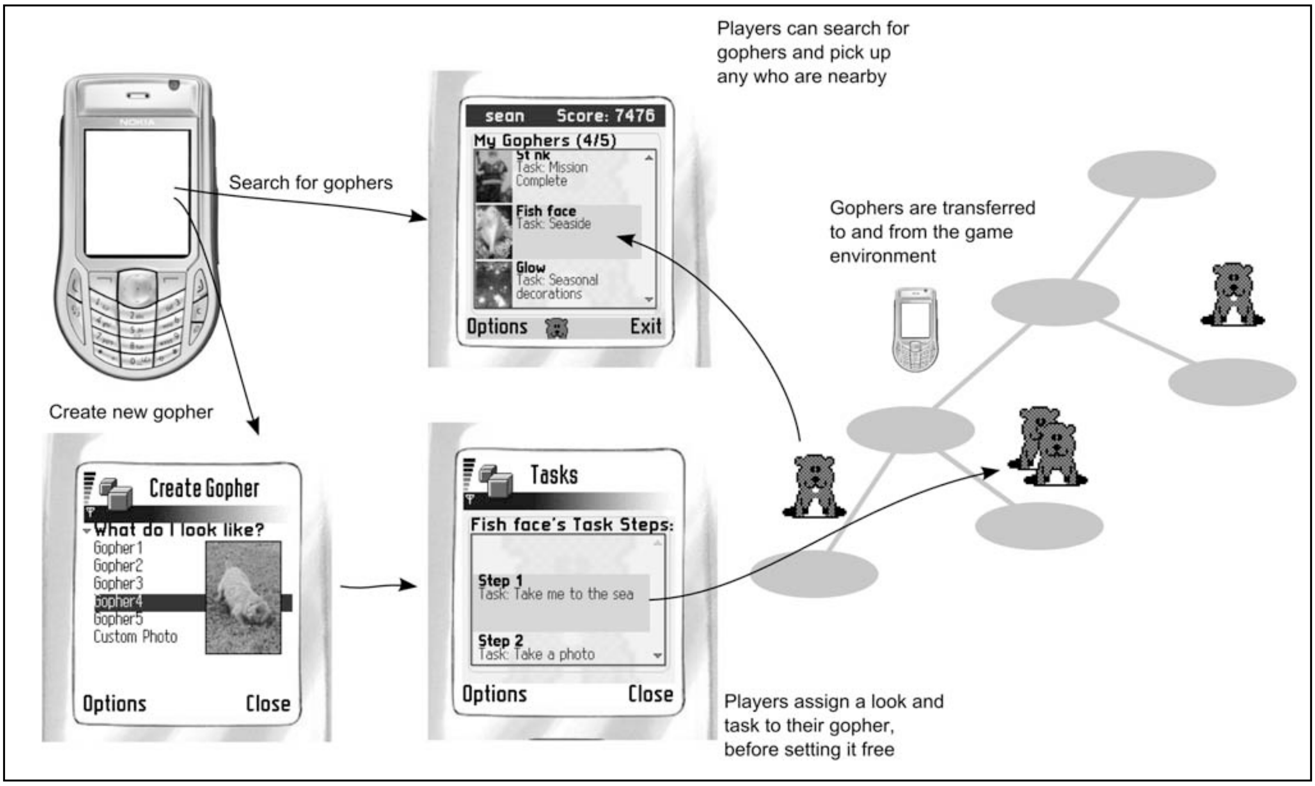
The concept of creating 'content trails' through situated media [4] provided inspiration for original game narrative. Additionally, the BackSeat Gaming initiative [2], have used dynamic content trails as the basis for a game narrative. Pets in the real world learn many things about their environment and as gophers continue on their missions, they continuously generate content trails in relation to their missions. These can be interpreted as an evolving narrative, to which players collaboratively engage in (in a similar style to TxTBook [16]).

Generating narratives autonomously from community play presents a number of advantages over scripted ones. Firstly, the story is completely open to interpretation of players, leading to original and unpredictable twists in the story. Secondly, it presents a far less expensive solution, in terms of system administration and moderation. Finally, the localised nature of many gophers means the narratives are also based on localised content, giving the stories a more personalised feel.

2.3 Virtual Pets and Social Networking

The possibilities of social networking in a virtual pet community have been highlighted by the success of products such as GoPets [7] and Nintendogs [13]. In the GoPets online environment, players can create pets who will travel around the online world, either on request of the player, or of their own accord. This feature, combined with the 'IKU' universal language, aim to allow for distributed online social networking.

In extending virtual pets into a pervasive, rather than online experience, it is possible to offer much more personalised networking. Two players may frequent the same café, to allow their Nintendogs to visit each other, for example. It is important to take note of implications these applications may have in terms of privacy; 'bark mode' (where stranger's dogs are allowed to 'visit' and exchange voice recordings) for example has potential


Players can search for gophers and pick up any who are nearby

Search for gophers

Gophers are transferred to and from the game environment

Create new gopher

Players assign a look and task to their gopher, before setting it free

Figure 1. Acquiring gophers (adapted from [3])

issue in regards to this – especially when considering children playing the game.

3 GAME EXPERIENCE

On starting the game, the player is presented with a scrolling list of the gopher pets they possess. A player can hold 5 gophers on the phone at any one time.

Initially, no gophers are present on the phone and the player has two options:

i) Search for Gophers: Searching will return an ordered list of gophers that are located at physically nearby locations. With each gopher returned, an indication of its mission and relative distance is given. Players can pick up any gopher they are interested in adopting. When the Gopher is picked up, it is transferred from its physical location to the phone. However, the further away a gopher is, the more expensive and longer it will take to arrive.

The notion of travel time enhances the feeling the player exists in a world inhabited by gophers, each of which has their own physical relation to the player, by reinforcing the mapping between the location of the Gopher and the real world. Gophers lie dormant at their locations until being picked up by a player, again strengthening the illusion that these are pets, who require an owner's assistance.

ii) Create a new Gopher: Players are given the capability to create a new gopher after acquiring sufficient points (by helping other players gophers). The player provides a customised camera photo and name, to represent the gopher creature. Following this, the player assigns the gopher a mission, for which it is the gopher's purpose to complete.

In keeping with existing virtual pets, the customisation gives the gopher a personalised feel. The additional ability to specify a real-world mission provides a continued interest for the player, long after the novelty of owning the pet has worn off.

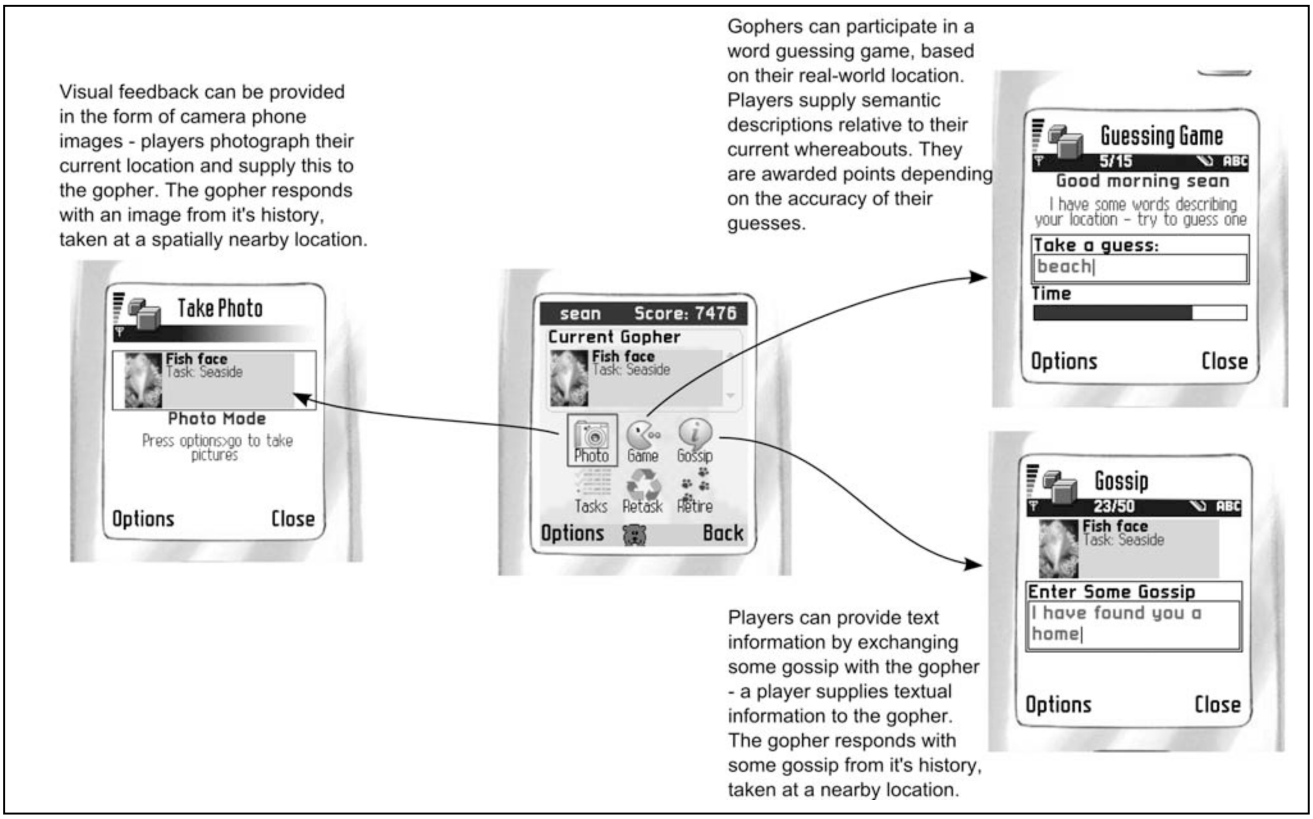
3.1 Assisting the Gophers

The ultimate aim of each pet is to successfully complete their assigned mission. Players can help gophers to do this by providing information relevant to their tasks. This is achieved through player-pet interactions. There are three modes of interaction: *photo mode, gossip* and *guessing game*, which are described in *Figure 2*. Each offers a unique way for the player to supply information and if a gopher has collected any recent content which may be of interest to the player, it will respond by returning this.

Like real pets, gophers enjoy human interaction and have a limited attention span. If they become bored they will abandon the player, through jumping from the virtual domain of the phone, back into the physical world.

3.2 Monitoring Gophers

Gophers can participate in many missions during their lifetime and can continue exploring long after a player has dropped


Visual feedback can be provided in the form of camera phone images - players photograph their current location and supply this to the gopher. The gopher responds with an image from it's history, taken at a spatially nearby location.

Gophers can participate in a word guessing game, based on their real-world location. Players supply semantic descriptions relative to their current whereabouts. They are awarded points depending on the accuracy of their guesses.

Players can provide text information by exchanging some gossip with the gopher - a player supplies textual information to the gopher. The gopher responds with some gossip from it's history, taken at a nearby location.

Figure 2. Interacting with gophers (adapted from [3])

them. Because of this, players tend to retain an interest with their gophers they have looked after in the past. The online blog viewer allows players to view all the interactions and encounters they have experienced.

The blog viewer provides the secondary function of allowing players to determine whether a gopher has completed their assigned mission. If this is the case, the gopher stands trial in jury service, where a panel of jurors, selected from the game players determine the success of the gopher's mission.

3.3 Limitations

Gophers are a simplistic representation of a virtual pet. Because the study concentrates on the pervasive aspects of the game and to retain a simple user experience, many of the classical interaction methods typical to existing virtual pets (such as feeding and petting), are not included in the simulation.

4 TECHNICAL CONSIDERATIONS

In implementing Gophers there were significant technical challenges, which needed to be met. The implementation of the game was achieved using Nokia Series 60 2nd Edition mobiles, without the need for any more specialist hardware (for example PDAs, GPS receivers). This allows the game to be played on standard equipment, without the risk excluding those from non-technical backgrounds.

The application makes use of GSM cell mast locations to approximate physical locations of players, gophers and in-game events. Through use of Placelab software [8], rather than expensive operator positioning services, it is possible to freely read the unique identifier of the mast to which a phone is currently connected. This generally represents a physically nearby mast, (although not necessarily the nearest in some circumstances [5]). Although the absolute geographical location of these masts is unknown, they are used in the game to calculate the relative distances between game events (for example the distance between a player and a nearby gopher they want to pick up). Through augmenting the application with this locative information, it is possible to achieve the game experience depicted.

Player and game states are all stored on a centralised PhP/SQL server, making wireless communications a key enabling technology. All interactions between player and game resulted in at least one HTTP call being made to this server via GPRS communications. The key factors in its suitability of this data transfer for mobile pet applications were cost, speed and availability. Tests revealed that 5632 server requests of differing length were made during the second trial of the game. The mean cost of a single transaction was calculated to be 0.029GDP. In terms of speed, a transaction took between 5 and 30 seconds depending on interaction type (logging in and sending/receiving photos proved particularly slow). Availability for the medium includes all areas covered by mobile mast signals. Other options for communication might include Bluetooth, or WiFi, but not without significant changes to the game architecture.

5 USER TRIALS

The game was assessed through ethnographic studies of two groups and this is reported in the ACE paper [3]. The main findings are summarised here.

i) Player Social Engagement: The study paper describes mixed success. At times Gophers was successful in engaging players through pervasive features and other times, less so. The most common locations for play overall were still the typical, socially isolated locations that traditional gamers tend to favour: watching TV, in the bedroom.

Nevertheless, the presence of a gopher on a player's phone encouraged them to play at times they would not have been able to with non-mobile game; there was even evidence for players travelling to complete tasks. Furthermore, when players did pick up and drop gophers, this tended to be in a spatially common location, where players would meet at the same time and exchange gophers in a social meeting, for example, when sitting in the school common room. These examples highlight the enhancement offered by locative play.

Content collected during the trial showed that players were keen to exchange information with their pets, involving them in their daily routine. The ability to swap photos and words proved to be popular methods interacting with gophers, with players using these as methods to record their current context, to chat with gophers, or provide the gopher with content relevant to their mission.

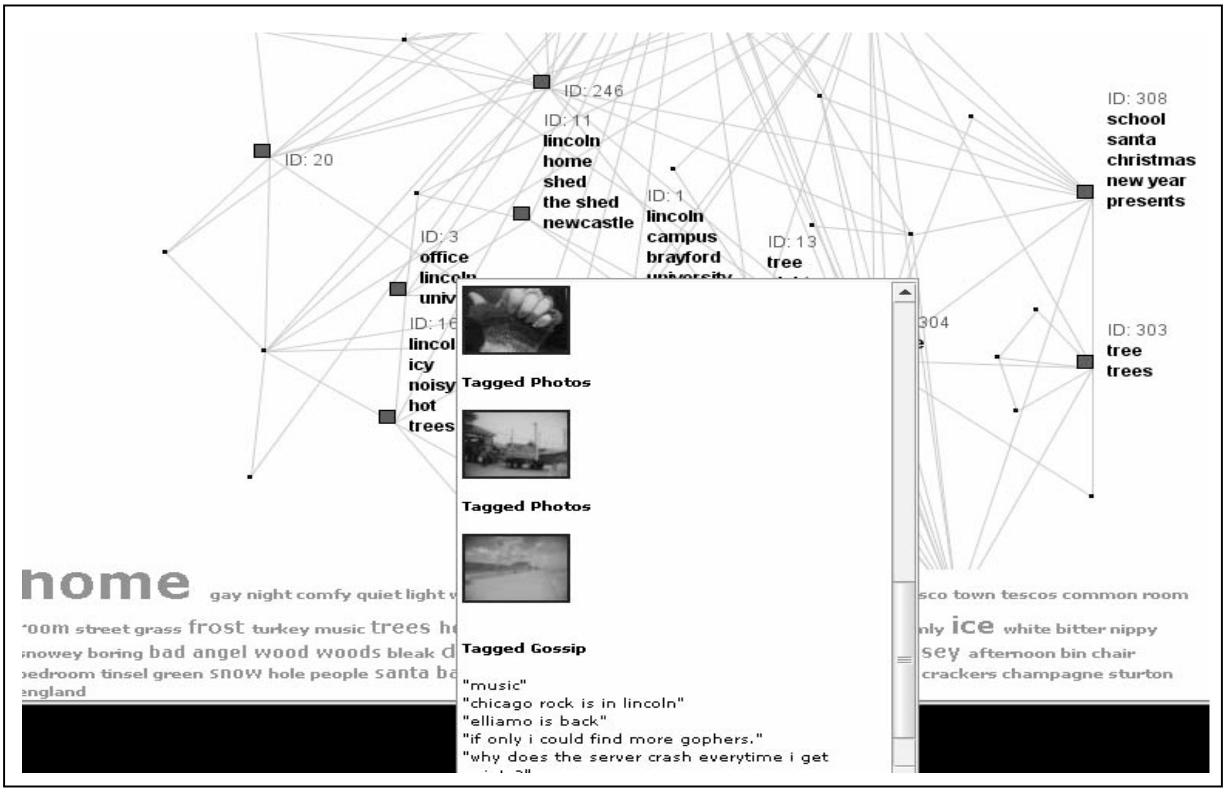
ii) Emotional Attachment and Mortality: Whereas most pets rely on a certain level 'affective blackmail' [9] to achieve an emotional bond between player and pet (for example, through linking a pet's development to the amount of attention it receives), Gophers uses the ongoing interest provided by the missions to similar effect. Comments left by some players showed a genuine emotional attachment to the creatures. In some ways, releasing the Gophers into the wild so that they could be picked up by other players, increased the bond with their owner. The fact that Gophers had a life independent of their owner and could visit other players before eventually returning to their owners phone had an *"absence make the heart grow fonder"* allure.

The importance of missions to users was further exemplified by the amount of time spent looking at blogs. The majority of respondents (79%), admitted to checking the blogs of their old gophers 'for fun', in addition to assessing mission progress. This demonstrates a continued interest in their pets, after they had finished directly interacting with them and further emphasises a bond between player and pet.

In early revisions of the game, gophers were removed from gameplay when their mission was completed. Although unintentional, this acted as a negative form of feedback and was another example where users showed attachment to gophers. This was deemed too harsh and future revisions gave positive encouragement for completion of tasks, through allowing players to assign a new task to the gopher once complete.

6 SITUATION AWARE

As players interact with Gophers, the dialog is used to build a connected graph that links spatial, temporal and semantic data. The graph consists of hierarchic labels, descriptions and images that players have submitted at each particular location. An example can be seen in *Figure 3*.



"music"
"chicago rock is in lincoln"
"elliamo is back"
"if only i could find more gophers."
"why does the server crash everytime i get

Figure 3. Example of a dynamic node graph

This continuously evolving collection provides a powerful link between the gophers, players and the real world. In the current iteration of Gophers, this information is simply used to return location relevant replies verbatim to the player during interactions with the pet.

The node graph shows a selection of cell locations and some of the data that has been exchanged with players at each site. The words underneath each ID are lists of words that have been guessed during the guessing game at that location, ordered with respect to frequency. The guessing game is designed to elicit words that are most indicative of the players' location and so these words may be useful in the future for creating context sensitive elements of play. Examining similarities between players (for example, whether many players submit the same word at the same location) and difference between players and also locations, may allow various categories of location to be created. In addition, it may also be possible to extract categories of player type (for example, those players who share a similar interest may label a certain location differently to those who do not share that interest). This could be used to enhance some of the social aspects of play. Photos and gossip provide less constrained user data and so are more difficult to incorporate into such an analysis. It may however be possible to show a player gossip or photos submitted by other players and examine the first player's response. This way, instead of the game spotting patterns in the data itself, the game would spot patterns in players' responses to the data. Again, this could be used to further enhance the pervasive, locative and social aspects of the play, further improving the virtual pet experience.

7 FUTURE WORK

An extension to this work is aimed at using the locative/semantic data held in the node-graph to create more situation sensitive behaviours in the pets. Through analysis of the data, it is expected to be possible to infer situational context and perhaps even behaviour of a gopher's owner. Availability of this information is particularly beneficial to pets, as it allows them to interact with their owner in a way which is more meaningful and appropriate to their current location, situation and owners behaviour. Essentially, this will allow gophers to become *contextually aware*, allowing them to know when to be attentive, passive and what behaviours are socially appropriate; a feature which is common to real life pet behaviour.

For example, a gopher may recognise they have been travelling along a familiar route at a fairly constant speed. In the past in these circumstances they may have received little attention from their owner. The gopher interprets this information and recognises that their owner is currently busy (travelling somewhere) and is therefore unlikely to want to be disturbed. As a result, the Gopher decides to sleep for the remainder of the journey.

Alternatively, a gopher could take a more active role. It may notice that is around 1:00pm and from previous experience knows that this is the time their owner normally likes to take a lunch break. Additionally, it recognises they are alone but in close proximity to a café frequented by people whom their owner often lunches with (a place often previously labelled with the concept 'food'). The Gopher may then indicate that it wants to go to the café, subtly orchestrating their owner's life (in a manner similar to that employed by real pets).

Nevertheless, acquiring data to meet these scenarios is a non-trivial task and requires 'meaning' to be extracted from source data. The filtering out of contextually relevant information is the focus of much on-going research.

8 CONCLUSION

Gophers aimed to provide an experimental concept study in the field of virtual pets. Through the implementation of the research study, we have demonstrated it is possible to apply pervasive computing theory to a virtual pet game. The trial itself yielded mixed results; the pervasive nature of the pets successfully promoted rich new interaction methods between player and pet, which move beyond those offered by existing virtual pet experiences. Additionally, narratives based upon real-world tasks, generated a strong bond between player and the gophers they created over a sustained period of time. These findings suggest there is much to be gained through extending the abilities of virtual the pets to offer social networking experiences, and our future work is directed in this area.

ACKNOWLEDGEMENTS

This work was funded in part by the PASION EU research consortium and by a studentship from the University of Lincoln. We gratefully acknowledge the assistance of Dr Steve Benford and Dr Adam Drozd from the University of Nottingham, for their provision of the Hitchers framework and for the loan of mobile phones. We also thank Dr Stacy Rowland and the VIth form students at Christ's Hospital School, Lincoln for their enthusiastic participation in the user trials.

REFERENCES

[1] S. Bjork, J. Falk, R. Hansson and P. Ljungstrand, 'Pirates! Using the Physical World as a Game Board', In *Proceedings of Interact 2001 IFIP TC.13 Conference on Human-Computer Interaction*, Tokyo, Japan, 2001.

[2] L. Brunnberg and O. Juhlin, 'Movement and spatiality in a gaming situation - boosting mobile computer games with the highway experience', In *Interact '03: Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction*, Zürich Switzerland, 2003.

[3] S. Casey, B. Kirman and D. Rowland, 'The Gopher game: A social, mobile, locative game with user generated content', To appear in *ACE'07*, Salzburg, Austria, 2007.

[4] S. Clarke and C. Driver, 'Context-aware trails', In *Computer*, 37(8):97-99, 2004.

[5] A. Drozd, S. Benford, N. Tandavanitj, M. Wright, and A. Chamberlain, 'Hitchers: Designing for cellular positioning', In *Ubicomp*, pages 279–296, 2006.

[6] M. Flintham, R. Anastasi, S. Benford, T. Hemmings, A. Crabtree, C. Greenalgh, T. Rodden, N. Tandavanitj, M. Adams and J. Row-Farr, 'Where on-line meets on-the-streets: Experiences with mobile mixed reality games', In *Proceedings Of The 2003 CHI Conference On Human Factors In Computing Systems*, Florida, USA, 2003.

[7] GoPets.
http://www.gopetslive.com/.

[8] J. Hightower, A. LaMarca, and I. E. Smith, 'Practical lessons from place lab', *IEEE Pervasive Computing*, 5(3):32–39, 2006.

[9] F. Kaplan, 'Free creatures: the role of uselessness in the design of artificial pets', In *Proceedings of the 1st Edutainment Robotics Workshop*, 2000.

[10] G. Kortuem, J. Schneider, J. Suruda, S. Fickas and Z. Segall, 'When cyborgs meet: building communities of cooperative wearable agents', In *Proceedings of the Third IEEE International Symposium on Wearable Computers*, San Francisco, CA, USA, 1999.

[11] S. Kriglstein and G. Wallner, 'HOMIE: an artificial companion for elderly people', In *CHI '05: extended abstracts on Human factors in computing systems*, pages 2094-2098, New York, NY, USA, 2005.

[12] Mopyfish.
http://www.mopyfish.net/.

[13] Nintendogs.
http://www.nintendogs.co.uk/.

[14] Pocketgamer.
http://www.pocketgamer.co.uk/.

[15] Tamagotchi.
http://www.tamagotchi.com/.

[16] TxTBook.
http://www.thetxtbk.com/.

[17] Vmigo.
http://www.vmigo.com/.

[18] You-Who.
http://www.age0.com/you-who/.

Virtual pets and electronic companions – an agenda for inter-disciplinary research

Shaun Lawson[1] and **Thomas Chesney**[2]

Abstract. The virtual pet and electronic companion genres of computer games and computing devices respectively are each held up as a success in terms of consumer appeal. Such products have had phenomenal recent success in terms of raw sales figures: Nintendogs alone for instance has in excess of 7 million sales worldwide after less than 2 years of it becoming available. Nintendogs builds on previous successes such as the Tamagotchi, Furby and the long-running Petz series. This paper discusses the recent commercial success of such games and devices as well as the surprising relative lack of interest by the research community in the same topic. We argue that virtual pets are a topic worthy of scientific investigation, present a number of research questions, and lay out an inter-disciplinary research agenda for addressing these questions.

1 INTRODUCTION

The virtual pet and electronic companion genres of computer games and computing devices respectively are examples of very successful commercial technological products. As examples of virtual (screen-based) pets we include software games such as Catz, Dogz, MoPets and Nintendogs, whilst as examples of electronic (embodied) companions we include devices such as Tamagotchis, Furbys (see Figure 1) and Sony Aibos. For shortness, unless otherwise stated, throughout the rest of this paper we will refer to both genres collectively as virtual pets[3]. A summary of worldwide and UK sales of a number of virtual pet products is given in Table 1.

As can be seen, millions of consumers worldwide have purchased these products, played with them, interacted with them, invested time in looking after them, and perhaps even become emotionally attached to them. Despite this huge financial and emotional investment by consumers, and an ongoing development and marketing investment by industry (new titles are appearing almost daily), academic interest in such products is virtually nil. This is surprising given the abundant activity in closely related fields such as social robotics [1], emotionally aware and affective computing [2], and the many diverse aspects of believable graphical agents [3][4].

Table 1. Estimated worldwide sales figures of a selection of commercial virtual pets and electronic companions

Virtual Pet or Companion	Manufacturer	Estimated Global Sales
Tamagotchi	Bandai (Japan)	>50,000,000
Furby	Hasbro (USA)	> 30,000,000
Nintendogs	Nintendo (Japan)	7,000,000
Petz series	various	2,000,000
Aibo	Sony (Japan)	<200,000

We believe that there is a set of fundamental, unanswered, questions centered on the commercial interest in virtual pets which is has hitherto been overlooked. Sales figures and the very fact that many virtual pet products are squarely aimed at children and younger people indicates to us that more attention should be paid to the effects, both positive and otherwise, that such products have on their users, owners and players. For instance, it is not known what benefits, companionship, or enjoyment that users gain from owning a virtual pet. Compared to the scarcity of published work in this area, there is, in stark contrast, an abundance of literature examining the benefits of owning *real* pets. For instance, studies have looked at how pets can positively effect people as they get older (e.g. [5][6]), how pets can alter the interaction between people when they meet for the first time (e.g. [7]), how they can help overcome the death of a close relative [8] and how owning a pet can be of benefit in a child's development [9]. It follows, and indeed it is often claimed by toy manufacturers (see below), that a virtual pet could possibly deliver some of these benefits, though to our knowledge, no studies have actually examined this.

Figure 1. An Emoto-Tronic version Furby from 2006 captured in the wild

This paper discusses the history of virtual pets and gives an outline of taxonomy of the existing products which are, or have been, commercially available. We discuss the interactions that virtual pets afford and the, often as yet anecdotal, evidence for peoples attachment and emotional involvement with them. We then review existing research efforts that have made some contribution to understanding the use and interest in virtual pets. We go on to present a research agenda to investigate the main questions surrounding virtual pet use and draw some conclusions. In particular we emphasize the need for inter-disciplinary research into the area of virtual pets; for instance we advocate an exploration of how analysis of the benefits of real pet ownership can be used to explore the benefits and effects of virtual pet ownership.

2 VIRTUAL PETS, A SHORT HISTORY

A virtual pet is an artificial companion that, typically, attempts to stimulate human-computer interaction by making the user feel

[1] Social Computing Research Centre, University of Lincoln, UK. e-mail: slawson@lincoln.ac.uk

[2] Nottingham University Business School (NUBS). e-mail: Thomas.Chesney@nottingham.ac.uk

[3] The term 'cyberpets' also gained popularity when Tamgotchis first appeared though this phrase now seems dated and unpopular.

responsible for it. Many virtual pets, visually at least, are often replicas of real animals such as cats and dogs though they can also frequently be abstract creatures such as Furbys (sometimes described as a cross between a hamster and a bird) and the wide range of fantasy creatures available in NeoPets.

Although the Petz series of PC games (see below) is acknowledged as being the first commercially available virtual pet product, the genre didn't really gain worldwide popularity until the late 1990s when Japanese toy manufacturer Bandai released the Tamagotchi electronic device. The original Tamagotchi was released in 1996 in Japan and in 1997 in Europe and the USA. About the size of a key ring, the device had a small black and white screen, three buttons, a speaker, a motion sensor and a microphone. The Tamagotchi creature itself appeared on the screen and had a varying appearance depending on its age. Users could feed, clean and play with their Tamagotchi, call it via the microphone and chase away predators by shaking the unit. The pet would evolve over time and would eventually either die or fly away. The Tamgotchi spawned the original fascination for 'cyberpets' as they became known and the product developed a small degree of notoriety with regard to their alleged demands for attention and overuse by school children to whom it was clearly targeted. It certainly became clear that many users became attached to their pet, with many actually mourning its death[4]. The Tamagotchi has, since its inception, gone through numerous model updates and is still, at the time of writing, being produced by Bandai– the latest, the Tamagotchi Connection Version 4, was released early in 2007.

Since the Tamagotchi there have been numerous copycat products, mostly aimed at children, and all appearing on sale in high street stores for a few tens of UK pounds. A commonality is that many such derivative games, unlike the Tamagotchi, feature simulations of real animals – mainly cats and dogs. Recent examples of these include Anipalz and Password Puppies.

Although the Tamagotchi was delivered in a format that required users to buy a complete electronic device it later became available as software that would run on gaming consoles such the Nintendo GameBoy. In a similar fashion, the virtual pet software in the Petz series has always required installation on a computer such as a Windows PC, or, much more recently the Nintendo DS handheld games console. The Petz series, which includes the games Catz and Dogz, uses animated instances of familiar pet animals as the user's virtual pet. Players, or users, or owners, can choose their pet at the pet shop, look after their health, teach them tricks and so on – exactly as one would with a real pet. The Petz series in this way actually feels more educational when compared with other products. Indeed, Ubisoft's Petz Executive Producer Tony Van, when interviewed about the recent release of the Petz series on the Nintendo DS, stated that:

> *"one value I always suggest is the player learning how to best take care of their pet, which translates to its use in the real world. This is valuable to both kids and adults, and if it results in one less abused animal in this world, that makes my job even more rewarding"*[5]

Van's claim that by playing a computer game which involves caring for a virtual pet, people (both children *and* adults) are able to train themselves to care for, and improve the welfare of, real animals, whilst presumably well-intended is currently unfounded.

Nintendogs, released by Japanese games company Nintendo in 2005 for its handheld games console the Nintendo DS, is one of the fastest selling games titles of all time and has received consistently high reviews by a video-gaming press usually dominated by adult oriented first person shooter and action games enthusiasts. The previous apparent quirkiness and child-only appeal of virtual pet games has largely been dispelled single-handedly by the game which is widely acclaimed as being radically mould-breaking. In actual fact, the title offers little more in terms of game-play than the game Dogz which precedes it by nearly a decade – however the overall package is highly polished and well-marketed. The Nintendo DS has two full colour screens – one which is touch-sensitive - which shows an animated puppy which owners must feed, water, walk, wash, groom, play with and train. The Nintendogs themselves are animated implementations of real breeds of dog (such as Labradors and Chihuahuas) and move in highly believable motion patterns. Nintendogs is unique in two aspects: firstly, users can actually touch their screen based pet through the use of the DS's touch screen, and secondly users may exploit the wireless network capability of the DS to exchange puppies with each other and allow Nintendogs to visit another device and play with each other.

The typical characteristics of Nintendogs owners are unclear – although it is easy to assume that the game is aimed towards children some of Nintendo's marketing for the game has clearly been adult oriented. Additionally, Nintendo recently claimed ([10] that 22% of Nintendogs owners are female compared to only 5% of players of their other early success for the DS platform, Mario Kart DS (a driving game). The games industry still appears to view female gamers as a largely hitherto untapped demographic and whilst early explicit attempts to exploit this potential market [11] were largely seen as unsuccessful, many recent games such as the Sims and Nintendogs have shown that certain styles of game-play (for instance, ones that encompass creativity and emotional attachment as well as, or even instead of, tangible goals) are indeed very appealing to female buyers. It is often assumed that the popularity of these games with female players has been accidental – however this view does seem naïve if one considers the careful, often very conservative, but ultimately successful strategies of the two games' publishers Electronic Arts (EA) and Nintendo and the burgeoning academic debate that is informing gender and gaming (e.g. [11],[12],[13]).

A related concern to "tapping into the female games market" in the games industry is exploiting the so-called 'casual-gamer' market. Casual-gamers are individuals who typically have no interest in the mainstream gaming titles which have to run on games consoles or high powered PCs but instead *do* have an interest in playing the occasional game whilst traveling to work on the train or bus or whilst relaxing in front of the TV or waiting for an appointment. Casual gamers are often generalized to be older people and female, whilst mobile phones and, to a lesser extent, mobile games consoles are viewed as the main computing platform which supports casual gaming (at least in public places). Example successful titles such as Dr Kawashima's Brain Training for the Nintendo DS support some of these presumptions. It is also not surprising that a number of high profile mobile-phone based virtual pet titles, by well known media corporations, seem to have been recently squarely aimed at the casual gaming market. The Walt Disney Internet Group (WDIG) recently have teamed up with original cyberpet innovators Bandai to produce My Little Dogs ~Kawaii Dogs, Sony BMG and Floodgate have released Mo-Pets, whilst I-Play (in the UK) have produced the My-Dog game. My Little Dogs in particular appears to take inspiration from the success of Nintendogs which in itself could also be argued is an example of the casual game genre. However, just as owning a real dog probably shouldn't be viewed as a casual commitment, many Nintendog owners appear to have abandoned the game due to aspects of the game-play which are very tying. The following are two excerpts taken from a thread[6] entitled "soft punishments" in the forum on the Game+Girl=Advance website:

[4] see for instance: http://starbulletin.com/97/06/18/features/rant.html and the virtual pet cemeteries at http://www.geocities.com/Heartland /Plains/2188/ and http://www.d-3.com/deadpet/

[5] Interview on Gamasutra website at http://www.gamasutra.com/php-bin/news_index.php?story=11736

[6] http://www.gamegirladvance.com/archives/2006/05/25/ soft_punishments.html

> *"I ultimately rid myself of … Nintendogs for this very reason. My maintenance -- things I had to do every time I started playing that took upwards of 15 minutes to complete -- started to drain my desire to do anything else. After a while I was *JUST* doing the maintenance"*
>
> *"the (nearly) mandatory maintenance of a Nintendogs session ultimately drove me to quit playing, esp after I got Mario Kart DS. Mario Kart can be played for a few minutes and put down."*

The second comment is particularly revealing in that it compares Nintendogs very unfavorably with a true casual game which can simply be picked up and "put down" with no emotional implications. The implication therefore is that virtual pets are not true examples of casual games.

A final, somewhat curious, example of the use of a computing platform to support screen based interaction with virtual pets is the idea of web-based systems. The US-based NeoPets is perhaps the best known example of this and claims, on its website[7], to have 70 million users, or owners of the range of fantasy creatures that are available. A related example is the much more recent South Korean based GoPets which, although primarily web-based in its functionality, supports downloading of virtual pets (cats and dogs) to a user's desktop where they then 'live'. Web-based virtual pet systems often seem to promote social networking between owners as opposed to looking after pets as the main thread of the game-play. This is not too surprising when one considers that it is unlikely that in real life a person would keep a true pet at a remote location. Some sites adopt out pets so that they can be included on other websites, whilst others allow breeding to create newer, often 'showier' creatures. Some sites even allow users to breed a pet for combat against other players.

A final category of virtual pet encompasses those products which are embodied electronic and often robotic, or animatronic, devices. One could loosely argue that a Tamagotchi is an example of this *embodied* genre but since the pet itself appears on the screen of such devices then we prefer to categorize these as screen-based. Sony's robotic dog Aibo [14] is perhaps the most well known of all embodied virtual pets. Significantly however, Sony announced a termination of all commercial and research activity in robotics early in 2006 [15] leaving a gap in the market for home entertainment robots which demonstrate sophisticated social behaviours and cognitive abilities. In terms of size and shape, Aibo was made to look like a real puppy. It would act like a puppy by exploring its environment (somewhat slower than a real dog), wanting to play, getting tired, angry and excited, needing sustenance (a battery re-charge), and could be trained by owners to do tricks. Aibo's price tag meant it was not targeted at children although similar, much cheaper systems inspired by it, such as the i-Cybie and Wow Wee's RoboPet, have since appeared in the marketplace. Although it was undoubtedly technological impressive and undeniably attention-grabbing in its curious pseudo-dog-like behaviour, it is not obvious the impact that the device has had on the marketplace as virtual pet – indeed global sales figures for the device are estimated to be less than 200,000 during its 6 year production run.

An embodied animatronic device with perhaps more impact is the Furby from US-based Tiger Electronics – certainly in terms of sales figures the Furby far outpaces the Aibo with global sales worldwide being in excess of 25 million in its first 12 months of availability alone in 1998. Indeed the impact, and notoriety, of the Furby is probably only matched by the Tamagotchi. Furbys, despite still costing only a few tens of UK pounds, however outperform Tamagotchis in interactive abilities – they are able to speak, move (their eyes, ears, mouth and feet) and even learn and repeat short spoken phrases. It is this last ability which caused the Furby notoriety in that it was perceived possible to use them as covert bugging devices[8]. The Furby is still available today, having gone through a number of face-lifts, and its popularity, although much weaker than when first released, appears stable with many adult users publicly discussing their interest in the devices. For instance, since its creation in 1999, the Yahoo group Adult Lovers of Furby (ALOF)[9] has had over 85,000 posts, with each week seeing over 100 new posts by its current members.

Figure 2. A Hasbro "TJ Bearytales" animatronic story-telling companion with his owner.

Whereas Furbys have seemingly proved popular with adults and younger people alike, a final genre of an embodied virtual pet or companion worthy of mention and which is aimed directly at children is the talking 'educational' toy. Included in this genre are the, now over 20-year old, Teddy Ruxpin teddy bear and, the much more recent, TJ Bearytales (see Figure 2), also a teddy bear, from US-based toy giant Hasbro. The Hasbro product is particularly sophisticated and is able to recite stories (from a cartridge) whilst synchronously moving its animated mouth, ears, eyes and nose and gesturing arms. These products have come under some criticism in that their reason for being seems to be to absolve parents' of their responsibility to read to their children.

3 SERIOUS STUDIES OF VIRTUAL PETS

Very little literature has appeared which describes any rigorous scientific investigations into peoples' use of virtual pets.

Some literature on the design of virtual pets has appeared (for example: see [16][17][18]) and, whilst human-virtual pet interaction has been described in the popular press (e.g. [19][20]), very few academic papers have examined the benefits of interacting with one. New media commentators such as Turkle in [21] have called for answers to the question of how we should interact with such devices claiming this is an important, even 'urgent', issue. Most recently in Isbister in [22] attempts to rationalise peoples' motivations in engaging with a virtual pet and suggests that the objective is to enjoy the pet's development as well as its moments of both connection and resistance to the player. In this way she identifies that virtual pets are relatively unique as autonomous agents since they evoke a high degree of time and emotional investment. Subrahmanyam et al in [23] discuss the shift from real life to simulation in the context of virtual pets but merely conclude, like Turkle, that systematic research is needed to assess the impact of such technology.

[7] http://www.neopets.com/

[8] CNN story "Furby a threat to national security?" January 13, 1999, online at http://www.cnn.com/US/9901/13/nsa.furby.ban.01/

[9] see http://groups.yahoo.com/group/ALOF/

In our own work, we have taken the approach of trying to assess whether people gain similar benefits in terms of companionship with virtual pets as they do with real pets. In [24] we show that people do indeed register feelings of companionship when trying to quantify how they feel about interactions with their pet. In order to do this, we deployed a well-known questionnaire-based measure to determine companionship from animals [25]. Given the marketing stance adopted by virtual pet manufacturers who clearly target younger people as the main consumers of their products, we have also tested the hypothesis that younger virtual pet owners will experience closer companionship with their virtual pet than older owners and show this to be true [26].

Aside from studies of commercially available virtual pets, only very few researchers have developed entertainment oriented software systems [27][28] or embodied robots [29] which are based on genuine animal behavioural (or ethological) studies and allow for social interactions between humans and autonomous non-human synthetic creatures (particularly dogs).

4 UNANSWERED QUESTIONS

Despite the huge commercial success of such products, fundamental, unanswered, questions remain as to the benefits, companionship, or enjoyment that users gain from owning a virtual pet. It is even unclear, for instance, as to whether different people play, or interact, with virtual pets for differing reasons. Additionally, there are many recent instances of virtual pet manufacturers claiming that the ownership of virtual pets in some way provides either useful training prior to, or a long-term substitute for, the ownership of a real animal. Many claims surrounded virtual pets are, at present, unfounded, and little, if any, academic work has examined such claims. We believe the questions given in Table 2 are the most important ones which arise from virtual pet use and that all such questions are currently unsolved.

Table 2. Unsolved issues surrounding virtual pets

Q	Research question
1	Why, fundamentally, do people buy and interact with different types of virtual pet – is it to get the same, some of same, or completely different benefits as they get from real pets?
2	Do people of differing age groups, backgrounds and gender view virtual pets in different ways and get different benefits (if any) from them?
3	Are virtual pets merely casual games that we treat in the same way as Tetris and Mario Kart?
4	Are virtual pets merely social lubricants and used to facilitate new interactions with other people and/or strengthen bonds between existing friends and peers?
5	What attributes of virtual pets are key to their commercial success? Is there an ultimate virtual pet requirements specification?
6	What educational benefits, if any, are there from interacting with virtual pets?
7	What health and ethical issues arise from virtual pet use – particularly when considering young and much older people?

In our own work we have already begun to adopt a multi-disciplinary approach to the understanding of virtual pets and companionship [24][26]. In particular we have looked at anthrozoological (human/animal interaction) studies which, for a number of years, have attempted to quantify the benefits humans receive from interacting with real pets and companion animals. We believe generally that it is only through investigations conducted jointly between computer scientists, companion animal scientists, social scientists and psychologists can we begin to understand some of the issues arising from virtual pet use. In the meantime, industry continues to release new titles and new innovations which we also need to keep pace with. As the designers of all virtual pet products are no doubt aware, an accepted consensus within anthrozoologic research is the quantifiable positive effects of human-animal relationships. Accordingly, Wilson [30] coined the term *biophilia* as “the connections that human beings … seek with the rest of life”, and argued that such cravings are determined by a biological need. However, to-date no link has been explored between such socio-biological theories and human interactions with artificial systems.

5 CONCLUSIONS

Sustained consumer interest in virtual pets and electronic companions seems to confirm the widespread appeal of interacting with artificial, albeit rather basic, representations of pet animals.

In this paper we have outlined a history/taxonomy of commercial virtual pets and posed what we believe are the main research questions currently surrounding such products. We are of the firm belief that much more research in this area is needed in order to better understand the phenomenal commercial success of virtual pets and that such research demands an inter-disciplinary approach.

REFERENCES

[1] Breazeal, C. *Designing Sociable Robots*. Cambridge MA, MIT Press. 2002.

[2] Picard, R. W. and J. Klein. Computers that recognise and respond to user emotion: Theoretical and practical implications. *Interacting with Computers* 14 (2), 141–169 (2002).

[3] Thomas, F. and Johnston, O. *Disney Animation: The Illusion of Life*. Abbeville Press. 1981.

[4] Maes, P. Artificial life meets entertainment: lifelike autonomous agents. *Communications of the ACM*. Volume 38 (11), pp.108-114 (1995).

[5] Mahalski, P.A. Jones, R. & Maxwell, G.M. The value of cat ownership to elderly women living alone. *International Journal of Aging Human Development* 27(4) pp. 249-260 (1988).

[6] Garrity, T.F., Stallones, L., Marx, M.B. & Johnson, T.P. Pet ownership and attachment as supportive factors in the health of the elderly. *Anthrozoos* 3(1) pp. 35-44 (1989).

[7] Wells, D.L. The facilitation of social interactions by domestic dogs. *Anthrozoos* 17, 340-352 (2004).

[8] Adkins, S.L. & Rajecki, D.W. Pets' roles in parents' bereavement. *Anthrozoos* 12(1) pp. 33041 (1999).

[9] Pattnalk, J. On behalf of their animal friends. Childhood Education Winter 2004/2005 pp. 95-100. 2004.

[10] Fils-Aime, R. Nintendo keynote speech at Montreal International Game Summit, November 2006.

[11] Cassell, J. and Jenkins, H. From Barbie to Mortal Kombat: Gender and Computer Games. MIT Press, London (1998).

[12] Bryce, J., & Rutter, J. The gendering of computer gaming: Experience and space. In S. Fleming & I. Jones (Eds.), Leisure cultures: Investigations in sport, media and technology (pp. 3-22). University of Brighton, Eastbourne, Leisure Studies Association (2003).

[13] Kafai, Heeter, Denner & Sun, Eds. Beyond Barbie & Mortal Kombat, New Perspectives on Gender and Computer Games. Forthcoming from MIT Press (2007).

[14] Pransky, J. AIBO - the No. 1 selling service robot. *Industrial Robot*, 28(1):24–26, 2001.

[15] Sony (2006), Q3 FY2005 Sony Group Earnings Announcement, 26 January 2006, accessed 17/04/06, http://www.sony.net/SonyInfo/IR/info/presen/05q3/qfhh7c000008adfe.html

[16] Kaplan, F. Free creatures: the role of uselessness in the design of artificial pets. Proceedings of the 1st Edutainment Robotics Workshop (2000).

[17] Breazeal, C. Function meets style: insights from emotion theory applied to HRI. *IEEE Transactions on Man, Cybernetics and Systems*. 34(2) pp. 187-194 (2004).

[18] Kusahara, M. The art of creating subjective reality: an analysis of Japanese digital pets. *Leonardo* 34(4) pp. 299-302 (2001).

[19] Hafnew, K. What do you mean, 'it's just like a real dog?'. New York Times May 25 2000.

[20] Herold, C. Buy a Puppy and Teach It Tricks, All Electronically. New York Times, August 27 p.7. 2005.

[21] Turkle S. Relational artefacts, children and elders: the complexities of cybercompanions: towards social mechanisms of android science. A COGSCI 2005 workshop, July 2005, Stresa, Italy, pp. 62-73. 2005.

[22] Isbister K. *Better Game Characters by Design: A Psychological Approach*. Morgan Kaufmann. 2006.

[23] Subrahmanyam, K. et al. The impact of computer use on children's and adolescents' development. Applied Developmental Psychology (22) pp. 7-30. 2001.

[24] Chesney, T. & Lawson, S. The Illusion of Love: Does a virtual pet provide the same company as a real one? *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*. (in press).

[25] Zasloff, R.L. Measuring attachment to companion animals: a dog is not a cat is not a bird. Applied Animal Behaviour Science 47 pp. 43-48. (1996).

[26] Lawson, S. and Chesney, T. The impact of owner age on companionship with virtual pets. To appear in Proc of 15th European Conference on Information Systems (ECIS 2007), St. Gallen, Switzerland, June 7-9 2007.

[27] Yoon, S-Y, Burke, R.C., Blumberg, B., Schneider, G.E., Interactive Training for Synthetic Characters, Proceedings of AAAI/IAAI 2000, pp.249-254 (2000).

[28] Tomlinson, B., and Blumberg, B. Social Behavior, Emotion and Learning in a Pack of Virtual Wolves. Proc of AAAI Fall Symposium, Nov 2-4, 2001.

[29] Fong, T., Nourbakhsh, I., and Dautenhahn K. A survey of socially interactive robots, *Robotics and Autonomous Systems* 42 pp.143–166 (2003).

[30] Wilson, E.O. *Biophilia*. Harvard University Press (1984).

Language, Speech & Gesture for Expressive Characters

Research into expressive characters, for example embodied conversational agents, is a growing field, while new work in human-robot interaction (HRI) has also focussed on issues of expressive behaviour. With recent developments in computer graphics, natural language engineering and speech processing, much of the technological platform for expressive characters — both graphical and robotic — is in place.

However, progress is hampered by the need to integrate work in various sub-fields of psychology, in natural language processing, speech and in computer graphics, carried out by many different groups in communities that do not always intersect. Other areas, such as integrating gesture and facial expression and affective state with language and speech, are less developed but vital to progress. The symposium aims to bring together psychologists, experts in natural language and speech technologies, researchers in embodied agents (graphical and robotic), affective computing and computer graphics and animation researchers.

Contribution we invited on the topics of speech and natural language processing for expressive characters, including: appropriate natural language processing architectures; natural language generation; dialogue systems and question answering, language and gesture coordination; language and facial expression coordination; language and action integration, emotional language; personality modelling, language and speech lip synchronisation and combination with facial expression; affect in speech synthesis and recognition. empirical studies of gesture and facial expression; frameworks for the specification and analysis of gesture and facial; expression for expressive characters; gesture and facial expression modelling and animation; evaluation of expressive characters

Ruth Aylett, John Glauert and Patrick Olivier (Symposium Chairs)

Programme committee: Ruth Aylett (Heriot-Watt University); Marc Cavazza (University of Teeside); Phil Heslop (Newcastle University); Suresh Manandhar (Unversity of York); Sally Jane Norman (Newcastle University); Patrick Olivier (Newcastle University); Catherine Pelachaud, (University of Paris 8); Thomas Rist (University of Applied Sciences Augsburg & DFKI); John Shearer (Newcastle University); John Glauert (University of East Anglia).

Coexpressivity of speech and gesture: Lessons for models for aligned speech and gesture production

Kirsten Bergmann and **Stefan Kopp**[1]

Abstract. When people combine language and gesture to convey their intended information, both modalities are characterized by an intriguing degree of coherence and consistency. For developing an account how speech and gesture are aligned to each other, one question of major importance is how meaning is distributed across the two channels. In this paper, we start from recent empirical findings indicating a flexible interaction between both systems and show that psycholinguistic models of speech and gesture production in literature cannot account for this interplay equally well. Based on a discussion of these theories as well as current computational approaches, we point out conclusions as to how a production model must be designed in order to simulate aligned, human-like multimodal behavior in virtual expressive agents.

1 Introduction

Humans intuitively combine language and spontaneous gesture to form multimodal utterances. In such utterances, words and gestures appear highly coordinated and closely intertwined - in other words *aligned* to each other by the human speaker. These alignments concern the meaning that the verbal and non-verbal behaviors convey, the form they take up in doing so, the manner in which they are performed, their relative temporal arrangement, as well as their coordinated organization in a phrasal structure of utterance. The results of these alignments are decisive of how meaning is constructed and communicated by the two modalities concertedly. In order to develop a systematic account of how speech and gesture align in multimodal communication three major questions have to be addressed. First, what kinds of meaning do people convey in concurrent speech and gesture to pursue their communicative intentions? Second, what form do speech and gesture take up to convey their meanings in context? This includes the question what particular gesture morphology speakers use to create, e.g., a coverbal depiction of spatial aspects of a referent? And third, how are speech and gesture organized across and within incrementally produced multimodal deliveries?

In this paper, we will try to shed light on these questions by addressing one pervasive phenomena, notably, the distribution of meaning across the two modalities. We review recent empirical findings concerning the coexpressivity of speech and gesture. We then discuss the implications these findings bear for both theoretical models of speech and gesture production as well as different approaches that were followed in modeling speech-gesture behavior in expressive agents. Conclusions will be drawn as to in which direction implementations of behavior generation in expressive agents should be heading.

2 Coexpressivity of speech and gesture

Across all kinds of description, researchers have found phenomena indicating that speech and gesture production are closely tied to one another. When produced along with speech in multimodal utterance, deictic gestures accompany referring verbal expressions or independently identify an object being referred to. In contrast, iconic gestures are thought to communicate mainly by virtue of their resemblance with what the speaker has in mind. They are intimately bound up with predication and are fully interpretable only in the context of simultaneous speech. These gestures were found equally likely to be either redundant with speech or to contribute information that is complementary to what is expressed in the verbal modality. More precisely, the relation between gestural and verbal content varies along a continuum of coexpressivity. That is, on the one extreme one can find gestures expressing more or less the same content as the verbal utterance they accompany, while the opposite extreme of the continuum are gestures encoding aspects of meaning that are not uttered verbally, in other words these gestures complement speech. Between those two extremes there are gestures with varying numbers of semantic features that are redundant or complementary to what is conveyed in speech.

Although each modality has its limitations with regard to the amount and kind of information it can readily express, several other factors seem to exert influence on the actual coexpressivity of speech and gesture. Based on a comparison of semantic features expressed by the gesture under consideration and its lexical affiliate, a recent empirical study of direction giving dialogs [3] has tested several hypotheses taking into account both individual meanings as well as the wider dialog context. First, refining previous results [5], the information status of objects has an impact on the choice of modality. Gestures with decreased redundancy and slightly increased complementarity are used to introduce entities, while it is exactly the other way around for gestures referring to evoked objects. Second, earlier findings [2] could be supported indicating a significant correlation between verbal encoding problems and the meaning distribution between speech and gesture. In detail, discourse markers and disfluencies result in a lower proportion of redundant and more complementary semantic features in the particular gestures. Third, communicative goals as they occur in route directions [9] are significantly influential. People tend to make use of redundant gestures when describing actions (e.g. reorientation, locomotion), while the number of complementary semantic features is increased when describing landmarks. When referring to landmarks with spatial orientation, however, speech-accompanying gestures contain less complementary aspects of meaning. Finally, no significant influence could be found for feedback signals of the addressee. Neither positive feedback nor

[1] Bielefeld University, P.O. Box 100131, D-33501 Bielefeld, Germany, email: {kbergman, skopp}@techfak.uni-bielefeld.de

interposed questions have an impact on the distribution of meaning.

The abovementioned findings demonstrate that the packaging and assignment of semantic features to one modality or the other is far from being fixed. Instead, there seems to be a flexible interaction between both systems (as proposed in [12]). The following statements summarize what is currently known about the interplay of the two systems:

(1) The information status of the referent seems to be influential in the sense that the introduction of new entities goes along with less redundancy and an increased number of complementary semantic features in gesturing. In different dialog acts, however, the influence of this has been less significant.

(2) Problems with verbal encoding result in more complementary and less redundant semantic features in gestures. People seem to compensate for these problems by enriching gesture with proportionately more information.

(3) The coexpressivity of speech and gesture is influenced by the communicative goals at hand. Instructions are accompanied by gestures with increased redundancy, while gestures describing landmarks are often characterized by more complementary aspects of meaning.

These findings indicate that the interplay between speech and gesture in forming multimodal utterances is complex and depends on a number of contextual aspects. This raises the question how we can, and to what extent we need to, account for this interplay when modeling natural multimodal behavior for expressive agents.

3 Psycholinguistic models of speech and gesture production

Starting with psycholinguistic models, several models have been put forward that, however, cannot account for the aforementioned findings equally well. Hadar and Butterworth's model [10] assumes that visual imagery becomes activated pre-verbally, while a set of semantic features to be realized linguistically is constructed during a separate step of conceptual processing. The visual image then translates into semantic features that can facilitate word-finding. Since the main focus of Hadar and Butterworth's model concerns the question how lexical retrieval is facilitated by visual imagery, two different kinds of facilitation are proposed. First, pre-verbal gestures serve to focus conceptual processing to a set of features. Additionally, they serve to refocus when semantic selection fails. Second, a post-semantic route provides for the case of phonological selection failure. Thus, aspect (2) is covered adequately by the model. However, the factors (1) and (3) are not addressed equally well. Since this model does not comprise a discourse history, it is not able to account for the information status' influence. Moreover, since the interplay between linguistic and imaginistic processes is realized for the linguistic unit of a word, the notion of underlying communicative goals cannot be considered in a sufficient way.

Krauss et al. [17] propose a model that shares the assumptions that an important function of iconic gestures is to facilitate lexical retrieval, and that they are produced pre-linguistically on the basis of spatial/dynamic representations in working memory. In contrast to Hadar and Butterworth [10], the hearing of the lexical affiliate serves as a signal to terminate the gesture. The question of whether or not a certain semantic feature is communicated by gesture, either redundantly or complementarily, is regarded as a conscious decision of the speaker as part of her communicative intention. A focused part of the discourse record is represented in the speaker's working memory, which is the source of gestural content. Since the speaker has access to a discourse model represented in long term memory via working memory, the influence of an entity's information status on the content distribution is supposable (1). However, there is no interplay between verbal and nonverbal processing except that the completed gesture provides input to the phonological encoder in order to facilitate lexical retrieval. In consequence, the necessary interplay to account for the factors (2) and (3) cannot be elucidated.

Similar to the model proposed by Krauss et al. [17] is de Ruiter's Sketch Model [8]. Although de Ruiter assumes that gesture is a communicative device, in contrast to Krauss et al., both models are alike in that they assume that gestures are generated before the linguistic formulation process takes place, as well as that speech and gesture production are independent processes to a large extent. Thus, concerning factor (3) the same argument as for the model proposed by Krauss et al. holds for the Sketch Model. It is not able to account for the influence of communicative goals, either. Moreover, both models rely on Levelt's language production model [18], providing the discourse record from long term memory via working memory. Information status of entities could thus be handled (1). However, in contrast to Krauss et al., de Ruiter includes verbal encoding problems and their influence on gesture. Speech failure is recognized and ”could be compensated for by the transmission of a larger part of the communicative intention to the gesture modality” [8, p. 293]. Thus, aspect (2) is explainable in an adequate way.

Kita & Özyürek proposed the Interface Hypothesis, according to which ”gestures originate from an interface representation between speaking and spatial thinking” [13, p. 17]. Moreover, ”gestures [do] not only encode (non-linguistic) spatio-motoric properties of the referent, but also structure the information about the referent in the way that is relatively compatible with linguistic encoding possibilities” [13, p. 17]. Building on this hypothesis, they developed the model of speech and gesture production shown in Fig. 1. A *Communication Planner* generates ”communicative intentions” making a first rough decision on the information to be expressed and deciding which modalities should be involved. The Communication Planner has access to a discourse model. Thus, it can process information status in a way as to accommodate our observation (1). The specifications of intent are sent to an *Action Generator* and a *Message Generator*. The Action Generator generates a spatio-motoric plan for the gesture to be performed. It has access to the part of working memory where relevant spatial imagery is active now, action schemata based on features of imagined or real space. The Message Generator, taking into account the communicative goal and the discourse context, formulates a propositional preverbal message that is sent to the *Formulator*. Both generators constantly exchange information, which also involves transformations between the two informational formats. Additionally, the Message Generator receives feedback from the Formulator whether a proposition is readily verbalizable or not. These interactions between the three components are thought to go on until equilibrium is reached. Not until this point, verbal formulation starts and the spatio-motoric representation generated by the Action Generator is sent to motor control for execution. With the posited interactions between speech and gesture-specific processes, Kita & Özyürek's model seems to be able to explain the influences of verbal encoding problems (2) as well as the overall communicative goals (3) on the coexpressivity of speech and gesture.

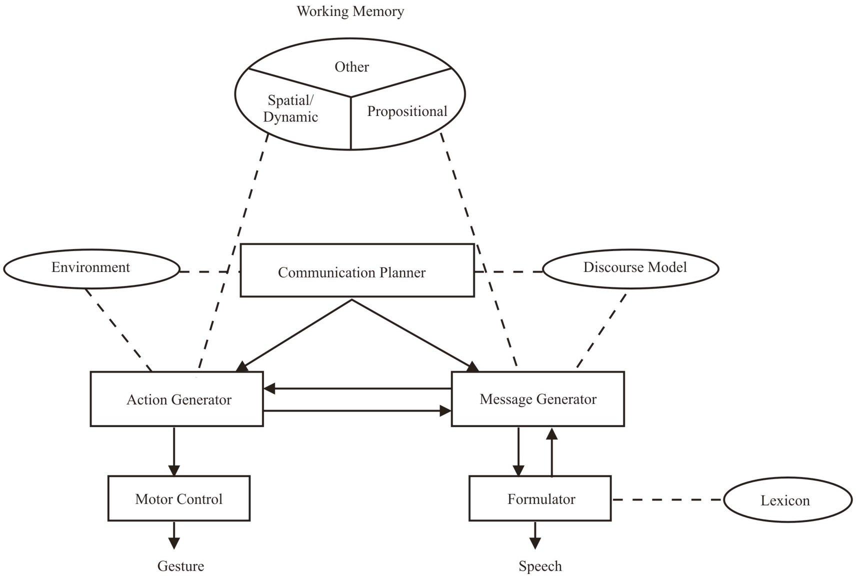


Figure 1. Speech and gesture production model proposed by Kita & Özyürek [13, p. 28]

4 Unfolding the Interface Hypothesis Model

The Interface Hypothesis proposed by Kita & Özyürek seems to be the only theory being able to integrate the different factors that were so far empirically found to influence the distribution of content across the modalities. With the goal in mind of building expressive multimodal characters on the basis of theoretically grounded models, now the question comes up if Kita and Özyürek's model is amenable to the requirements of modeling computational simulations. We consider a generation example to illustrate advantages as well as shortcomings of the model. Since most of the empirical results described in section 2 arised from direction giving dialogs, we try to realize a multimodal utterance from this domain.

Let us assume a direction giver wants to describe that a church is located left of the street a direction follower is to take. According to categories of communicative acts that can be found in route directions (cf. [3]), the corresponding communicative intention would be to describe a landmark, namely the church, with its spatial orientation viewed from the route perspective (first person walking the route), that is left. Among other decisions, the Communication Planner is to figure out which of the modalities should be involved. As described in [3], dialog acts belonging to the category ”landmark with spatial orientation” are accompanied by one or even two gestures in the majority of cases. Thus, rough specifications of the content to be communicated need to be passed to both the Action Generator and the Message Generator.

According to Kita & Özyürek's model, the Action Generator would now access relevant parts of visuo-spatial memory and extracts some salient visuo-spatial properties to be put into gesture, e.g. the church's SIZE ant its RELATIVE POSITION as seen from the correct perspective. At the same time, the Message Generator retrieves propositional information from working memory, for example, the semantic feature of ENTITY. The now following interaction between Action and Message Generator is decisive for the distribution of meaning across the modalities.

The underlying goal of the communicative act has to be taken into consideration to account for our factor (3). A description of landmarks with their spatial orientation tends to go with RELATIVE POSITION as the only redundant semantic feature in the majority of cases, while there occur only few complementary features. Assuming that this semantic feature has been retrieved from spatio-motoric memory only, we must conclude that the Action Generator shares this piece of information with the Message Generator or, in other words, triggers the related information in propositional memory.

Finding (1) tells us that, when introducing a new entity, additional features like SIZE and RELATIVE POSITION tend to be communicated as complementary information in the speech-accompanying gesture. This is not surprising as the mere existence of an entity is clearly not conveyable by gesture per se, which can only pick up and depict some of its spatial aspects. Notably, however, these features tend to not appear in speech in these cases. This suggests that either the Communication Planner or the Message Generator, both of which having access to discourse history information, decide to verbally ”focus” on the ENTITY aspect. That is, SIZE and RELATIVE POSITION, though found to be salient and hence being selected, must be assigned to the Action Generator.

Additionally, the abilities of both modalities to actually put semantic features into surface behavior (either verbal or gestural) is likely to exert a ”bottom-up” influence on this decision. We know that problems of verbal encodability can lead to complementary information in gesture (2), i.e. a transcoding and shifting of the particular semantic feature(s) from Message Generator to Action Generator. It stands to reason that, the other way around, the availability of linguistic resources for encoding meaning features may also lead to their verbalization. For example, we can assume that the Formulator disposes of English constructions for introducing an entity in the right tense, for referring to the church and the direction follower, *viz.* the addressee, and for denoting the spatial relation ”left-of” between the two. The availability and context-depending activation of a ”left-of” construction may imply verbal realization of the RELATIVE POSITION feature. Such an activation may be due to priming as in the alignment account of Pickering & Garrod [21], proposing that encountering an utterance that activates a particular representation makes it more likely that this representation will be used in a subsequently produced utterance. In contrast, the SIZE feature as transcoded from visuo-spatial information may be hard to express verbally in a way as provided for by open slots of the other selected constructions. This is signaled back to the Message Generator, and on to the Action Generator. Notably, the gesture path in Kita & Özyürek's model is posited to be devoid of such bottom-up effects. It is worth to be discussed whether there may be visuo-spatial properties of a referent that are hard to put into gesture, especially with regard to depictive gestures like drawing an 2D outline in space. This may give rise to a need for a feedback channel from Motor Control to the Action Generator (much like between Formulator and Message Generator). Altogether, these interaction processes could result in a distribution of content such that the gesture will convey the semantic features RELATIVE POSITION and SIZE, whereas the existence of the ENTITY and its RELATIVE POSITION will be assigned to verbalization.

The "microplanning" process within the Formulator now has to generate the verbal part of the utterance, something like "*There will be a church to your left.*". On the other hand, the semantic features to be communicated by gesture have to be mapped onto a set of morphological features, such as hand shape, hand orientation, hand position, and trajectory of movement. This mapping constitutes a significant step from the semantic representation of content towards the realization of a gesture, and it is of high relevance for a concrete simulation of multimodal behavior generation. Kita & Özyürek's model, as most psycholinguistic models, does not make any provisions as to how this mapping is achieved. De Ruiter has accounted for this mapping by introducing a dedicated "Gesture Planner", which could supplement Kita & Özyürek's Motor Control in this regard. Discussed in more detail in the following section, for now, we take a set of morphological gesture features for depicting the semantic features RELATIVE POSITION and SIZE of the church for granted.

Finally, in order to integrate speech and gesture into a single performance, the timing of speech and gesture has to be planned in detail. Typically the gestural stroke lines up in time with the specific linguistic segments that are coexpressive with it [19] (note the discussion in McNeill (2005) of a difference between the lexical affiliate of a gesture and its coexpressive speech). In our example, the gesture was derived from the semantic features RELATIVE POSITION and SIZE, as they apply to the church at hand, along with the communicative goal of describing a landmark with its spatial orientation. The expressive phase of the planned gesture thus has to be synchronized with the phrase "church to your left". For this reason, if one adopts the information-processing paradigm as in Kita & Özyürek's model, an extension becomes necessary. The results of the Gesture Planner and the Formulator cannot be passed on to the respective motor systems in separation. Instead, there must be anticipation and representation of the timing facts holding between the two modalities, allowing the modality-specific realization systems to prepare for this synchrony demands (see below).

5 Current state of computational approaches

The implementation of behavior generation in expressive agents is an appropriate means of testing and evaluating any model. A lot of multimodal expressive agents have been built, and generating convincing speech and gesture with them has proven to be a major challenge. Generally, one can demarcate three stages in this production process starting from a general communicate goal or intention of the agent: selecting the content that needs to be conveyed, planning the surface form of multimodal behaviors that can realize the communicative act, and finally rendering synchronized synthetic speech and gesture animations to actually perform these behaviors. Several implemented systems are described in the literature, each focusing on certain steps along this generation process.

5.1 Gesture planning

The BEAT system [6] employed a set of behavior generators to select and schedule conversational behaviors like hand gesture, head nod, or pitch accents. Relying on results from an empirical study [4], specialized generators supplemented the verbal description of actions as well as of rhematic objects features with gestures that were drawn from a lexicon.

The REA system [5] was able to successfully generate context-appropriate natural language and gesture in an embodied conversational agent. REA extended a natural language grammar formalism to handle constituents to be uttered in different modalities, and it was able to generate some iconic gestures, coordinated with the meaning of the linguistic expression they accompany and the discourse context within which they occur. However, whole gestures were lexicalized like words, selected using a lexical choice algorithm and incorporated directly into sentence planning. While this approach allows for context-dependent coordination with speech, it does not allow for the natural generative power of gestures, e.g., that can be formed to express new content. More recent work ([11]) has addressed the adjustment of expressive qualities of gestures, but sticked to the usage of lexicons of gesture templates that could be parameterised to certain extents.

The NUMACK system [14] (see figure 2) has tackled the formation of iconic gestures based on the assumption of systematic meaning-form mappings. This approach adopts the notion that iconic gestures communicate mainly in virtue of their resemblance to visuo-spatial properties of the entity they depict. To account for how iconic gestures are able to express meaning, this work linked gestures to their referents by assuming an intermediate level of abstraction and representation that accounts for a context-independent level of visuo-spatial meaning. Gesture generation was tackled based on the assumption that there are prevalent patterns in the ways the hands and arms are used to create iconic, gestural images of the salient, visual aspects of objects/events, and that such patterns may account for the ways human speakers derive novel gestures for objects they are describing for the first time. However, only sparse empirical evidence for this view could be obtained [15].

From a simulation point of view, this concept proved sufficient to create a range of direction giving gestures in the NUMACK system. Separable, qualitative image description features (like shape, orientation, or principal extent) were used to describe the meaningful geometric and spatial features of both a gestures' morphology and the entities to which a gesture can refer. A gesture planner (GP) was responsible for planning the morphology of a gesture by composing sets of one or more morphological features that convey the IDFs, which are part of the current communicative intention. Similar to a sentence planner for language, the GP drew upon a input specification of domain knowledge, plus a set of form feature entries that connect (conjunctions of) IDFs to (combinations of) morphological features. When receiving a set of IDFs as input the GP searches for all combinations of form feature entries that can realize them, and combines them by iteratively filling a morphology feature structure for a gesture. In result, the GP provides a set of gestures, each of which annotated with the IDFs it encodes. Based on this information, the sentence planner combined them with words in order to derive multimodal utterances. Note that the GP may also output an underspecified gesture if a morphological form feature does not meaningfully correspond to any of the disposed IDFs, i.e., it remains undefined by the selected patterns.

Figure 2. Numack [14] and Max [16] as examples for expressive virtual agents able to synthesize multimodal behavior

5.2 Multimodal synchronization

In other previous work, we have developed the virtual human Max (see figure 2) based on the Articulated Communicator Engine (ACE, for short) for behavior realization [16]. ACE allows to create virtual animated agents, and to synthesize for them multimodal utterances including speech, gesture, or facial expressions. Input descriptions are formulated in MURML, an XML language for succinctly defining multimodal utterances. The ACE production model aims at creating a human-like flow of continuous multimodal behavior. To this end, it tries to simulate the main mutual adaptations that appear to take place between speech and gesture, when humans try to achieve synchrony between the coexpressive elements in both modalities. An incremental process model allows for handling cross-modal interactions at different levels of an utterance, corresponding to decisive points in multimodal behavior generation. It is based on an empirically suggested segmentation hypothesis [19], that continuous speech and gesture are co-produced in successive segments, each expressing a single idea unit, together forming a hierarchical structures of overt gesture and speech.

ACE takes *chunks* of speech-gesture realization, as produced in trouble-free utterance, to be pairs of an intonation phrase and a coexpressive gesture phrase (see [3] for empirical evidence for this). Within each chunk, the coexpressivity of the gesture and a word or sub-phrase is evidenced by temporal synchrony between them, often accomplished by stretching single gesture phases or inserting dedicated hold phases in the flow of movement [7], [20], [19]. In order to simulate how humans strive to meet this synchrony constraint, ACE accounts for cross-modal adaptations that take place either within a chunk or, at a higher level, between two successive chunks. Within a chunk the synchrony between certain words and the stroke is mainly accomplished by the gestures' adapting to the timing of speech, which in turn runs mostly unaffected by gesture ("ballistically"). In producing a single chunk, the intonation phrase is therefore synthesized in advance. Information about absolute phoneme timings retrieved from a text-to-speech system (TTS) is used to set up timing constraints for co-verbal gestural or facial behaviors. Appropriate behaviors are then planned on-the-fly by means of procedural animation. For a dynamic gesture, to a post-stroke hold after a normally executed stroke phase or to additional repetitions of the stroke.

The synchrony between speech and gesture in the forthcoming chunk is anticipated at the boundary between two successive chunks. First, the onset of the gesture phrase co-varies with the position of the nucleus and, secondly, the onset of the intonation phrase co-varies with the stroke onset ([7], [20], [19]. In consequence, movement between two strokes depends on the timing of the successive strokes and may range from the adoption of intermediate rest positions to direct transitional movements (so-called co-articulation effects). Likewise, the duration of the silent pause between two intonation phrases may vary according to the required duration of the preparation for the next gesture. These adaptation effects are simulated during a phase in which the next chunk is ready for being uttered ("lurking") and the preceding chunk is "subsiding", i.e., done with executing its meaning-bearing parts (intonation phrase and gesture stroke).

For example, suppose that Max has just completely uttered the intonation of chunk, has performed the corresponding gesture stroke, and is now moving his hands back to a rest position. In the next chunk, which is to be seamlessly connected, the linguistic elements that are coexpressive of gesture are located relatively early in the intonation phrase, and the gesture requires under the current movement conditions an extensive preparation. Thus, movement needs to start early in order to meet within natural velocities the mandatory timing of stroke onset. ACE thus creates a fluent gesture transition after an only partial retraction, due to the position of the coexpressive speech within the next verbal phrase. Additionally, the vocal pause between the intonation phrases is stretched as needed for the speech-preceding preparatory movement due to the current movement context.

6 Conclusions

How must a computational model be conceived to better account for the empirical data, while being able to produce behavior that exceeds previous approaches? When compared with this model, we can draw the following five conclusions as to how the design of computational models needs to go beyond the aforementioned uni-directional three-staged process.

First, we propose to extend the concept of feedback in-between the generation and formulation processes. First, not only encoding problems should be effectual, but also the existence of primed linguistic constructions. Secondly, the bottom-up flow of information should take pace both in the speech as well as the gesture pathway to account for alignment of motor representations, either through self-priming (resulting from a gesture performed before) or through perceptuo-motoric processes (observing someone else doing a gesture). This may eventually result in a grounding of information distribution or packaging in lower levels of grammatical encoding and action schemas, and it may provide a way of modeling the impact that gestures can have on the conceptualization for running speech as proposed by Alibali & Kita [1].

Second, we suggest two different, complementary approaches to address the mapping of meaning onto gesture morphology: Implementing a top-down combinatorial search through combinations of surface morphology features, and assuming a bottom-up guidance by gestural motor schemas that are able to fulfill abstract depictive strategies. While the former can directly map semantic features onto form features like finger aperture or movement trajectory, in the latter approach, the availability of action schemas and their possible parameters pick up and bind features of visuo-spatial meaning that can be conveyed.

Third, cross-modal interaction must not be limited to the level of Action and Message Generation. An interplay between Formulator and Gesture Planner is needed for the coordination that enables the temporal synchrony of the verbal part of the utterance and its accompanying gesture(s). Additionally, interactions at the lower levels of phonation and motor control may account for compensation of potentially occuring timing lags during execution (e.g. pre-stroke holds in gesture).

Fourth, it is tempting to think of microplanning, i.e. the forming of speech and gestures to encode given meanings, as a form of negotiation between Action Generator, Message Generator, Formulator and Communication Planner, in which an "optimal" collection of semantic features is selected that can readily and most efficiently be uttered. However, it seems more plausible to assume that the stages come up with an approximative solution that can be found in the time available, e.g. words and gestures that carry most but not all of the meaning at hand. Rich feedback about what has been successfully encoded would then allow for picking up with an appropriate next utterance.

And finally, as one consequence, the evolving discourse context must comprise information not only about which meanings have been (successfully or not) communicated, but also the words and gestures that have been employed for this. This may be realized by decaying activations of knowledge structures involved, from propositional and visuo-spatial to linguistic and motoric representations.

These five conclusions frame our ongoing work on building a model for aligned speech and gesture production, which we pursue in the newly established collaborative research center (CRC) 673 "Alignment in Communication". An empirical study is underway intended to isolate the relevant aspects of meaning representations and their mapping onto gesture morphology, possibly based on a set of different depictive strategies (e.g. drawing 2D outlines or pantomiming an action to refer to a structurally coupled object). Our experimental setup combines a direction giving task for a Virtual Reality stimulus, which allows for fine control of the to-be-communicated content, in conjunction with eye-tracking to obtain information about which entities subjects perceptually focused. Building upon the empirial findings, we will devise a computational model of aligned speech-gesture production that can be implemented and tested in a simulation prototype based on our virtual human Max.

7 Acknowledgements

This research is partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 "Alignment in Communication".

REFERENCES

[1] Martha W. Alibali and Sotaro Kita. On the role of gesture in thinking and speaking: Prohibiting gesture alters childrens problem explanations, (submitted).

[2] Janet Bavelas, Christine Kenwood, Trudy Johnson, and Bruce Philips, 'An Experimental Study of When and How Speakers Use Gestures to Communicate', *Gesture*, **2:1**, 1–17, (2002).

[3] Kirsten Bergmann and Stefan Kopp, 'Verbal or visual: How information is distributed across speech and gesture in spatial dialog', in *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, eds., David Schlangen and Raquel Fernandez, pp. 90–97, (2006).

[4] Justine Cassell and Scott Prevost, 'Distribution of Semantic Features Across Speech and Gesture by Humans and Computers', in *Proceedings of the Workshop on Integration of Gesture in Language and Speech*, (1996).

[5] Justine Cassell, Matthew Stone, and Hao Yan, 'Coordination and Context-dependence in the Generation of Embodied Conversation', in *First International Conference on Natural Language Generation*, (2000).

[6] Justine Cassell, Hannes Vilhjalmsson, and Timothy Bickmore, 'BEAT: The Behavior Expression Animation Toolkit', in *Proceedings of SIGGRAPH 01: Los Angeles*, (2001).

[7] Jan Peter de Ruiter, *Gesture and speech production*, Ph.D. dissertation, University of Nijmwegen, 1998.

[8] Jan Peter de Ruiter, 'The Production of Gesture and Speech', in *Language and Gesture*, Cambridge University Press, (2000).

[9] Michel Denis, 'The Description of Routes: A Cognitive Approach to the Production of Spatial Discourse', *Current Psychology of Cognition*, **16**, 409–458, (1997).

[10] Uri Hadar and Brian Butterworth, 'Iconic Gestures, Imagery, and Word Retrieval in Speech', *Semiotica*, **115**, 147–172, (1997).

[11] B. Hartmann, M. Mancini, and C. Pelachaud, 'Implementing Expressive Gesture Synthesis for Embodied Conversational Agents', in *Gesture in Human-Computer Interaction and Simulation*, eds., S. Gibet, N. Courty, and J.-F. Kamp, (2005).

[12] Judith Holler and Geoffrey Beattie, 'How Iconic Gestures and Speech Interact in the Representation of Meaning: Are Both Aspects really Integral to the Process?', *Semiotica*, **146/1**, 81–116, (2003).

[13] Sotaro Kita and Asli Özyürek, 'What Does Cross-Linguistic Variation in Semantic Coordination of Speech and Gesture Reveal?: Evidence for an Interface Representation of Spatial Thinking and Speaking', *Journal of Memory and Language*, **48**, 16–32, (2003).

[14] Stefan Kopp, Paul Tepper, and Justine Cassell, 'Towards Integrated Microplanning of Language and Iconic Gesture for Multimodal Output', in *Proceedings of the 6th International Conference on Multimodal Interfaces*, pp. 97–104, New York, NY, USA, (2004). ACM Press.

[15] Stefan Kopp, Paul Tepper, Kim Ferriman, Kristina Striegnitz, and Justine Cassell, 'Trading Spaces: How Humans and Humanoids use Speech and Gesture to Give Directions', in *Engineering Approaches to Conversational Informatics*, ed., T. Nishida, John Wiley, (to appear).

[16] Stefan Kopp and Ipke Wachsmuth, 'Synthesizing Multimodal Utterances for Conversational Agents', *Computer Animation and Virtual Worlds*, **15(1)**, 39–52, (2004).

[17] Robert M. Krauss, Yihsiu Chen, and Rebecca F. Gottesman, 'Lexical Gestures and Lexical Access: A Process Model', in *Language and Gesture*, ed., David McNeill, Cambridge University Press, (2000).

[18] Willem J. M. Levelt, *Speaking: From Intention to Articulation*, MIT Press, 1989.

[19] David McNeill, *Hand and Mind - What Gestures Reveal about Thought*, University of Chicago Press: Chicago, 1992.

[20] Shuichi Nobe, 'Where do most spontaneous representational gestures actually occur with respect to speech?', in *Language and Gesture*, ed., David McNeill, Cambridge University Press, (2000).

[21] Martin Pickering and Simon Garrod, 'Toward a mechanistic psychology of dialogue', *Behavioral and Brain Sciences*, **27**, 169–226, (2004).

A semantic description of gesture in BML

Nicolas Ech Chafai[1, 2] and **Catherine Pelachaud**[1] and **Danielle Pelé**[2]

Abstract. Our aim is to animate Embodied Conversational Agents. In this paper we propose a semantic description of gestures based on two levels of gesture movement: the action elements, and primitives of movements. In the context of BML, a markup language of human behavior, this description proposes symbolic alternatives to the geometric ones, enhancing the possible use of gesture specifications from the semantic content of its feature.

1 INTRODUCTION

Since 1990's, several Embodied Conversational Agents (ECAs), have been developed [2, 4, 9] aiming at giving autonomy and interactivity when conversing with users and other agents. Most of them use specific representation languages to describe agent's behaviour, mental and emotional state. Wishing to be able to integrate each other works and to mutualize works, some researchers started to establish common representation languages (RL) for behaviour and gestures. This work takes place in a project called BML (for Behaviour Markup Language), whose first task is working toward establishing a unified language for behavior description [5].

In our work and toward the elaboration of this gesture repository, we aim to enhance BML specifications with a description of gestures that preserves its semantic content: when a subject produces a gesture, there is some information encoded in that gesture that is relevant to interpret its meaning. For example, the axial location and direction of a gesture is mainly relevant when we perform a gesture expressing temporal relations. Since RLs aim at to be player independent, and for a large domain of applications, we think that the semantic content of gestures allows us to interpret each gesture entry with the level required by each specific application. For example, some platforms (computers, mobiles, internet application) may want to animate only the meaningful and relevant part of a gesture, whereas other platforms would want to animate the more precise gesture. If the animated character tries to produce a deictic gesture by pointing an object with her hand, but her hand is not available for a gesture (*eg.* if the character holds an object in her hand): the semantic information on this gesture tells us that the direction is meaningful, and we have the possibility to produce another gesture with the same meaning. There is as many possible utility of this semantic content as works we could proceed to interpret this semantic content. We propose to use symbolic values for the description of gestures, toward this aim.

Another important requirement for the RL of gestures is that we should be able to encode all the co-verbal gestures into that language. To this end, we aim at the definition of gesture primitives: movements, configurations, locations, and directions [3, 5, 6, 7, 11]. The composition of all elements included in such primitives should encompass all the gestures considered in co-verbal communication.

First, we present a general overview of BML (section 2). Then, based on Calbris's work, we present the notions we use for the description of gestures satisfying two requirements: a symbolic description (section 3), and a definition of primitives (section 4), for gestures.

2 THE BEHAVIOR MARKUP LANGUAGE

Before we present the Behavior Markup Language (BML), let us have a look on SAIBA.

2.1. The SAIBA framework

SAIBA (for: Situation, Agent, Intention, Behavior, Animation) corresponds to a project aiming at the specification of a unified multimodal behavior generation framework for the animation of Embodied Conversational Agents (ECAs) [5]. By now, some languages already exist [10] that represent the human behavior, but each of these languages has its own aims of application, and are commonly specific to this application. As mentioned in [5], SAIBA project follows three main objectives; the framework specified in SAIBA has to:

- be independent of a particular application or domain;
- be independent of the employed graphics and sound player model;
- represent a clear-cut separation between information type (function-related versus process-related specification of behavior).

The framework is three-stages: (1) the planning of the communicative intent; (2) the planning of the behavior that realizes this communicative intent; (3) the production of this behavior in an ECA application. Two markup languages describe the transitions between these stages. First, the Function Markup Language FML specifies the semantic units associated with a communicative event: if an agent has to show fear, or joy, or if its purpose is to convey an order, to show the acceptance, etc., these functions have to be encoded in FML. Second, since an ECA has to perform a particular behavior to convey these functions or emotions, a markup language is proposed by SAIBA to

[1] University of Paris 8 {n.chafai, pelachaud, @iut.univ-paris8.fr}

[2] France Télécom R&D {danielle.pele, @orange-ftgroup.com}

describe this behavior to the end of the animation of the ECA. This markup language is called BML for Behavior Markup Language, and we focus on this topic in the following of the article. We give a graphical representation of these two steps:

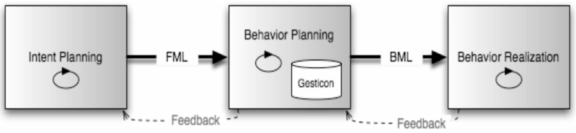


Figure 1. SAIBA framework for multimodal generation [5]

2.2. Focus on BML

The Behavior Markup Language (BML) is a representation language for the behavior an ECA has to produce. If the ECA has to speak, performing gestures during her speech, we encode which text it has to pronounce, and which gestures it has to perform into the BML. For the need of sharing gestures specifications, the gestures are described into a repository called *gestuary* [10] that BML refers to.

At this stage, one of the main problematic of BML is the resolution of the multimodal *synchronization and timing management*. Kopp *et al.* [5] propose a specification based on synchronization points, and on conditions and events. Using identifiers, we can synchronize with a gesture g1 by pointing to g1.stroke for example. Using events, the <wait> behavior implies a modality to be performed, to wait for an event to occur. For example, a modality can <wait> for the end of a gesture g2.

One of the other problematics is the modalities specification of: face, head movements, posture, gaze, gestures, which have particular form features that BML have to define. Our work looks at gesture description: each entry in the gesture repository have to define which are the physical parameters of a gesture, without taking into account the meaning of this gesture, but nevertheless without leaving out the meaning of the gesture *feature*. We detail this hypothesis in the remaining of the article.

3. A SYMBOLIC DESCRIPTION FOR GESTURES

3.1. Introduction

Toward the animation of an ECA called Greta, Pelachaud [9] has developed an editor for gestures based on a key frame description: each key frame of a gesture is described in term of gesture configuration using a system of conventions for the description of the sign languages called HamNoSys [11], and the system computes the interpolated frames. Lebourque *et al.* [7] also do uses HamNoSys conventions, considering its representation of movement, that distinguishes: straight, curved, wavy, and elliptical moves. From our point of view, the main problem of these types of representation is that they do not preserve the meaning of the encoded gestures, and restrains their possible uses. Thus, we use symbolic description for gesture, and express gesture movements in term of Action Elements (AEs). An AE considers the gestures from a subjective point of view: when a subject performs a gesture, she does not move her arm from a point to another one following a curved movement for example, but she performs the action of opening her arm, or rounding the arms, etc. At each corporeal element of gestures a list of AEs is possible; each element gives: the form of gesture movement, but also do preserve its semantic content.

Some researchers already have establish a list of actions describing gesture movements [1, 8]. For example, here is a list of arm symbolic feature extracted from Calbris [1]:

-direction = toward oneself, others, space;

-handedness = one hand, two hands, hand in interaction;

-actions = to fold, round, rise, fall, dangle, tend, cross, move aside, shake, fell, throw, etc.

These lists are not exhaustive, but do give a first view on the work we have to proceed. The elements of the action list (that we call: AEs) are the units of meaningful movements. That is, each arm movement that is meaningful for the gesture should be composed of one or more AEs. The arms, hands, and fingers, have a specific list of AEs. We present this feature for hand movements that are related to the self, or to the other.

3.2. Categories of gestures

The description of gestures in [1] gives a large repository organized into a taxonomy. The main dichotomy for this taxonomy is the distinction between the gestures produced in straight lines or planar surfaces, and the *gestures produced in curved lines or curved surfaces*. For the moment, we have studied only gesture from the first category. For this category, the taxonomy distinguishes gestures according to they are either:

a. planar: directed to the self, to the other, or in the space;

b. closed on something;

c. closed on a fist;

d. opening or closing;

Most of the gestures we produce during a conversational interaction are related to the first sub-category (planar). In this work, we are interested in gestures that are planar and related to the self or to the other; these sub-categories have common properties that we encompass in our study. We collect the action elements from the repository related to these subcategories, and we give examples for these AEs.

3.3. The Actions Elements for hand movements related to the self, or to the other

We have collected 19 Aes related to the self or to the other for gestures in straight line or planar surface from Calbris [1].

To locate: the hand localises a dirt on the face, or a part of the body that is meaningful.	*To push:* the hand pushes the back of the ear to imitate somebody with unstick ears.
To support: the hand supports the jaw to signify that we dream.	*To isolate*: the hand isolates or canals the mouth to yell.
To touch: the hand touches the jaw to signify a mistake.	*To dry*: the reverse of the hand dries the forehead to signify « it was short ».
To press/compress: the hand compresses the heart for the painful, the forehead when we are sick.	*To excavate*: the hand excavates the jaw to signify we are underweight.
To hide: the hand hides the eyes when we feel shame, or the mouth when we are surprise.	*To convey*: the hand conveys a kiss to other to signify a lovely « goodbye ».
To hit: the hand hits the forehead to signify « how idiot I am »	*To pinch*: the fingers pinch (and gloss) the chin to draw a beard.
To gloss: the hand glosses the back of the ear to draw a long hair.	*To hold*: the hand holds the chin when we are thinking.
To rule: the hand rules out the forehead to signify we are sick out of something, or the neck to signify we are feed up.	*To cut*: in french, the hand cuts the arm on different parts to signify « we give him *this* much, but he wants *this* much »
To hang: the thumb hangs the chin to laugh at somebody.	*To enclose*: the hands enclose the eyes to imitate the photograph.
To pull: the hand pulls the temples on forehead sides when we speak about a lifting.	

Table 1. Description of the Action Elements

Most of these AEs is related to the face or a part of the head, since the face is the center of the thinking, the forehead and the skull the centers of the memory, ideas, and mental pain, the eyes the center of the vision, etc., and thus the head is the center of the feelings and sensations.

At this point, we have to establish how to use these AEs toward the realization of our gesture editor. Some of these AEs have a similar configuration of the hand but are related to the face in differents manners (we localize and we touch with a planar hand, but we usually localize with the top of the fingers whereas we usually touch with our phalanxes); some of them share the same « function » of the hand to express different meanings (when we compress, push, or touch a part of the face, we use the same function of the hand which is tactile). These functions of the hand are related to its properties, and to the relevant dimensions of the hand when we perform a gesture.

In Table 2, we specify from which property of the hand the AEs derive (first column), the different forms for its expression (second column), and the possible AEs (third column). For example, the *reverse* of the hand is *convex* and thus can express the *excavation* of the jaw for someone who is underweight.

Function	**Forms**	**Action Element**
Convexity	- reverse of the hand	- excavate
Concavity	- palm	- isolate - enclose
Planarity	- palm - reverse of the hand - top of the phalanxes - bottom of the phalanxes	- support - press - hide - gloss - dry - convey - pull
Prehension	- inner hand - thumb - other fingers	- pinch - hold
Tactile	- palm - reverse of the hand - fingers	- hit - touch - hang - push - pull - locate - compress
Section	- section of the hand	- cut - rule

Table 2. Properties of the hand

Figure 2 shows the functions and forms of the hand:

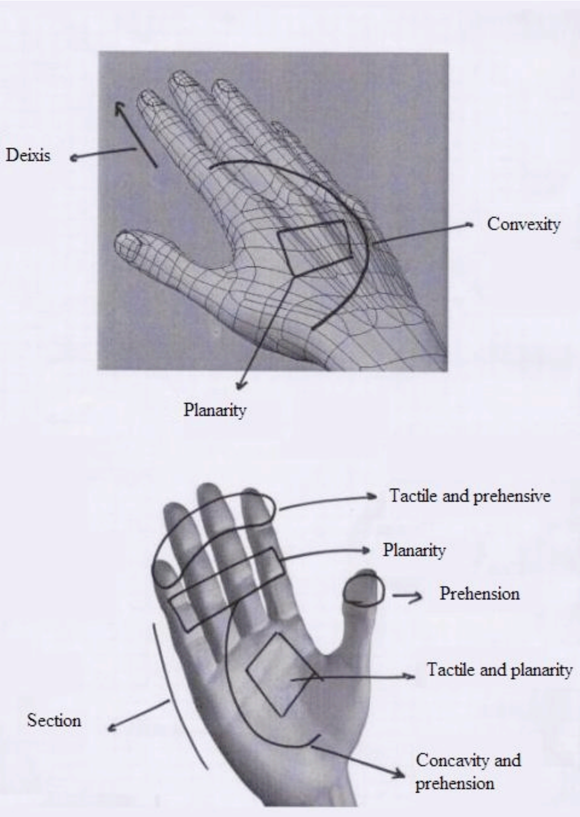


Figure 2. The functions and forms of the hand.

4. PRIMITIVES OF GESTURE FEATURES

The actions elements AEs we have introduced in the previous sections are described in term of symbolic values. Thus, we can observe two restrictions of this feature. First, *the completeness of AEs*: the list of action units is the minimum list of gesture element actions that are meaningful. Nevertheless, this list still corresponds to high level information on gesture movements. If we want a designer to compose new actions, we have to propose primitives of movements able to compose such actions. Second, *the precision of AEs*: an AE does not have an explicit numerical behaviour. That is, we do not impose an ECA to perform a specific AE with some specifics numerical values. Thus, there is a little freedom how we can interpret each element of the AEs. If we want to perform a precise gesture (with precise values), we have to express the primitives of movements with explicit values. To preserve the flexibility of our model, the primitives are expressed in term of measure units, a gesture equivalent of the facial animation parameters units in MPEG-4: *eg.* the expansion of a specific gesture in space is expressed in function of the ECA body dimensions, whereas self-contact locations of gestures are expressed in term of referent points (based on HamNoSys).

5. GLOBAL OVERVIEW

To summarize, the different elements we introduced are taking place into two different components in the process of an ECA animation:

a. from the point of view of a gesture repository: gestures are composed of a list of AEs, which are themselves composed of primitives of movements;

b. from the point of view of an ECA system: each primitive of movement is converted into a specific parameterised animation. The composition of primitives of movement into AEs allows the ECA to produce all the gestures defined in the gesture repository.

6. CONCLUSION

The schema that we have proposed enables to enhance BML specifications with semantic content for gestures. We express this semantic content in term of action elements that describe gestures from a subjective point of view, and that are defined with primitives of movements and measures units to allow the flexibility of the representation. In the future, we plan to extend the work on action elements to a larger part of co-verbal gestures, particularly to co-verbal gestures produced in space, and to complete the work on movement units that is still in progress.

REFERENCES

[1] G. Calbris, *Contribution à une analyse sémiologique de la mimique faciale et gestuelle française dans ses rapports avec la communication verbale*. PhD thesis, volume II (1983)

[2] J. Cassell, T.W. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H.H. Vilhjalmsson, and H. Yan, Embodiment in conversational interfaces: Rea. In: *Proceedings of CHI'99* (1999)

[3] B. Hartmann, M. Mancini, and C. Pelachaud, Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis. *Computer Animation*, Genève (2002)

[4] S. Kopp, B. Jung, N. Lessmann, and I. Wachsmuth, Max – a multimodal assistant in virtual reality construction. In *KI Zeitschift* (German Journal of Artificial Intelligence), Special Issue on Embodied Conversational Agents (2003)

[5] S. Kopp, B. Krenn, S. Marsella, A. Marshall, C. Pelachaud, H. Pirker, K. Thórisson, and H. Vilhjalmsson, Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. In *IVA '06*, Los Angeles, CA. (2006)

[6] A. Kranstedt, S. Kopp, and I. Wachsmuth, MURML: A Multimodal Utterance Representation Markup Language for Conversational Agents. *AAMAS'02 Workshop on Embodied conversational agents – Let's specify and evaluate them!*, Bologna, Italy (2002)

[7] T. Lebourque, and S. Gibet, A complete System for the Specification and the Generation of Sign Language Gestures. *Gesture Workshop '99*, Gif-sur-Yvette, France (1999)

[8] D. McNeill, and E.T. Levy, Conceptual representations in language activity and gesture. In R. Jarvella and W. Klein (eds.), *Speech, place, and action*, England (1982)

[9] C. Pelachaud, Multimodal expressive embodied conversational agent. *ACM Multimedia*, Brave New Topics, Singapor (2005)

[10] C. Pelachaud, B. Krenn, M. Lamolle, H. Pirker, and M. Schröder, *D6e: Report on representation languages*,

Deliverable NoE Humaine, see: http://emotion-research.net (2006)

[11] S. Prillwaitz, R. Leven, H. Zienert, T. Hanke, and J. Henning, Hamburg notation system for sign languages: An introductory guide. In: *International Studies on Sign Language and Communication of the Deaf*, volume V. Signum Press, Hamburg, Germany (1989)

Let’s shake hands!
On the coordination of gestures of humanoids

Zsófia Ruttkay and Herwin van Welbergen

1. INTRODUCTION

Hand gestures are important means of expressivity for humanoids. In this paper by humanoid we understand user-controlled or autonomous virtual humans (VHs) or conversational agents (ECAs) [4] as well as human-like robots [18]. We cover both cases of human-humanoid and humanoid-humanoid interaction. The semantics, the morphology, the variations in performance of gestures reflecting cultural, affective and other characteristics of the speaker [8]. as well as general gesture movement laws [6] have been addressed Our focus in this paper is the issue of *coordination* of hand gestures to external signals. One type of coordination, alignment of speech-accompanying gestures to the speech, has been studied extensively, and different design principles have been formulated and implemented for specific applications with virtual humans [11, 13, 25]. In these cases, the phonological synchrony rule [15] have been taken as basis, usually resulting in gestures timed to the speech – even if it is generated by TTS. An exception is [24], where in assembly tasks, where a physical manipulation may be accomplished in a shorter or longer time, the speech is aligned to the manipulative hand gestures. Another domain where two-handed gestures play a role is sign language [9]. Also, mechanisms for fast planning for deictic gestures have been proposed [14]. Our ongoing research extends these works in the following aspects:

- We propose a coordination scheme which is *more general*, allowing to take into consideration external events such as tempo indication or perceived state information about the interlocutor of the ECA.
- We allow the declaration of coordination requirements on a *low level of granularity*, looking at different stages of gestures. Such a refined approach makes it possible to perform experiments on e.g. expressivity and style, and to include timing strategies as a means to fine-tune the gesturing behavior of a humanoid.
- Our main interest is in *reactive scheduling and planning of gestures* with reference to an environment influencing the timing of the gestures.
- We are using the (still under development) *BML language* for the formulation of scheduling requirements. As BML is meant to become a general-purpose markup language [3], our testing and extension of its constructs contributes to the development of this unifying language.

One may wonder if it is necessary to endow humanoids with the capability of such subtle coordination. What are the application contexts where such coordination is needed?

In present applications typically an ECA is either ‘alone’, or at least not paying attention to the (real or virtual) partner while talking. If he does, it is via eye contact, and not body contact. Indeed, modeling gaze behavior [1] during conversation addresses a similar, albeit much simpler coordination problem, where no physical constraints are present to influence the gazing behavior of the parties.

Current humanoids are not adaptive and robust enough in their reactive gesturing behavior. However, consider a virtual world inhabited by multiple humanoids, either autonomous or as avatars driven by real people’s intentions. A very natural ‘act’ in such environments too is greeting, to initiate conversations, or indicate (e.g. in games, simulations) their relationship. Quite long ago hand shaking was one of the ‘wishes’ formulated as something a humanoid should be capable of [22]. Yet, it has not become a practice. Also, the subtle, reactive coordination assumes (real or simulated) perception of humanoids – a topic which is getting more attention recently [20].

Though our own field is virtual humans, we emphasize that a subtle, reactive coordination scheme can be applied for robots, where physical contact is a plus dimension of communication [16, 19], especially for building a common ground and expressing emotions [18, 3].

HMI, University of Twente, e-mail: zsofi@cs.utwente.nl

In our paper first we give an ontology of types of coordination problems related to hand gesturing. Then we demonstrate the problems on the basis of two examples: clapping as a rhythmic 2-handed gesture; and hand-shaking which is a single-handed gesture of two persons, which 'makes sense' only if the required coordination takes place. Hand gestures like these fall outside the usual speech-driven hand gesturing of humanoids, both concerning the coordination problems involved and the application context. Then we explain our own multimodal planning system to deal with the subtle coordination problems. Finally we outline the extra requirements these coordination tasks pose on the BML language. Our work is done parallel to gathering empirical data on relevant human-human gesturing, by recording and analyzing captured motion data, and possibly also other nonverbal signals gained from the recorded video. The proposed coordination strategies are to be demonstrated with virtual humans, and their gestures are to be tested. The present work is a step towards our ultimate goal; that is to build a multi-layer behavior engine for virtual humans, particularly for serious game applications.

2. ONTOLOGY OF COORDINATION OF HAND GESTURES

In order to characterize the coordination problems related to hand gestures, the following aspects need to be specified:

1. Origin of signals involved in the coordination
   a. an event or signal of the world;
   b. one or more other modalities of the humanoid self (speech, other hand, gaze,…);
   c. one or more modalities of another humanoid or a real human.
2. The flexibility of the timing of signals involved in the coordination
   a. a signal is an inflexible signal, if its given timing cannot be changed;
   b. the timing of the signal is flexible within certain constraints.

Table 1 gives an overview of the 3x2 cases, one of which – flexible signals from the word – is empty.

The concept of an inflexible signal is crucial: that is the signal which is to be taken 'as is', no alteration of its given timing is possible. In general, 'signals of the world' are such: one cannot change the tempo of a recorded music or the trajectory of a ball to be caught (but you may apply different strategies to catch it). In speaker-listener situation, the listener's feedback signals are the flexible ones, to be aligned with the speaker. (It is another question that the speaker may interpret the feedback so that he needs to talk slower, as the listener cannot follow his fast speech.)

But in cooperative tasks, both parties are flexible in order to achieve the common goal. On the other hand, from time to time one of them takes the initiative and expects the other to comply. Hence it may change from time to time if one or the other's hands (or other modality) is the leading, non-flexible signal.

Flexibility **Origin**	**Inflexible**	**Flexible**
World	- pointing at a *moving object* - clapping to *rhythm of music*	-
Humanoid's own modality	- gesture aligned to *speech* which is taken as leading signal	- *gaze and hand* coordination
Other humanoid's modality	- back-channeling as listener to a *speaker* e.g. by head nods	- *hand shake* - *two hands* involved in taking over an object

Table 1: Overview of signals for gesture coordination, the signal categorized is shown in italics.

Our earlier work on a listening conductor [2] made it clear that even a conductor-conducted musician relationship, it is not always the conductor who's hand movements are the leading signals – occasionally, he must adjust his conducting to the (too slow) music produced by the player. A similar, subtle 'game' happens often, albeit usually in an unconscious way, in case of interpersonal manipulative or communicative hand gestures, like carrying a heavy bag together with somebody, or shaking hands to greet.

3. COORDINATION IN TWO EXAMPLES

In order to perform a gymnastic exercise in a given rhythm, a hand gesture like a clap above the head is to be repeated according to the (may be changing) tempo 'dictated' by music, or a metronome. In this case, the metronome or music is a fixed signal of the world, and the hand gesture of the humanoid is flexible, which needs to be synchronized to tempo. The freedom in the synchronization is in how the total time is to be distributed among stages of the hand gesture. Note that in physical exercises it is important to be specific about the scheduling of the phases of the gesture, demanding e.g. that the stroke part of the gesture is done much faster, the continuity of the motion requires that no hold times are used, etc. We have gathered mocap data of joints and video and audio, and analyzed the synchronization strategies for tempo changes [23]. In a nutshell, we found that:

- The phonological synchrony rule was valid for counting while clapping.
- The clapping movement is often sped up just by decreasing the path distance decreases linearly with the clapping speed.
- A pre-stroke hold can be used as a slowdown strategy.
- The standard deviation of the relative phase angle between the left and right hand increased with the clapping frequency. No significant increase of the mean was observed.
- For our right handed subjects the motion was asymmetrical, the right hand was moving ahead in phase compared to the left.
-

Another example is the hand shaking of two (virtual) humans. There is little literature about how hand shake takes place among humans. Some studies address the variety in greeting gestures, depending on the social, gender and cultural characteristics of the people meeting [5, 7] and coordination in social interaction, in general [10]. As of the 'Western handshake', some normative guidelines are available for every-day scenarios [26]. The importance of establishing gaze contact, the strength of the grasp applied, the duration and number (2-4) and performance characteristic of 'pumping motions' such as to be performed from the elbow, are mentioned. Some social connotations of coordination and timing are noted. Particularly, the person initiating the hand shake is seen as the more dominant, socially higher-ranked - in Western business-like situations.

The coordination of this common greeting act is, in fact, a subtle process, where both participants are involved. One of the parties takes the initiative, and extends his right hand, to the 'normal' position of hand shaking start. If the other is to accept the hand, then he reaches towards the hand of the other. When does this movement start? What should be the target of the hand of the interlocutor? If the partner's hand is already in the hand shake start position in front of him, then obviously, the partner's hand is the target. But if the interlocutor started to move his hand while the partner's hand is still in motion (which is often the case in real life), the movement of the partner's hand is observed unconsciously: you do not grab a hand still in acceleration, but adapt your hand's motion in way that you both 'arrive' at a point where the hands hardly move and the palms are close enough to be embraced by the other's fingers. Once the hands are joined, they may kept sill, or a few 'pumping' motions take place, and at some point the parties extend their fingers and thus each of them may withdraw his hand. In reality, important factors of the entire process (who should start the hand shake, how long should it take) are controlled by social protocols and by visual and haptic feedback. The latter are used to accommodate to the special geometrical or other characteristics of the partner, such as size or position (e.g. seated).

The hand shake may be coordinated with other modalities: as soon as the hand contact is established, gaze contact should be established too, and kept as long as the hands are joined. Sometimes the partner does not respond to a hand shaking initiative, or keeps the hand beyond the will of the other. In the first case, how long should a humanoid wait with his extended hand to be grabbed by the partner? In the latter case, how to escape from such a situation? In reality, application of force may be enough, however, usually some social protocol is applied. The too long kept hand may be interpreted as sign of extra interest, or establishment of power relationship, and may be acknowledges or refused by a new communicative action (speech), also with the goal of ending the hand shaking.

Currently, we are busy with recording mocap and video data in situations where two persons need to great each other by handshake in a spontaneous, natural way; and in situations where one of the parties (an experimenter) tries to influence the handshake in different ways, e.g. being too slow with response (including no response at all), influencing the number of pumping motions and the duration of holding the other's hand. Besides eliciting motion characteristics, we are looking at the timing of gaze behavior.

4. REPRESENTATION AND PLANNING OF SYNCHRONIZATION

BML is a multimodal generation language, describing synchronization between speech and animation on such a level that it can be used as input for the final process of multimodal generation [11]. We extended BML, e.g. with *Observers* to monitor *outside actions*, that is ones not related to the humanoid's own modalities and synchronize to those too.

We have developed an interactive environment where the user may specify explicitly the tempo of repetitive gestures like claps, and tell the tempo of preparation and stroke, and how to distribute the remaining time (if any) between holds before and after the clap stroke. We implemented a demonstrator where a virtual character performs the clap sequence according to the specification, see figure 1.

Besides direct timing prescriptions, the amplitude of the motion may be prescribed too, which has consequences on timing. The amplitude-duration of stroke relationship is based on empirical tests with humans. Such a constraint between amplitude and duration is considered as characteristic of the clap as a gesture. However, when planning clapping, this constraint is to be taken into account in addition to the explicitly declared constraints.

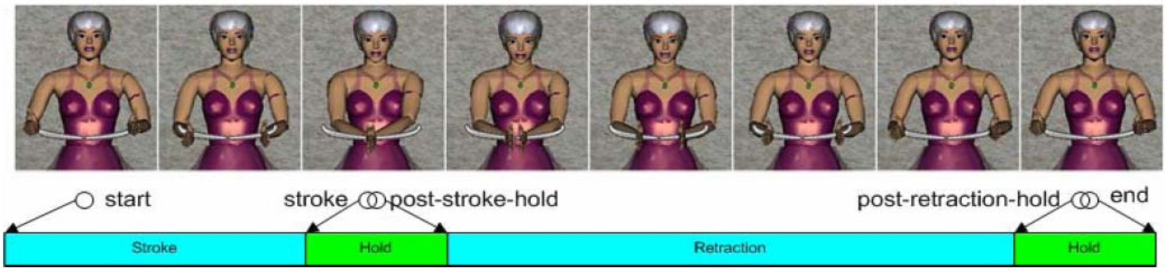


Figure 1: Duration of stages of a clap.

Currently, there is no built-in mechanism to prevent or correct inconsistent, infeasible specifications. Partial (underconstrained) specifications will be planned according to a 'default' strategy. E.g. if the tempo is to be slowed down, the duration of preparation and stroke will be slowed down proportionally, as default. However, the user may specify, for instance, that in the slow tempo the stroke should not be too slow, and the remaining time is assigned automatically to a hold.

For the clap gesture we used alignment points, identifying phases of hand gestures as the start and end of the entire gesture, the stroke, the pre- and post-stroke-hold, and retraction.

The timing constraints are expressed in BML linking (some of) these alignment points to each other and/or to the rhythmic signal of the world produced by a metronome. As a result of planning, all the alignment points of the subsequent claps get mapped to real time values, and the resulting animation is produced by the low-level animation engine which, in our current demonstrator, time-warps a default clap animation of the given amplitude.

5. DISCUSSION

The close look at clapping as (metronome) signal-driven inter-personal two-handed, and hand-shaking as a inter-personal cooperative gesture serve as studies for possible other cases where the success of some action is based on coordinated hand movement. We left unspecified in our framework how the signal from the outside world or the 'other' party is perceived and interpreted. In case of handshake, peripheral computer vision, or, if applied for robots, tactile feedback would be needed. Within our planning framework, it remains an issue how to propagate back updated, perceived information from the low-level planner to the higher level gesture scheduler?

Another interesting and difficult issue is the cognitive aspects of multi-party hand gesture performance. As mentioned in the introduction, in real life it depends on social and other characteristics of the partners who the 'leading' party will be. Also, deviations from a 'standard' performance convey meaning, intentions and believes of each party with respect to the other. Hence the subtle, perceived characteristics of the performance of a cooperative gesture may have effect on the cognitive aspects, as emotional state, believes, goals, user modeling. This issue has been raised at the recent BML meeting in a specific form, namely how to prepare a humanoid for failure in gesturing, e.g. due to malfunctioning of the lowest-level animation mechanisms. The issue of gapping the low-level gesturing and the mind via a feedback mechanisms is essential for interpreting behavior of partners of humanoids, and learning about the (potential) communicative partners.

REFERENCES

1. Argyle, M. and M. Cook: Gaze and mutual gaze, Cambridge University Press, 1976.

2. P. Bos, D. Reidsma, Zs. Ruttkay and A. Nijholt: Interacting with a virtual conductor. Proc. of 5th International Conference on Entertainment Computing, Cambridge, UK (September 2006), no. 4161 in Lecture Notes in Computer Science, Springer Verlag. 2006, pp. 25–30.

3. C. Breazeal: Social interactions in HRI: the robot view, **.** Systems, Man and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 34:2, 2004, pp. 181-186.

4. Justine Cassell: Embodied Conversational Agents. The MIT Press, April 2000.

5. A. Duranti: Universal and Culture-Specific Properties of Greetings, Journal of Linguistic Anthropology, Vol. 7, No. 1, 1997, pp. 63-97.

6. S. Gibet, J.-F. Kamp and F. Poirier: Gesture Analysis: Invariant Laws in Movement. In Gesture-based Communication in Human-Computer Interaction, LNCS/LNAI, Volume 2915, 2004, pp 1-9.

7. Paul E. Greenbaum and Howard M. Rosenfeld: Varieties of touching in greetings: Sequential structure and sex-related differences, Journal of

Nonverbal Behavior, Volume 5, Number 1, 1980. pp. 13-25.

8. B. Hartmann, M. Mancini, C. Pelachaud, Implementing Expressive Gesture Synthesis for Embodied Conversational Agents, Gesture Workshop 2005, LNAI, Springer, 2005.

9. Matt Huenerfauth: Representing Coordination and Non-Coordination in American Sign Language Animations. Behaviour & Information Technology, Volume 25, Issue 4, 2006, pp. 285-295.

10. A. Kendon: Movement Coordination in Social Interaction: Some examples described. Acta Psychologia, 32, 1970. pp. 100-125.

11. B. Krenn, H. Pirker: Defining the Gesticon: Language and Gesture Coordination for Interacting Embodied Agents, Proc. of the AISB-2004 Symposium on Language, Speech and Gesture for Expressive Characters, Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour, University of Leeds, UK, 2004, pp.107-115.

12. S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, H. Vilhjálmsson: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. Proc. of IVA 2006, pp. 205-217.

13. Stefan Kopp and Ipke Wachsmuth: Model-based animation of coverbal gesture. In Proceedings of Computer Animation, Washington, IEEE Computer Society, 2002, pp. 252-257.

14. J. Lester, J.Voerman,, S.Towns, C.Callaway: Deictic believability: Coordinated gesture, locomotion and speech in lifelike pedagogical agents, Applied AI, Vol. 13. No. 4/5. 1999. pp. 383-414.

15. D. McNeill: Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, Chicago, 1995.

16. C. L. Nehaniv, K. Dautenhahn, J. Kubacki, M. Haegele, C. Parlitz, and R. Alami: A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction. Proc. of ROMAN 2005, pp. 371- 377.

17. Z. M. Ruttkay, J. Zwiers, H. van Welbergen, and D. Reidsma. Towards a reactive virtual trainer. In Proceedings of the 6th International Conference on Intelligent Virtual Agents, Springer, 2006. pp. 292–303.

18. Sidner, C. and C. Lee and C. Kidd and N. Lesh: Explorations in Engagement for Humans and Robots, Artificial Intelligence, May 2005.

19. P. Olivier, D.G. Jackson & C. Wiggins.A Real-world Architecture for the Synthesis of Spontaneous Gesture, Proceedings of 19th annual conference on Computer Animation and Social Agents (CASA2006), Geneva, Switzerland, 2006.

20. C. Peters: Evaluating perception of interaction initiation in virtual environments using humanoid agents. Proceedings of the 17th European Conference on Artificial Intelligence, 2006. pp. 46--50.

21. Michihiko Shoji, Kanako Miura, and Atsushi Konno: U-Tsu-Shi-O-Mi: The Virtual Humanoid You Can Reach, SIGGRAPH 2006 Emerging technologies

22. D. Thalmann, R. Boulic, Z. Huang, and H. Noser, Virtual and Real Humans Interacting in the Virtual World, Proc. International Conference on Virtual Systems and Multimedia `95, pp.48-57.

23. H. Van Welbergen and Zs. Ruttkay: On the parameterization of clapping, Proc. of Gesture Workshop 2007, Lisbon, Portugal, to appear.

24. I. Voss and I. Wachsmuth: Anticipation in a VR-based anthropomorphic construction assistant. In J. Jacko and C. Stephanidis (eds.): Human-Computer Interaction, Theory and Practice *(Part I)*, London: Lawrence Erlbaum Associates. 2003, pp. 1283-1287.

25. I. Wachsmuth and S. Kopp: Lifelike Gesture Synthesis and Timing for Conversational Agents. In I. Wachsmuth and T. Sowa (eds.): Gesture and Sign Language in Human-Computer Interaction Berlin: Springer (LNAI 2298), 2002, pp. 120-133.

26. Marta Wilson and Sharon Flinder: The Business Protocol Advantage, http://transformationsystems.com/Assets/BusProtocol.pdf

On the simulation of interactive non-verbal behaviour in virtual humans

John Shearer[1] and Patrick Olivier[1] and Marco De Boni[2]

Abstract Development of virtual humans has focused mainly in two broad areas – conversational agents and computer game characters. Computer game characters have traditionally been action-oriented – focused on the game-play – and conversational agents have been focused on sensible/intelligent conversation. While virtual humans have incorporated some form of non-verbal behaviour, this has been quite limited and more importantly not connected or connected very loosely with the behaviour of a real human interacting with the virtual human – due to a lack of sensor data and no system to respond to that data. The interactional aspect of non-verbal behaviour is highly important in human-human interactions and previous research has demonstrated that people treat media (and therefore virtual humans) as real people, and so interactive non-verbal behaviour is also important in the development of virtual humans. This paper presents the challenges in creating virtual humans that are non-verbally interactive and drawing corollaries with the development history of control systems in robotics presents some approaches to solving these challenges – specifically using behaviour based systems - and shows how an order of magnitude increase in response time of virtual humans in conversation can be obtained and that the development of rapidly responding non-verbal behaviours can start with just a few behaviours with more behaviours added without difficulty later in development.

1 INTRODUCTION

Interactive non-verbal behaviour is important in human-human interaction, but has to date been given very limited attention by the virtual human AI community. AI in games has been focused more on game play with attention only recently towards non-verbal communication in games such as Half-Life 2 [1]. Previous games had non-verbal communication limited to cut-scenes. AI researchers have been focused on conversation for a long time, but mainly under a natural language processing paradigm – that is trying to understand spoken (or more often textual) language and respond appropriately[2]. More recently virtual humans capable for full body expression have been developed and these have proved engaging[3, 4, 5]. Their limitation has been that similar to simpler text-based or speech-based systems their only input has been typed or spoken speech. The non-verbal behaviour has therefore only been based on the textual input and output, ignoring the important behaviour in the non-verbal modality (though [6] and similar research attempts with some success to predict non-verbal behaviour based on the speech modality only as [7, 8, 9] show significant redundancy between the modalities). Non-verbal behaviour, especially gesture, has been given attention under a computer control paradigm [10] and also inform interactive system as a whole [11], but little attention has been given to the actual development of virtual humans that utilise non-verbal behaviour as both input and output, especially in a fast control loop. The notable exception to this is [12] who use head nod detection for conversational feedback – to inform the flow of conversation..

The introduction of more complex data streams to virtual humans introduces difficulties with the analysis of this data and also with determining appropriate behaviour based on this input data. Present AI systems in virtual humans are either very simple rule based systems, such as those in computer games or imitation agents [13], or highly complex natural language processing (NLP) systems that attempt to fully understand the context of spoken or more usually typed language and search for appropriate actions. Fully modelling the world and searching for appropriate actions has been possible due the limited form of data input. The additional complexity and unpredictability of non-verbal behaviour input introduces similar problems to AI systems for virtual humans that were approached in the 1980's for AI systems for robot control. The use of a full sense-process-act cycle for the AI systems was too complex and more importantly too *slow* for real-time systems (such as robotics, or interactive virtual humans). All virtual systems at present have a response time of at least half a second, and many much more (text systems usually only respond when new text is input).

In comparison with robotics AI history, the real-time behaviour of virtual humans is still in the first stage of development (sense-plan-act - which worked for robotics in simulated or highly restricted environments, and is still appropriate in many circumstances). In order for virtual humans to be interactive in real-environments their behaviour response time need to be reduced by an order of magnitude – towards that of humans in normal conversation. That is, they need to response immediately to a users behaviour, which is not to say that their full response must be immediate, but that there must be *some* immediate response. We propose drawing on the further developmental stages of real-time robotics AI systems to provide inspiration for virtual human AI systems – specifically subsumption architecture[14] and behaviour based systems, moving towards more hybrid systems[15] drawing on the strengths of present virtual human AI systems with the addition of simpler fast response behaviours. These stages of robot

[1] Newcastle University, Culture Lab, Kings Walk, Newcastle University, NE1 7RU, UK, email: john.shearer@ncl.ac.uk; p.l.olivier@ncl.ac.uk

[2] Unilever Corporate Research, Unilever R&D Colworth, Sharnbrook, Bedford, MK44 1LQ, UK, email: marco.de-boni@unilever.com

AI systems made it possible, in addition to increased response times, to build up robot behaviours step-by-step with increased reliability and robustness using less computing power than previously thought possible. We believe that it is possible to build up a fully interactive virtual human using a hybrid approach of behaviour based systems and the more traditional virtual human techniques, but at this state the focus in on developing early prototypes that interact in simple ways before moving towards more complex systems.

The next section provides more detail and history of the development of AI control systems in robotics along with the advantages and disadvantages of these approaches. Section 3 shows how these developments can be applied in virtual humans and discusses the importance of conversational state in interactions and that the relative context-freeness (from the specific high-level conversation meaning) enables that behaviours can be modulated by the conversational state without awareness of that high-level context. We then provide some details on the present state of development our behaviour based virtual human system and discuss how it is possible to initially build as system with just a few behaviours, with further behaviours being able to be added at a later a date without difficulty. Finally moving on to some approaches to evaluating these virtual humans, both in their entirety and piecewise (i.e. evaluating *which* behaviours are important).

2 DEVELOPMENTAL HISTORY OF AI CONTROL SYSTEMS IN ROBOTICS

Norbert Weiner in the late 1940s developed the field of cybernetics – the "marriage of control theory, information science, and biology that seeks to explain the common principles of control and communication in both animals and machines"[16] – which affirmed the notion of situatedness – the strong two-way coupling between an organism and its environment[16]. It is this strong two-way coupling that seems to be missing from present state-of-the-art virtual humans. There is, of course, two-way coupling in all virtual humans. The difficulty lies with the limited strength of that coupling. The focus of this paper is on the limitation of the coupling in terms of the limited sensory input and the limited response speed – both contributing to the limited strength of the coupling. We should note at this point that there are other factors that reduce the strength of the coupling as compared with that of real human-human interactions, such as the lack of physicality, realism, etc in virtual humans.

Following on from Weiner's work W. Grey Walter designed and constructed some of the earliest robots using simple sensors and actuators (and entirely analogue computing), with strong coupling between those sensors and actuators [17]. These simple machines, consisting merely of two sensors (a photocell and a bump sensor), two actuators (motors), and two "nerve cells" (vacuum tubes) were capable of surprisingly complex behaviour – seeking light, heading towards a weak light, back away from a bright light, etc. For whatever reasons this work was not strongly continued until revived almost 30 years later by Braitenberg [18] as a series of thought experiments, which were eventually transformed into true robots. MIT's Media Lab built twelve such robots and demonstrated a large variety of simple behaviours, including a timed shadow seeker, an indecisive shadow-edge finder, a paranoid shadow-fearing robot and a driven light seeker [19].

It is generally held that the start of artificial intelligence (AI) as a separate field was associated with a summer research conference held at Dartmouth University in 1955, with the original proposal indicating that an intelligent machine "would tend to build up within itself an abstract model of the environment in which it is placed. If it were given a problem it could first explore solutions within the internal abstract model of the environment and then attempt external experiments" [20]. From this point onwards the dominant approach in robotics and AI research for the next three decades was this representational knowledge and deliberative reasoning approach - representing hierarchical structure by abstraction; and using "strong" knowledge employing explicit symbolic representational assertions about the world.

In [21], Brooks claimed that "planning is just a way of avoiding figuring out what to do next", and while that is perhaps a little extreme, it does embody the idea of behaviour based systems and exemplifies the reaction against the traditions of classical AI. At this point also, advances in robotic hardware made it feasible to test the behaviour based approaches in real robots. The area of distributed artificial intelligence (DAI) developed at or around the same time as behaviour based systems in robotics. The idea that multiple competing or cooperating processes (or demons/daemons, or agents) could generate coherent behaviour [22, 23, 24], and Arkin states "individual behaviours can often be viewed as independent agents in behaviour based robotics, relating it closely to DAI" [25].

Approaches and techniques for robotics control can be depicted in on a spectrum from deliberative system to reactive systems as in Figure 0 ([25], page 20). As discussed previously, other than in computer games the focus for humans has been towards the deliberative end of this spectrum – developing virtual humans with well developed high-level level intelligence abilities, but as shown in the diagram these more cognitive process have a slower response time. As each person knows from their own normal lives, interactions with other people are made up of a whole set of different responses that sit along the deliberative-reactive spectrum, and all these varied responses are important for a smooth and useful interaction, not just the high-level responses. Therefore, a virtual human (like most present day ones) that only exhibits high-level intelligence is missing out on important low-level intelligence, which is also important. The relative importance of the levels of intelligence is clearly variable and is not under discussion here, but it is clear from a long history of work is psychology that these lower-level intelligence responses, such as eye-contact, intonation, gesture, back-channel speech, are highly important in human interactions, and therefore also in human-virtual human interactions [26, 27, 28, 29, 30]. The structure of human (and other animal) brains reflects this continuum from simple to complex behaviours and while the physical separation of different parts of the brain for different behaviours was part of the inspiration for

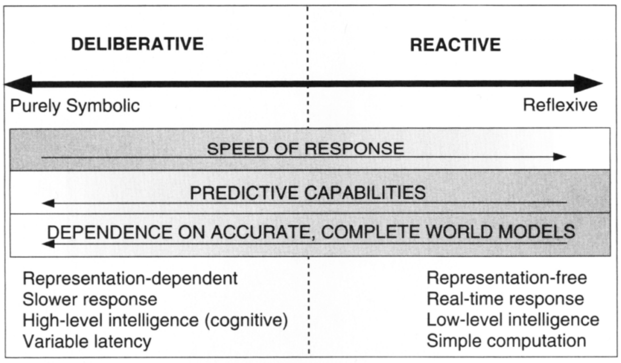


Figure 0 - Robot control systems spectrum

behaviour based systems, behaviour based system do not claim to be a replication or model of the human (or any animal) brain, merely drawing on them for ideas.

Robots (or virtual humans) utilising deliberative reasoning require relatively complete knowledge about the world and tend to struggle in more dynamic worlds where data that the reasoning processes uses may be inaccurate or have changed since last reading. More importantly, the deliberative reasoning process is frequently slow. Behaviour based systems or reactive systems were developed to attempt to solve some of the apparent drawbacks of deliberative systems – namely a lack of responsiveness in unstructured and uncertain environments.

A reactive control or a behaviour is a simply a tight coupling between perception and action to produce timely responses in dynamic and unstructured worlds. A behaviour based system is a collection of behaviours (perception-actions) pairs that cooperate/compete to produce more global behaviour. The obvious difficulty with having multiple behaviours is how to choose which behaviours should take control in times of conflict. The approach usually used in behaviour based systems in simply a priority system where higher priority behaviours win out over lower priority behaviours. The idea of one behaviour winning out over another (lower priority] behaviour also applies, in addition to behavioural outputs, to behavioural inputs. That is, rather than there existing a separate conflict “resolver” choosing between the outputs of behaviours A (high priority) and B (low priority), view behaviour A as inhibiting, or replacing the outputs of B. It is then, a relatively small leap to imagine that behaviour A could also inhibit or replace the *inputs* of B. This is the idea of Brooks’ “subsumption architecture” [14].

Within the field of robotics behaviour based systems saw significant success before running into the problem that almost inevitably, without any high-level or abstract representations the systems were incapable of the more complex behaviours that we wanted. The obvious next step was a hybrid between the two where behaviour based systems provide the fast, reactive control, while the deliberative systems provide the slower higher level cognitive control[15]. And it would be perhaps fair to say that many people would not view a robot or a virtual human with *only either* fast reactions *or* high level cognitive behaviours as intelligent – it would be both.

3 BEHAVIOUR BASED ARCHITECTURES FOR VIRTUAL HUMANS AND CONTEXT-FREE BEHAVIOURS

A behaviour based system consists of a set of behaviours, some of which can subsume (override or replace) the inputs and/or outputs of others (inhibition is simple overriding with nothing). We can view even slow high-level cognitive processes as behaviours, and therefore present deliberative virtual human control systems are simple behaviour based systems with one (or a few) complex behaviours, and furthermore a hybrid system is also just a behaviour based system. Behaviour based systems as applied to robots usually apply the behaviours directly to drive systems (motors, etc.). While this is possible in virtual humans (to control joint angles, muscle forces, etc.), it is also possible for a behaviour based system to control at a higher level – i.e. control the various animations that a virtual human may already have. This is the main adaptation needed to apply behaviour based systems to virtual humans.

Within human interactions the lower-level behaviours are predominantly unaware of the deeper meanings in an interaction and are consistent across different interactional contexts. In other words whether an interaction involves talking about the weather; discussing the latest cricket result; who ate all the pies; or solving world hunger, the majority of human interactional behaviours are still present and the same – i.e. people still look at each other (enough, but not too much); they still nod in agreement (in western cultures); and still give back-channel speech encouragers, etc. Of course, not all these behaviours are present all the time and are sometimes affected by high-level context, for the most part they are not. That said; these behaviours are influenced by the conversational state. This is the state of conversation from the simple state of whose turn it is to speak, to the deeper levels of state such as “Bill is speaking, but Ted is trying to break into the conversation”. These conversational states influence the various behaviours that are active (or their form). For example, as Ted is trying to break into the conversation, Bill will have increased behaviour(s) that try to hold the turn. In other conversational states Bill will have other behaviours enabled and disabled.

As one might expect the conversational state is again just a more complex behaviour or set of behaviours, with transitions between states caused by sensory input. So, this fits nicely into the whole behaviour based model – the conversation state behaviour modulates (or subsumes) some of the lower level behaviours.

Before moving on to some implementation details of behaviour based systems with virtual humans we should note that the idea of having rapidly interactive virtual humans has been worked on in the field of animation, especially by Perlin [31]. The main limitation of this work is that it was not grounded in behaviours and behavioural responses that real people use and that it did not investigate the scalability of the solutions (which behaviour based robotics has). It was found that character that react quickly and variedly to people were engaging and appeared to portray personality.

4 DEVELOPMENTS WITH BEHAVIOUR BASED VIRTUAL HUMANS

In practice when connected together a set of behaviours create a directed graph between input, output, and processing elements. The ideas of subsumption (one element overriding another's inputs and/or outputs) can be implemented by redirecting the edges within that graph. The idea of a graph of processing elements has been implemented in a variety of multimedia processing frameworks. Both DirectShow [32] on Windows and GStreamer [33] on Linux and Windows connected elements into pipelines or directed graphs. Additionally, the EyesWeb open platform [34] utilises a directed graph approach to supporting multimodal expressive interfaces and multimedia interactive systems and uses a visual programming paradigm whereby elements can be placed and connected together in a GUI. This visual programming paradigm is also present in both DirectShow and GStreamer. The advantage of EyesWeb is that it includes significant elements for performing both complex vision (OpenCV [35]) and audio processing, which is needed in order for a virtual character to respond to real-world sensory data.

For our early investigations into using behaviour based systems to control virtual humans, our virtual human [36] was adapted to be accessible from EyesWeb and we then designed simple vision and audio processing graphs (or pipelines) to control the character. We found that it is easy to create simple reactive behaviours, and the response time of the system is fast as it is only limited by the processing speed and the latencies of the hardware – there is no high level processing occurring at this point. It is no surprise that the main difficulties lie with the vision and audio processing – i.e. managing to detect the right things, but it is easy to add significant sensory capability in this system. The actual behavioural parts are straightforward, and it is simply a matter of moving some of the connections to subsume lower level behaviours. The follow on stage involves adding a larger set of detectable human behaviours and responses behaviours, followed by the modulation of these behaviours by the conversation state behaviour. We will also be using additional sources of interactional data, such as eye tracking. Further work will be reported at a later date, but behaviours are independent apart from their inputs and outputs being subsumable. Therefore adding additional behaviours does not invalid the previous ones – they can just be added in, subsuming other behaviours when needed.

5 EVALUATING BEHAVIOUR IN BEHAVIOUR BASED VIRTUAL HUMANS

General evaluation of virtual humans has been relatively limited to date [37, 38] and is dependent upon definitions of what metrics make a "good" virtual human and this varies with context. Within any specific domain metrics can be created to measure the important aspect within that domain, for example, how much people like the virtual human. But, the focus in this paper is not on evaluating virtual humans generally, but on how to evaluate a) whether a virtual human with these additional simple, fast-acting behaviours is better, and b) which of those behaviours help the most. Both these evaluations could be run together. Assuming one had an appropriate metric, the virtual human could be tested with a variety of combinations of behaviours on and off - including all behaviours expect the high-level cognitive behaviours off (i.e. a virtual human like present ones), vice versa (how good is a virtual human with *only* simple behaviours?), and any other combinations. Statistical analysis will allow the determination of the quality contributions of the individual behaviours. The knowledge of which behaviours are important will be useful not only to inform which behaviours to focus on in terms of development or in more limited systems, but also useful to inform (or be a test bed for) areas such as psychology which behaviours are especially important in human-human interactions. This could be especially useful for people suffering from various forms of autism – both to inform which behaviours they could focus on, but also to provide a transparent systems where they could see how and why it responds as it does. Finally, we haven't discussed or tried how these virtual human would respond to each others more varied set of behaviours. This is something that could be highly interesting to investigate in the future, and how interactions that are interesting or realistic to a third party observe could be based on *only* simple behaviours.

REFERENCES

[1] Valve Corporation, *Half-Life 2*, Valve Corporation, Bellevue, Washington, 2004.

[2] J. Weizenbaum, *ELIZA - a computer program for the study of natural language communication between man and machine*, Communications of the ACM, 9 (1966), pp. 36-45.

[3] J. Cassell, H. H. Vilhjálmsson and T. Bickmore, *BEAT: the Behavior Expression Animation Toolkit*, *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, ACM Press, 2001.

[4] P. Tepper, S. Kopp and J. Cassell, *Content in Context: Generating Language and Iconic Gesture without a Gestionary*, *Workshop on Balanced Perception and Action in ECAs at AAMAS*, 2004.

[5] J.-C. Martin, R. Niewiadomski, L. Devillers, S. Buisine and C. Pelachaud, *Multimodal complex emotions: Gesture expressivity and blended facial expressions*, International Journal of Humanoid Robotics, Special Edition "Achieving Human-Like Qualities in Interactive Virtual and Physical Humanoids" (2006).

[6] H. Yan, *Paired Speech and Gesture Generation in Embodied Conversational Agents*, *Media Arts and Sciences, School of Architecture and planning*, Massachusetts Institute of Technology, Cambridge, MA, USA, 2000.

[7] A. Kendon, *Gesticulation and speech: Two aspects of the process of the utterance*, in M. R. Key, ed., *The Relation*

Between Verbal and Non-Verbal Communication, Mouton, The Hague, The Netherlands, 1980.

[8] D. McNeill, *Gesture and thought*, University of Chicago Press, Chicago, IL, 2005.

[9] Y. Xiong and F. Quek, *Gestural Hand Motion Oscillation and Symmetries for Multimodal Discourse: Detection and Analysis*, *Computer Vision and Pattern Recognition for Human Computer Interaction (CVPRHCI)*, Monona Terrace Convention Center, Madison, Wisconsin, USA, 2003.

[10] M. M. Cerney, *From gesture recognition to functional motion analysis: Quantitative techniques for the application and evaluation of human motion*, Iowa State University, Ames, 2005.

[11] H. Gunes, M. Piccardi and T. Jan, *Face and body gesture recognition for a vision-based multimodal analyzer*, *Pan-Sydney area workshop on Visual information processing: CRPIT '36*, Australian Computer Society, Inc., 2004.

[12] C. Sidner, C. Lee, L.-P. Morency and C. Forlines, *The Effect of Head-Nod Recognition in Human-Robot Conversation*, *Human-Robot Interaction*, Salt Lake City, Utah, USA, 2006.

[13] S. Kopp, T. Sowa and I. Wachsmuth, *Imitation Games with an Artificial Agent: From Mimicking to Understanding Shape-Related Iconic Gestures*, in V. Camurri, ed., *Lecture Notes in Computer Science*, Springer-Verlag, Berlin, 2004.

[14] R. A. Brooks, *A robust layered control system for a mobile robot*, Robotics and Automation, IEEE Journal of, 2 (1986), pp. 14-23.

[15] E. Gat, *On Three-layer architectures* in D. Kortenkamp, R. P. Bonnasso and R. Murphy, eds., *Artificial intelligence and mobile robots: case studies of successful robot systems* MIT Press, 1998.

[16] N. Wiener, *Cybernetics; or, Control and communication in the animal and the machine*, Wiley; Hermann et Cie, New York, Paris,, 1948.

[17] W. G. Walter, *The Living Brain*, W W Norton & Co, 1953.

[18] V. Braitenberg, *Vehicles, experiments in synthetic psychology*, MIT Press, Cambridge, Mass., 1984.

[19] D. W. Hogg, F. Martin, M. Resnick and Massachusetts Institute of Technology. Epistemology & Learning Research Group., *Braitenberg creatures*, Epistemology and Learning Group MIT Media Laboratory, Cambridge, MA, 1991.

[20] J. McCarthy, L. Minsky, N. Rochester and C. E. Shannon, *A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE*, 1955.

[21] Brooks, *Planning is just a way of avoiding figuring out what to do next*, Massachusetts Institute of Technology, 1987.

[22] G. Selfridge and U. Neisser, *Pattern Recognition by Machine*, Scientific American, 203 (1960), pp. 60-68.

[23] L. Erman, F. Hayes-Roth, V. Lesser and D. Reddy, *The Hearsay II Speech Understanding system: Integrating Knowledge to Resolve Uncertainty*, Computing Surveys, 12 (1980), pp. 213-53.

[24] M. L. Minsky, *The society of mind*, Simon and Schuster, New York, 1986.

[25] R. C. Arkin, *Behavior-based robotics*, Mit Press, Cambridge Mass, 1998.

[26] S. Duncan and D. W. Fiske, *Face-to-face interaction : research, methods, and theory*, L. Erlbaum Associates, Hillsdale, N.J. , 1977.

[27] M. L. Knapp and J. A. Daly, *Handbook of interpersonal communication*, SAGE Publications, Thousand Oaks, CA, 2002.

[28] L. Cerrato and M. Skhiri, *Analysis and measurement of communicative gestures in human dialogues*, *AVSP*, St. Jorioz, France, 2003.

[29] D. Efron, *Gesture and environment*, King's Crown Press, New York, 1941.

[30] D. McNeill, *Hand and mind : what gestures reveal about thought*, University of Chicago Press, 1992.

[31] K. Perlin, *Real Time Responsive Animation with Personality*, IEEE Transactions on Visualization and Computer Graphics, 1 (1995), pp. 5-15.

[32] Microsoft Corporation, *Microsoft DirectShow 9.0*, 2007 http://msdn2.microsoft.com/engb/library/ms783323.aspx.

[33] *GStreamer*, (2007), http://gstreamer.freedesktop.org/.

[34] A. Camurri, P. Coletta, A. Massari, B. Mazzarino, M. Peri, M. Ricchetti, A. Ricci and G. Volpe, *Toward real-time multimodal processing: EyesWeb 4.0*, *AISB 2004 Convention: Motion, Emotion and Cognition*, Leeds, UK, 2004.

[35] Intel Corporation, *OpenCV*, http://sourceforge.net, 2005.

[36] I. Haptek (2007), http://www.haptek.com/corporate/.

[37] Z. Ruttkay and C. Pelachaud, eds., *From Brows to Trust*, Kluwer Academic Publishers, Dordrecht, 2004.

[38] D. M. Dehn and S. v. Mulken, *The impact of animated interface agents: a review of empirical research*, International Journal of Human-Computer Studies, 52 (2000), pp. 1-22.

The CereVoice Characterful Speech Synthesiser SDK

Matthew P. Aylett [1] and **Christopher J. Pidcock** [2]

Abstract. CereProc®Ltd. have recently released a beta version of a commercial unit selection synthesiser featuring XML control of speech style. The system is freely available for academic use and allows fine control of the rendered speech as well as full timings to interface with avatars and other animation.

With reference to this system we will discuss current state-of-the-art commercial expressive synthesis, and argue that underlying current approaches to sythesis, and current commercial pressures, make it difficult for many systems to create characterful synthesis. We will present how CereProc's approach differs from the industry standard and how we have attempted to maintain and increase the characterfullness of CereVoice's output.

We will outline the expressive synthesis markup that is supported by the system, how these are expressed in underlying digital signal processing and selection tags. Finally we will present the concept of second pass synthesis where cues can be manually tweaked to allow direct control of intonation style.

1 INTRODUCTION

CereVoice®is a unit selection speech synthesis software development kit (SDK) produced by CereProc Ltd., a company founded in late 2005 with a focus on creating characterful synthesis and massively increasing the efficiency of unit selection voice creation. The system is designed with an open architecture, has a footprint of approximately 70Mb for a 16Khz voice and runs at approximately 10 channels realtime. The system is a diphone based unit selection system with pre-pruning and a Viterbi search for selecting candidates from the database similar to systems described in [3, 1, 4].

Speech synthesis has progressed enormously since the trademark Stephen Hawking voice which was based on synthesis developed in the mid-eighties. Current systems are acceptable for reading neutral material such as bank balances but sound unacceptable if you use them to read longer texts or more personal information.

We believe this is caused by current approaches to voice building. Most state-of-the-art synthesisers use unit selection to synthesise speech. This approach is based on recording a large database of speech and concatenating small sections of speech together to create new utterances.

The process for recording the database is time consuming (20-30 hours of studio time) and resource intensive. Thus, for commercial systems, a strong focus is made on creating neutral multiple-use voices. In addition, in order to improve concatenation there is an emphasis on reducing the variance of the speech within the database leading, for example, to requesting the source speaker to alter their natural speaking style to make it unnaturally neutral.

This results in voices which are completely inappropriate for expressive characters.

This leads to a vicious circle: commercial synthesis companies don't produce expressive voices so commercial customers can't develop systems using expressive voices. In turn, this forms the perception that there is no market for expressive voices and thus commercial synthesis companies don't create them.

2 EXPRESSIVE SYNTHESIS: Breaking the Deadlock

Four key elements are required for breaking the vicious circle of dull speech synthesis:

1. Voice building must be made more efficient.
   If it becomes possible to build a voice with 10 hours or 6 hours of studio time the incentive for building more voices and making them more expressive is greater. In addition it becomes possible to record a wider variety of speech styles while maintaining a sufficient commercial standard.
2. Control of speech style
   In order to make use of the variation recorded in the voice, it needs to be categorised, or automatically coded, when the voice is built, and the system needs to be able to select material based on this categorisation during synthesis.
3. Semi-automatic synthesis
   Although we don't yet understand how to completely control expressive voices we can use a limited amount of manual intervention to create expressive and characterful cues and prompts. Inserting automatic synthesis between these stock phrases is a pragmatic way of generating expressive dynamic synthesis.
4. Development of applications which require characterful synthesis
   In order to move the technology forwards we need pressure from innovative application developers who can see and harness the enormous potential of characterful synthesis.

CereProc has addressed the first issue by developing a completely automatic voice generation and capture system. This has made the general voice building process more efficient and allows more risks to be taken in the generation of expressive voices. For example a George Bush voice was successfully developed completely automatically from web based material.

In addition CereProc reduces the amount of material required for sound coverage using a process we term 'voice bulking' where unusual diphones (the basic unit used in the synthesis) can be synthetically generated offline. This allows more material to be recorded for prosodic and speech style coverage.

The ability to select and mimic speech styles is accomplished with the use of a rich XML control language. A special tag within this control language also allows the manual manipulation of the synthesis

[1] CereProc Ltd. and CSTR, University of Edinburgh, email: matthewa@inf.ed.ac.uk

[2] CereProc Ltd.

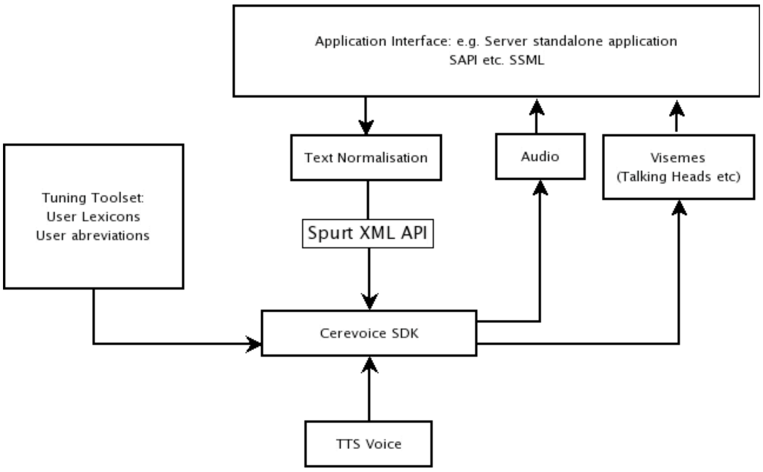


Figure 1. *Overview of the architecture of the CereVoice synthesis system. A key element in the architecture is the separation of text normalisation from the selection part of the system and the use of an XML API.*

process by allowing the user to cycle through the selection of sounds made for a particular word. This allows a simple manual method for discarding the units selected for a word and selecting an alternative set. In many cases this simple operation of discarding unwanted synthesis is sufficient for selecting synthesis which the user finds more appropriate.

Finally, by making the system freely available to the academic community as well as allowing innovative commercial enterprises to take part in an extensive beta test program, CereProc hopes that application developers will make use of this functionality and in turn drive the technology forward.

Despite this, perhaps the most important aspect of creating characterful voices is the simple intention of doing so. In many systems variation in speech style is removed in order to make smoother concatenation easier. CereProc, in contrast, prefers to retain the variation and put more effort in to developing the concatenation process. We have also found that users will accept minor concatenation errors if the voice has more personality. Given that many commercial voices have very few concatenation errors but have a speech style so dull and repetitive that extended synthesis becomes unacceptable, CereProc has found that commercial leverage can be gained by trying to offer voices which sound more characterful and give a stronger impression of a personality behind the voice.

2.1 Overview of the System

CereVoice is a new faster-than-realtime diphone unit selection speech synthesis engine, available for academic and commercial use. The core CereVoice engine is an enhanced synthesis 'back end', written in C for portability to a variety of platforms. The engine does not fit the classical definition of a synthesis back end, as it includes lexicon lookup and letter-to-sound rule modules, see Figure 1. An XML API defines the input to the engine. The API is based on the principle of a 'spurt' of speech. A spurt is defined as a portion of speech between two pauses.

To simplify the creation of applications based on CereVoice, the core engine is wrapped in higher level languages such as Python using Swig. For example, a simple Python/Tk GUI was written to generate the test sentences for the Blizzard challenge.

The CereVoice engine is agnostic about the 'front end' used to generate spurt XML. CereProc use a modular Python system for text processing. Spurt generation is carried out using a greedy incremental text normaliser. Spurts are subsequently marked up by reduction and homograph taggers to inform the engine of the correct lexical variant dependent on the spurt context.

3 CONTROLLING EXPRESSIVE SPEECH IN CEREVOICE

The CereVoice front end takes text and generates a series of XML objects we term spurts. The spurt is a section of speech surrounded by pauses. XML markup can be inserted into the input speech and is maintained in the spurt output. The CereVoice system allows a very wide variety of XML markup to control synthesis. Industry standard SSML markup [6] is converted by the front end into a 'reduced instruction set' of XML with a clear functional specification.

In addition, a set of XML markup is allowed which can change the selection process in the system, for example the ability to alter pitch targets. Tags used to alter selection are used in conjunction with tags which cause a change in the speech using digital signal processing to create different speech styles.

The speech styles are based on the activation-evaluation (AE) space, Figure 2. Here emotional states are described in terms of a value varying from very active to very passive and a value varying from very positive to very negative. Within CereVoice 1.2 (alpha) the perception of the emotional content of the speech in terms of the AE space is controlled by four speech style tags: happy (active/positive), calm (passive/positive), cross (active/negative), sad (passive/negative)[3].

Each tag gives a perception of emotion fairly central to each quarter of the AE space. Variation across the positive/negative plane is created by recording two extra sub-sets of data from the speaker. In the first the speaker is requested to produce speech with an unusually relaxed voice quality and in the second set with an unusually tense voice quality. The extent speakers are able to modify their speech in this way, and its relationship to their normal speaking style, varies. This in turn can affect how strongly a change is perceived when the tags are applied. For example, CereProc's Scottish voice 'Heather'

[3] Subject to UK patent application: 0704205.4

was chosen specifically for her cheerful and relaxed speaking style. For this reason the movement into the positive side of the AE space for this voice is less marked than towards the negative side of the space.

Variation across the active/passive plane is achieved using digital signal processing. In general a higher average pitch, a slightly faster speech rate, and increased speech volume make the speech sound more active and whereas the converse make the speech sound more passive.

The intensity of the perceived emotion across the active/passive dimension can thus be altered by changing the underlying control tags that make up the speech style. However their are severe limitations to the extent this is effective. Only a certain degree of modification can be carried out on the speech before it begins to sound unnatural rather than more active or more passive. Thus it is not possible to generate hyper-emotional such as fury, bliss, despair. In contrast, if the changes are too small the change a change in speech style is not perceived at all.

Despite the difficulty of subtle control and the inability to reach edges of the AE space, the use of the tags can be very effective. Much work in altering the perceived emotion of synthetic speech generated using the unit selection approach has concentrated on comparing identical sentences with differences in pitch, duration, rate and voice quality. This is because the content of the sentence has a strong effect on a subjects perception of the emotion in the speech. For a scientific evaluation of the importance of the different cues for the perception of emotion the effect of content is a confounding factor. Fortunately, as a pragmatic engineering solution for adding emotion to a voice, it acts a strong reinforcement to the underlying effects of the speech tags.

This reinforcing effect can be further improved if the negative/positive voice subsets also focus on covering negative and positive vocabulary items.

Overall, the positive/negative voice quality data, the ability to effect unit selection based on pitch and duration features, and the application of rate, pitch and duration changes using digital signal processing act a little like an artists pallet. Creating satisfying emotional characteristics using this functionality is still extremely difficult, just as being able to paint a picture is difficult no matter how many expensive brushes and paints you may have. Making this functionality available in a state of the art commercial synthesiser is, however, a critical step in making characterful synthesis possible.

4 SECOND PASS SYNTHESIS

The vast proportion of speech audio currently used in computer applications is in the form of recorded prompts. This alone demonstrates that although fully automated synthesis is required for completely dynamic content, much content is, in fact, not that dynamic at all. Currently, users of speech synthesis have used markup such as SSML [6] to manually control exactly how synthesis is realised. However the format of much of this markup stems from earlier diphone based synthesis systems rather than database approaches. CereVoice, however, will accept markup which allows users to control the inner working of the selection process. Such manual intervention is an effective stop-gap technique for competing with natural pre-recorded prompts.

Second-pass synthesis is a post-hoc method of tuning the synthesis output to improve the perceived quality of the output. A Viterbi search is used to find the 'best' sequence of states. In CereVoice it is possible to ask the engine to prune out a section of the best path found during the Viterbi search and to rerun the Viterbi over that section to find a less optimal alternative or *variant*. The next variant approach can be applied to a whole utterance or, more usefully, focus on a problem word or diphone. In the case of changing a single word or diphone in a larger utterance, units not within the the variant section are 'locked' to prevent modification of units that are considered acceptable. A new variant is selected by running the Viterbi search then pruning out the rejected selection of units. The pruning out of rejected units is cyclical, continuing until the requested variant number is found. Inside an XML spurt, a word can be enclosed by a 'usel' tag containing a variant attribute to force this behaviour. For example <usel variant='0'> is equivalent to no tag, and <usel variant='6'> would be the sixth alternative according to the Viterbi search. Fig. 3 shows a schematic of this process.

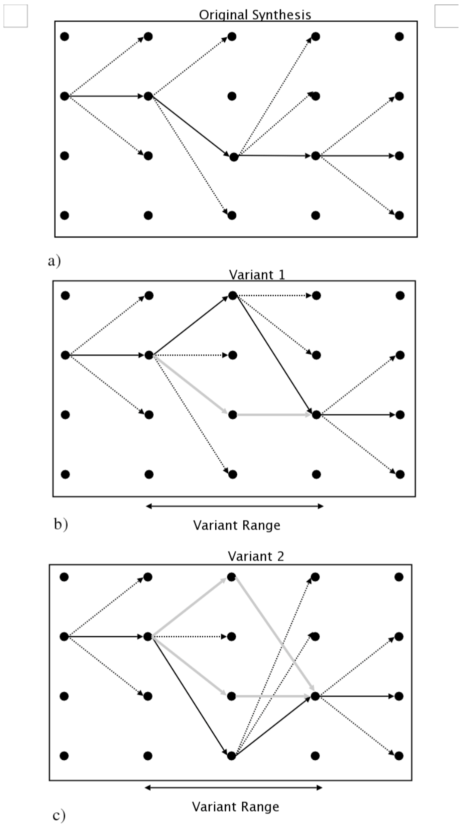


Figure 3. *Schematic diagram of the CereVoice variant tag process. a) The best path chosen by the Viterbi is shown as a black line. b) The unit in row 3 column 3 is rejected and the variant tag requests the next alternative. The path going through the unit is pruned out and a second path marked in black is selected. c) The new unit at row 1 column 3 is also rejected, the process is run again, a final acceptable unit at row 4 column 3 is selected.*

Below is an example of text marked up with variant tags.

```
The <usel variant='2'>Fruitto</usel> de
Mare featured, calamari served with <usel
variant='1'>tomatoes</usel>, peppers,
artichoke, avocado and, again, frisee.
```

Investigating efficient manual methods for improving synthesis addresses a crucial research question; given the database, how good could the synthesis become if our search algorithms produced optimum quality speech? In order to supply synthesis for entertainment there is a requirement for building fast, good quality characterful voices, often within specific domains. It is currently unclear what the degrees of freedom are for minimising the size of col-

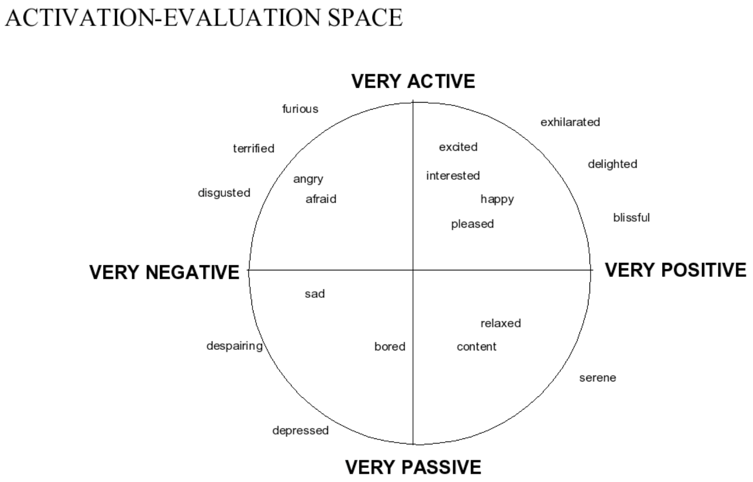


Figure 2. *Activation-Evaluation space. Adapted from [5] in turn adapted from [2]*

lected databases. Previous work which has tried to improve the quality of voices made from small databases has made use of information from a different voice with a larger database, either by using voice-morphing e.g.[3] or the larger voices prosodic model e.g.[4]. In contrast, second-pass synthesis allows us to answer the question of whether critical errors in the synthesis are caused by the poverty of the search algorithm or whether they are caused by database sparsity.

5 CASE STUDIES

In order to demonstrate the use of the XML control language we will present two case studies which show how they can be used. The first is an example of how the underlying tags in our Scottish voice are used to position the speech within the AE space for the 'happy' tag. The second is how we can use manual intervention to tailor a short paragraph of speech synthesised using our George Bush voice. Examples of the audio for these two case studies are available at http://www.cogsci.ed.ac.uk/∼matthewa/AISB2007.html.

5.1 Case Study 1: The Happy Tag

In order to explore how we create our happy speech style tag we will start by synthesising material which should be spoken happily in this example the sentence 'What a lovely day.' As discussed earlier, attempting to alter the emotional bias of the content is extremely difficult and will not be attempted here.

The baseline for this sentence is synthesised with the raw text:

```
What a lovely day.
```

The first stage in the process is to bias the unit selection to choose units from the calm voice quality section of the database. This is accomplished using a *genre* attribute within the unit selection tag *usel*.

```
<usel genre='calm'>What a lovely day.</usel>
```

This makes a major impact on the material selected and immediately produces a more positive sounding utterance. It sounds cheerful but not as upbeat as we might like. In order to make it sound more active we can in turn: increase the average pitch by 5 hertz,

```
<usel genre='calm'><sig f0='+5'>What a
lovely day.</sig></usel>
```

increase the amplitude. The value '2.0' used here does not directly increase the amplitude by two times its original value. In order to prevent clipping the speech is also compressed so that higher volume sections are not amplified as much as quiet sections.

```
<usel genre='calm'><sig f0='+5'
amplitude='2.0'>What a lovely
day.</sig></usel>
```

and increase the speech rate.

```
<usel genre='calm'><sig f0='+5'
amplitude='2.0' rate='1.05'>What a lovely
day.</sig></usel>
```

The combined effect is quite subtle but reasonably effective. The effects of the digital signal processing are more pronounced if you compare it do doing the opposite with the speech, i.e. reducing the pitch. lowering the amplitude and slowing the speech rate. The effect of this is to produce a stronger feeling of calm.

```
<usel genre='calm'><sig f0='-5'
amplitude='0.5' rate='0.95'>What a lovely
day.</sig></usel>
```

it is not possible to use digital processing techniques to make increase the percept of happiness much more than this. For example if we continue to increase pitch, amplitude and rate it begins to sound strange.

```
<usel genre='calm'><sig f0='+15'
amplitude='3.0' rate='1.2'>What a lovely
day.</sig></usel>
```

In our commercial system these underlying control tags are bundled into a SSML style tag <voice emotion='happy>.[4]

5.2 Case Study 2: George Bush Discusses HRI

The CereProc George Bush voice was created using audio trawled from the web. Unlike CereProc voices, where the design and capture of the audio is within our control, there is no guarantee of having appropriate coverage of phonetic material or that the acoustics will

[4] In our current system we use a lower amplitude increase in this bundled tag because we are currently unhappy with audible artifacts at the higher level described here.

be at the same standard we expect from a bespoke recording environment. In addition the transcriptions lifted from the web can be slightly inaccurate and that can cause quite serious synthesis errors.

For this reason the George Bush voice offers an excellent example of how we can remove mistakes and improve synthesis with a little manual intervention.

The text we chose to synthesise was taken from the first two sentences of the description of scope of the special session in AISB on language, speech and gesture for expressive characters.

```
Research into expressive characters, for
example embodied conversational agents, is a
growing field, while new work in human-robot
interaction (HRI) has also focused on
issues of expressive behaviour. With recent
developments in computer graphics, natural
language engineering and speech processing,
much of the technological platform for
expressive characters  both graphical and
robotic  is in place.
```

The raw synthesis of this material using the George Bush voice was reasonably acceptable but did contain some errors. Below is a marked up version of the text which gives a better rendition. The superscript beside each tag links to an explanation for its insertion below.

```
Research <lex phonemes='ih2 n t uw1'> into[1]
</lex> <usel variant='1'> expressive[2]
</usel> characters, for example embodied
conversational agents, is a growing field,
<break type='4'/>[3] while new work in <sig
rate='0.8'> human-robot[4] </sig> <lex
phonemes='ih1 n t er0 ae1 k sh ax0 n'>
interaction[5] </lex> <break type='0'/>[6]
(HR <usel variant='3'> I[7] </usel>) has
also focused on issues of expressive <lex
phonemes='b ax0 hh ey1 v y er0'> behaviour[8]
</lex>. With recent developments in computer
graphics, natural language engineering and
speech processing, <break type='4'/>[9] much
of the technological <usel variant='1'>
platform[10] </usel> for expressive characters
both graphical and robotic  is in <usel
variant='2'> place[11] </usel>.
```

The explanations for the additional tags are as follows:

1. The default stress on 'into' is to reduce it (i.e. 'inter' rather than 'intoo'). We override the pronunciation and thus the reduction with this tag.
2. There is a error caused by the database which produces something which sounds more like 'ixpressive' than 'expressive'. The variant tag discards this selection and the next selection does not have the error.
3. A comma normally generates an intermediate phrase break. In this case a the more final break '4' is appropriate. (Replacing the comma with a full stop would have had the same effect).
4. 'human-robot' is an unusual compound. A human speaker would typically make this more salient and the same effect can be achieved by using digital signal processing to slow the speech rate down by 20%.
5. It is hard to select the correct stress of syllables like 'in' in 'interaction'. By using the phoneme tag we have increased the stress from the default of secondary to primary by adding a '1' on the phone 'ih'.
6. The bracket creates a non-final phrase break by default. This has been removed by using a break of type '0' which prevents an odd pause before the acronym.
7. getting the stress right in acronyms is difficult. We want the voice to say hc**I** not h**C**i. We reject the first 0-2 variants of 'I'for being too reduced and use the variant '3' version.
8. The voice is a general American voice and doesn't have a lexical entries for British spellings. This is the US pronunciation of the word 'behavior'.
9. See note 3.
10. Again getting the stress right on compounds is difficult. We preferred the stress on the variant '1' to the original.
11. George bush doesn't have very much phrase final intonation in his speeches. Like many politicians he has learnt the trick of not sounding finished as he talks. Variant '2' was the first variant with a satisfying phrase final intonation.

This may seem a lot of manual work to get your synthesis to sound better. However, bear in mind we are using a voice that is not designed for this sort of synthesis. Most of the changes are actually using appropriate phrasing (spoken language has shorter sentences than written language), ensuring pronunciation is correct and fixing the odd concatenation error with a variant tag.

In this case, its also worth bearing in mind that getting George Bush into the recording studio and get him to say it perfectly is intractable, and even with more accessible voice talents re-recording material is a resource intensive and troublesome job.

Even if voices are constructed from limited prompt material, as the original prompts will be generated perfectly, we believe it is almost foolish not to use a synthesis solution to allow greater flexibility. After all, it offers more control and the possibility of creating new material without having to re-record.

6 CONCLUSION

Speech synthesis is a key enabling technology for pervasive computing. For many areas a key requirement is that the user is communicating with something which can simulate character and personality. Much current speech synthesis, although of a high standard for generating neutral speech, falls far short of what is required for giving character to avatars and speech based systems. Although there is much we do not understand in the generation of expressive speech it is possible to generate limited expressive speech and to further increase its effectiveness by offering more manual control of the speech rendered when required.

By making this technology freely available to the research establishment we hope to increase the awareness of this functionality, improve it and discover the extent it can produce innovative applications and user experiences.

REFERENCES

[1] Robert A.J. Clark, Korin Richmond, and Simon King, 'Festival 2 - build your own general purpose unit selection speech synthesiser', in *5th ESCA Workshop in Speech Synthesis*, pp. 147–151, (2004).

[2] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schröder, 'Is disfluency just difficult?', in *ISCA Tutorial and Research Workshop on Speech and Emotion*, (2000).

[3] A.J. Hunt and A.W. Black, 'Unit selection in concatanative speech synthesis using a large speech database', in *ICASSP*, volume 1, pp. 192–252, (1996).

[4] John Kominek, Christine L. Bennet, Brian Langer, and Arthur R. Toth, 'The Blizzard challenge 2005 CMU entry - a method for improving speech synthesis systems', in *INTERSPEECH*, pp. 85–88, (2005).

[5] R. Plutchik, *The Psychology and Biology of Emotion*, Harper Collins, New York, 1994.

[6] Paul Taylor and Amy Isard, 'SSML: A speech synthesis markup language', *Speech Communication*, **21**, 123–133, (1997).

Expressive Synthesis of Read Aloud Tales

Virginia Francisco and **Pablo Gervás** and **Mónica González** and **Carlos León** [1]

Abstract.

An important challenge for text-to-speech is to get a synthesized voice that sounds as similar as possible to human voice. However, nowadays the voice generated by synthesizers sounds artificial and this is the main cause of rejection by users. In this paper we propose a solution for modeling emotions in the FESTIVAL synthesizer by controlling the parameters of the system. We have chosen Fairy Tales as the domain for the synthesized text, because emotions play a fundamental role in the speech of such stories. We also present an evaluation process for the resulting voices of our prototype, and we show the results we have obtained in our first experiments, as well as the conclusions for those results.

1 Introduction

Often it is not possible to read text written on a screen. Some users who might not be able to access a particular textual system (children, or blind people) could access the information stored in a computer, if it was spoken, and all users may experiment a better user experience. However, nowadays most of the information in the computers is stored as text, and this impedes the retrieval of the content.

From this point of view, translating from written text to phonetical sounds that can be listened and understood by humans can be very useful. This process is known as *text-to-speech* (TTS), but it still presents many problems. One of the main challenges for *text-to-speech* is to get a synthesized voice that sounds as similar as possible to the human voice. Nowadays the voice generated by synthesizers sounds artificial. This is the main cause of rejection of this kind of systems by humans. To make the TTS systems more *user friendly* and, in this way, more useful for people, we have to generate voices that can express emotions, just like humans do.

There is much information in the way we speak that is not present in written text. It is very important for the generation of emotional voice to generate clear emotions, so that there will be no confusion for the listener. However, there is an important lack of emotions in the usual *text-to-speech* systems. In some domains, this might not be very crucial, but when narrating a fairy tale, for instance, synthesized voice must express emotions, because these emotions carry much information that should not be ignored.

This project arises to explore the possibility of modeling emotions through control parameters in an existing synthesizer when reading tales aloud. There are many theories which try to define emotional scales, and the choice of a specific scale determines the emotions that we try to distinguish. Another important challenge is to analyse the acoustic characteristics of human voice production at different emotional states in order to try to reproduce the same characteristics in the synthesizers.

To obtain the parameters that must be passed to the synthesizing system, we have to carry out an analysis of the emotional components of significant chunks of audio to create a model of that emotional speech. Once we have this model, the next task is to test the results.

The work presented in this paper extends previous work carried out with a different synthesizer [10]. The main goal of the present work is to change the synthesizer engine used in the previous system for another one which generates a more natural voice and allows us to control more voice parameters than the previous synthesizer. In the Conclusions of this paper we compare the results obtained in the present work with the results of the previous one.

2 Previous Work

The first studies about emotional speech were written by Fairbanks and Pronovost [9]. Even though this line of work gave rise to a great amount research and published articles, there is still a lot of important aspects to cover. The complexity of affective speech starts with the concept of emotion. Nowadays there are many theories about emotions, each of them with a different interest. Sometimes these theories are contradictory and it is difficult to integrate all of them in a single one.

Research on expressive synthesized speech has been carried out by Cahn[4], Murray and Arnott [18] for the English language, Burkhardt for German [2], Mozziconacci [17] for French and Montero [15] for Spanish.

2.1 Meaning of the Word “Emotion”

Emotions are defined as a flexible mechanism for the adaptation to a changing environment [21]. There are mainly two types of emotions [6]:

- Extreme emotions: This term denotes an emotion fully developed, which is intense and incorporates most of the aspects which are considered relevant in a emotion.
- Underlying emotions: It denotes the type of emotional colouring which is part of most of the mental states.

2.2 Clasification of Emotions

For the study of emotional speech we need to decide which emotions we are going to model, and how we are going to represent them. There are different methods in order to represent emotions [6]:

- *Emotional categories*. It is the most common method for the description of emotions. The method of Emotional categories uses

[1] Departamento de Inteligencia Artificial e Ingenieria del Software, Universidad Complutense de Madrid,Spain, email: virginia@fdi.ucm.es, pgervas@sip.ucm.es, monica.glez.jenal@gmail.com, cleon@fis.ucm.es

emotion-denoting words, or category labels for indicating emotions. Several approaches have been proposed in the literature for reducing the number of emotion-denoting adjectives:

  - *Basic emotions*. There is general agreement that some full-blown emotions are more basic than others. The number of basic emotions is usually small so it is possible to characterize each emotional category in terms of its intrinsic properties [7].
  - *Super ordinate emotion categories*. Some emotional categories have been proposed as more fundamental than others on the grounds that they include the others. Scherer [22] and Ortony suggest that an emotion A is more fundamental than another emotion B if the set of evaluation components of the emotion A are a subset of the evaluation components of the emotion B. Cowie and Cornellius [7] give a short overview of recent proposals of such lists.
  - *Essential everyday emotion terms*. A pragmatic approach is to ask for the emotion terms that play an important role in everyday life. The approach is exemplified by the work of Cowie [8], who proposed a Basic English Emotion Vocabulary.

- *Descriptions based on psychology*. The appraisal of a stimulus determines the significance of stimulus for the individual, and triggers an emotion as an appropriate response [1].
- *Descriptions based on evaluation*. Emotions are described from the point of view of the evaluations involved [19].
- *Circumflex models*. Emotional concepts are represented by means of a circular structure [20] such that two emotional categories being close in the circle represents the conceptual similarity of these two categories.
- *Emotional dimensions*. Emotional dimensions [6] represent the essential aspects of emotion concepts. Evaluation (positive/negative) and activation (active/passive) are the main dimensions; sometimes they are augmented with the power dimension (dominant/submissive). This approach is very useful because it allows measurement of the similarity between different emotional states. Another important property of this method is the relative arbitrarity in naming the dimensions.

2.3 Obtaining Prosodic Rules for Emotions

There are a lot of researches to obtain the prosodic rules which take part in the generation of emotional voice. These rules are obtained in different ways:

- Extracting it from the existing literature [3, 18].
- Analizing a corpus [16].
- Obtaining the optimum values from the systematic variation of the parameters in the synthesis [2, 17].

In the present work we are going to combine these three methods in order to obtain better results in the hope that the weaknesses of each individual approach are reduced by their combination.

2.4 Data Sources for Emotional Voice

The identification of the prosody associated to each emotion must be obtained empirically. There are different sources that have been used in the past in order to generate an emotional voice data base:

- *Actors*. The oldest and the most frequently used technique is to obtain recorded data from actors. The main advantage of that method it is that all the emotions can be reproduced using the same sentence [17] or the same pseudo-sentence composed of words with no sense [14]. This way the phonetics, prosody and voice quality can be compared in the same sentence with different emotions. Another advantage of this method is the facility of obtaining extreme emotions. A disadvantage of this technique is that the actor can reproduce a stereotype of the emotion which do not correspond with the emotion obtained spontaneously.
- *Expressive reading of emotional material*. It is a variant of the previous method suggested by Nick Campbell [5]. Campbell proposed to have readers that read texts with an appropriate verbal content with the emotion which is expected to be transmitted.
- *Production of emotions*. Subjects are urged to cause an emotion by means of the so-called MIPS (*Mood Induction Procedures*) [11].
- *Natural occurrences*. A research of Klaus Scherer, Bob Ladd and Kim Silverman [22] deals with the spontaneous generation of emotions.

Each one of these methods varies with respect to the control on the voice signal, from more to less control. These methods can be ordered in the following way: *actors*, *expressive reading of emotional material*, *production of emotions* and *natural occurrences*. Each of these is better or worse depending on the domain of the study. For the researching of extreme emotions the most appropriate is the use of *actors*. On the other hand for the researching of underlying emotions the best method is the observation of *natural occurrences*. In the case of studies centered on the speaker, the best choice is the *production of emotions*.

2.5 Prosody and Emotions

In all researches the global parameters of the prosody, like the base frequency, the scale of the base frequency and the speech rate, are treated like universals, at least when the number of emotional categories is small. The most interesting acoustic variables for voice synthesis are the ones that can be controlled through a voice synthesis system.

For modeling a system able to generate an emotional voice it is necessary to have a correspondence between the emotions and the values of the characteristics of the voice.

2.6 PRAAT

PRAAT[2] is a free, stable, scriptable and user-friendly scientific software program, designed and continuously developed by Paul Boersma and David Weenink at the Institute of Phonetic Sciences of the University of Amsterdam. It can run on a great variety of operating systems and allows to perform a great variety of tasks, which is why it is used in a wide range of situations such as phonetics classes, pronunciation improvement teaching and emotional voice synthesis research.

PRAAT does not only allow speech analysis but also speech synthesis, including articulatory synthesis. It can be used to manipulate speech as well as to create high-quality representations that show the parameters of the analyzed voice. These outputs can be spectrograms, intensity contours or even pitch and formants graphics. PRAAT's seemingly endless possibilities also include functions for learning algorithms, segmentation, labeling and listening experiments, filters, sound recording and a lot of other functionalities that are continually expanded by its users. This is why PRAAT is among

[2] http://www.fon.hum.uva.nl/praat

the most popular free downloadable speech analysis software packages and the reason why we chose to use it for our research.

For the present work the main advantage of PRAAT is the generation of high quality graphs in which pitch, spectrogram, intensity, formants, record pulses . . . can be visualized.From these graphics and analyzing different records we can establish how to change the voice characteristics in order to express emotions.

2.7 FESTIVAL

The synthesizer employed for our emotional story teller is FESTIVAL 3. FESTIVAL is a speech synthesis system that offers full text-to-speech through a number of APIs, such as the Scheme API, the Shell API, the Server/client API, the C/C++ API and the Java and JSAPI. It uses the UniSyn residual excited LPC diphone synthesizer, the CMU lexicon and letter-to-sound rules trained from it. The intonation was trained from the Boston University FMRadio corpus and the duration for this voice also comes from that database 4. It is multilingual and includes many voices. For our research we chose to use the default voice *kal diphone*, which is an American English male speaker.

The system is written in C++ and uses the Edinburgh Speech Tools for low level architecture and has a Scheme (SIOD) based command interpreter for control that we used to transform our SABLE marked up texts into audio files. We employed the FESTIVAL 1.95-beta version of the system and made it run on Cygwin[3], which is a Linux-like environment for Windows.

2.8 SABLE

SABLE[4] is an XML (Extensible Markup Language)/SGML (Standard Generalized Markup Language) based markup scheme for text-to-speech synthesis. It was developed to address the need for a common, system-independent TTS control paradigm. The aim of the Sable Consortium is to merge the STML (Spoken Text Markup Language) standard, developed by Bell Labs and the Edinburgh University, and Sun's JSML (Java Speech Markup Language). There are different groups involved or interested in this project, such as the Edinburgh University, Bell Laboratories, AT&T, Sun Microsystems and the Carnegie Mellon University. The 0.2 version of the Sable specification was released in March 1998 and FESTIVAL contains a basic implementation of it in its standard distribution since its 1.3.0 version. Although we found that not all tags have been implemented yet in the FESTIVAL 1.95-beta version we used for this research, the specification has been a useful guideline. The set of text description and speaker directives tags we finally used to mark up our texts are a subset of those implemented by the FESTIVAL 1.95-beta version that allowed us to modify the voice parameters our previous research efforts had proved to be relevant.

2.9 Evaluation Paradigms

There are several paradigms of evaluation for the emotional voice, the three most used are:

- *Forced choice*: This type of evaluation has been used in a lot of researches of generation of emotional voice [3, 15, 17]. The procedure is to give to evaluators a finite set of possible answers which includes all the emotions that have been modeled. The advantages of this approach are that it is easy to carry out, it provides a simple recognition measurement and it allows to compare different researches. On the other hand it has a disadvantage because it does not provide information about the quality of the stimulus from the point of view of naturalness and veracity.
- *Free choice*: The answer it is not restricted to a close set of emotions [18, 23]. It is very useful when the aim of the evaluation is to find unexpected phenomena.
- *Free choice modified*: Murray and Arnot [18] and then Stallo [25] introduced some modifications to the previous paradigm: introduced distraction categories, the “others” category, neutral texts with emotional texts. The difference between the recognition of the neutral text and the emotional text is taken as a measurement of the impact of prosody in the perception.

2.10 The Previous System: EMOSPEECH 0.1

The details of the previous system can be consulted in [10]. This system used FREETTS[5] as synthesizer, which is a voice synthesizer engine written entirely in Java, based on FLITE[6], and derived from FESTIVAL and FESTVOX[7]. FREETTS allows variations in the following voice parameters: pitch, pitch range, volume and rate.

It does not allow modifications of the parameters half way through a sentence, nor different assignations of parameters to different part of a sentence. This was found to be an important disadvantage for further work. For this reason, we have developed a new system, which will be explained in later sections.

3 Our Proposal

Fairy tale narration has been chosen as the domain of the application, because it is considered to be an environment where emotions clearly take part in the communication effort. Tales try to summarize the emotions that most of the children experiment in their way to maturity: happiness, sadness, anger, fear, envy. . . When reading a tale, one tends to exaggerate. The voice of the person or persons who read the tale will be as important a tool as the words themselves for a child to infer the emotions of the characters. Therefore we can affirm that the emotions expressed in a tale are extreme emotions, not underlying ones.

3.1 Emotions in the Fairy Tale Narration

In order to explain the voice markers which make our tale lively and personalized, we are going to use Scherer's research [21]. By means of the personality markers the speaker externalize some characteristics and the listener perceives it and assigns these characteristics to the speaker.

The classification of the emotions expressed in speech that satisfies best the requirements of a fairy tale teller is the basic emotions classification, because when a tale is being told, we usually exaggerate the emotions, so a small set of extreme emotions is enough. We have selected five basic categories in order to model the emotions: *happiness*, *sadness*, *fear*, *anger* and *surprise*.

In the tales synthesized by our story-teller there is only a narrator speaking and there are no dialogues. So we have decided that emotions are related to fragments of the tale and we have selected the sentences as emotional units. The narrator will try to impress the

[3] http://www.cygwin.com/

[4] http://www.cstr.ed.ac.uk/research/projects/sable/sable_spec2.html

[5] http://freetts.sourceforge.net

[6] http://www.speech.cs.cmu.edu/flite/

[7] http://www.festvox.org

emotion with which the sentence has been tagged, so that his voice will transmit sadness when he is reading a sad sentence and happiness when reading a happy one.

3.2 Tales Marked Up with Emotions

The input of our system are tales marked up with basic emotions. In order to get tales tagged at the same time as we generate them, an existing module for automatic story generation [12] has been modified. This module generated a conceptual representation of fairy tales and its corresponding text by means of natural language generation techniques. The input of the module are the actions which take part in the story plot and the semantic information about characters, locations, attributes and relations involved in the actions. From this input the story is generated automatically.

The marking up of tales in our generator is carried out in the *lexicalization* stage of the natural language generation process, where it is decided which specific words and phrases should be chosen to express the domain concepts and relations which appear in the messages. Given the basic linguistic structures used by the generation module, the mark up is done by phrases. The result of the *lexicalization* stage is a list of messages with their correspondent lexical forms and the emotion they are going to be marked up with. A final stage of *surface realization* assembles all the relevant pieces into linguistically and typographically correct text.

Two elements of the tales are taken into account when deciding the emotion associated to each sentence: characters and actions in which the characters are involved.

3.2.1 Emotions Associated to Characters

Using the traditional distinction between good and evil, the characters in our stories are supposed to be involved in good, bad or neutral situations. For each case, one of the basic emotions is associated to the character. For the tale "Cinderella" the emotions in Table 1 have been considered for the main characters.

	Good	**Bad**	**Neutral**
Cinderella	Happy	Sad	Neutral
prince	Happy	Sad	Neutral
father	Neutral	Neutral	Neutral
mother	Neutral	Neutral	Neutral
stepmother	Angry	Happy	Neutral
stepsisters	Angry	Happy	Neutral

Table 1. Emotions associated to characters

As they are the villains of the tale, for the "stepmother" and "stepsisters" the emotion assigned for the good situations is *angry*. For the hero and victim, the assignment is just the opposite.

3.2.2 Emotions Associated to Actions

The actions are considered as good, bad or neutral situations. When choosing the emotion associated to the message representing the action, the characters involved in it are taken into account. There is a type of action that must be treated in a special way. These are the surprising actions, that are always assigned the *surprise* emotion, not taking into account their arguments. The information about the type of action is specified in the story plan received by the generation module as input.

3.3 Voice Parameters and Emotions

In order to obtain the parameters of the prosody we have analyzed recorded material of tales read by actors. This is how we have identified the relation between the parameters of the voice and the different emotions. We have used actors because we are going to deal with extreme emotions and the employment of actors is the best choice when the aim of the research are extreme emotions.

We have used PRAAT in order to analyze the tales read by actors. With PRAAT we have obtained the pitch base line, the pitch range and the rate related to each of the emotions which take part in the tale.

The aspects of the voice that act as emotional identifiers are: pitch, volume, voice quality and rate. For this research we have assumed that the voice aspects that are necessary for modeling the different emotions are: pitch baseline, pitch range, volume and rate, we have not used the voice quality. The answer to the question "Is voice quality fundamental for the generation of voice with emotions?´´ is not unanimous. We have assumed that the voice quality is not fundamental for the generation of emotional voice.

To obtain the values of these parameters for every emotion, we have consulted Scherer [21] and the results of the analysis of emotional material generated by actors. Finally we have obtained the optimal values through the systematic variation of the parameters during synthesis. Table 2 shows the rules of the synthesizer for the basic emotions.

	Volume	**Rate**	**Pitch Baseline**	**Pitch Range**
Anger	$+10\%$	$+21\%$	$+0\%$	$+173\%$
Surprise	$+10\%$	$+0\%$	$+25\%$	$+82\%$
Happiness	$+12\%$	$+25\%$	$+35\%$	$+27\%$
Sadness	-30%	-10%	-12%	-40%
Fear	$+10\%$	$+12.5\%$	$+75\%$	$+118\%$

Table 2. Configuration Parameters for Emotional Voice Synthesis

3.4 EMOSPEECH 0.2

In the new version of our system, EMOSPEECH 0.2, we have changed some design and implementation aspects. We have used SABLE as a language for the control of the TTS engine, this way the configuration of the different voice parameters for the expression of emotions is independent of the synthesizer engine use. In the previous system we have no control language, we made all the changes directly in FreeTTS.

The input file of our system is a XML file in which every sentence is marked up with the five basic emotions or with the neutral emotion. The file is generated automatically with the modified system for automatic story generation mention before. A sample part of a marked tale is given in Table 3.

Based on the XML file we generate a SABLE file. In order to make this transformation we apply the rules for the synthesizer, which we have obtained from the Scherer researches, the analysis of emotional material and the systematic variation of the voice parameters, to the XML file. For each of the sentences of the previous tale

```
...
<Neutral> Gretel ate the window. </Neutral>
<Surprise> The witch came out of the house. </Surprise>
<Fear> The witch locked Hansel up. </Fear>
<Neutral> Brave Hansel was locked
          in the cage. </Neutral>
<Angry> Hansel made Gretel work
        very hard. </Angry>
<Surprise> She tricked the witch. </Surprise>
<Neutral> The witch was locked in the oven. </Neutral>
<Neutral> Gretel released Hansel. </Neutral>
<Neutral> She was pretty. </Neutral>
<Surprise> Hansel came out of the cage. </Surprise>
<Happy> Pretty Gretel found the treasure. </Happy>
...
```

Table 3. Fragment of a Marked Up Tale

we apply these rules and we generated automatically the SABLE file given in Table 4.

```
...
<MARKER MARK="Neutral">
Gretel ate the window.
</MARKER>

<MARKER MARK="Surprise"> <VOLUME LEVEL="+10%">
<RATE SPEED="+0%"> <PITCH BASE="+25%" RANGE="+82%">
The witch came out of the house.
</PITCH> </RATE> </VOLUME> </MARKER>

<MARKER MARK="Fear"> <VOLUME LEVEL="+10%">
<RATE SPEED="+12.5%"> <PITCH BASE="+75%" RANGE="+118%">
The witch locked Hansel up.
</PITCH> </RATE> </VOLUME> </MARKER>

<MARKER MARK="Neutral">
Brave Hansel was locked in the cage.
</MARKER>

<MARKER MARK="Angry"> <VOLUME LEVEL="+10%">
<RATE SPEED="+21%"> <PITCH BASE="+0%" RANGE="+173%">
Hansel made Gretel work very hard.
</PITCH> </RATE> </VOLUME> </MARKER>

<MARKER MARK="Surprise"> <VOLUME LEVEL="+10%">
<RATE SPEED="+0%"> <PITCH BASE="+25%" RANGE="+82%">
She tricked the witch.
</PITCH> </RATE> </VOLUME> </MARKER>

<MARKER MARK="Neutral">
The witch was locked in the oven.
</MARKER>
...
```

Table 4. Fragment of SABLE file obtained from the mark up tale.

The SABLE file is processed by the FESTIVAL text-to-speech system which returns an audio file in which the tale is read aloud taking into account the different emotions marked in the input text.

4 Evaluation

It is not easy to evaluate this kind of systems, because there is not “good” or “bad” output. In order to evaluate our work we carried out two different tests with two different types of audio files. Fifteen evaluators have taken part in this experiment.

With our evaluation we have tried to measure two main aspects: how well the emotions modeled are recognized by the evaluators and how much the meaning of the reading text influences the emotional recognition.

In order to obtain these measurements we have made some distinctions about the texts that are going to take part and the type of tests that are going to be performed.

The texts that have taken part in the evaluation are two:

- *“Hansel and Gretel”* tale. We have selected this tale because it is a tale that has all the modeled emotions and because it is the tale with most emotional sentences from all the tales generated.
- Sentences without emotional content read aloud with each of the five modeled emotions were reproduced by PRAAT without articulating any word, and these intonated sentences were marked up by the evaluators.

So we measured the two main aspects commented before: on the one hand how well the the evaluators recognized the emotions modeled and on the other hand how much the meaning of the reading text influences the emotional recognition.

We have carried out two types of tests:

- *Test of free choice*: Evaluators can assign to each of the sentences any emotion they consider that best suits the voice they are listening.
- *Test of force choice*: Evaluators have to choose one of the six modeled emotions (five basic emotions and the neutral emotion).

We have made this distinction in order to determine if the emotions are only well distinguished among the five basic emotions or are well distinguished among all the range of emotions.

4.1 Free Choice Results

The graph in Figure 1 shows the percentage of sentences marked with the correct emotion, in each of the tests carried out with the two types of audio files (tale and intonated sentences), grouped by emotions. This figure seems to indicate that the emotions with a high success percentage are Neutral (54%), Fear (46%) and Sad (45%) in the case of *“Hansel and Gretel”* tale and Surprise (38%) in the case of the intonated sentences.

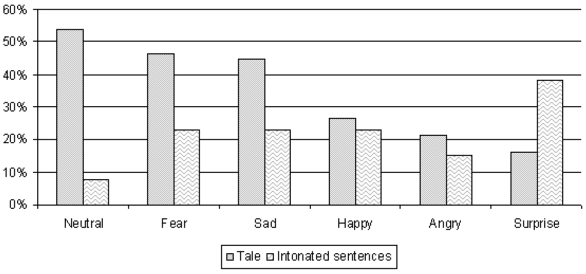


Figure 1. Percentage of sentences marked up with the correct emotion in the free choice tests.

In Tables 5 and 6 the main confusions can be seen.

	Neutral	Sad	Surpr.	Worry	Hysteria	Bored
Angry	×					
Fear				×	×	
Happy			×			
Neutral		×				
Sad	×					×
Surprise	×					

Table 5. Confusion between emotions in the tale in free choice test.

	Ne.	Sa.	Su.	Ca.	An.	Bo.	Ex.	Hy.
Angry	×			×			×	
Fear			×		×			×
Happy			×		×			
Neutral		×		×		×		
Sad					×	×		
Surprise	×	×				×		

Table 6. Confusion between emotions in the intonated sentences in free choice test. The headings correspond, from left to right, to: Neutral (Ne), Sad (Sa), Surprise (Su), Calm (Ca), Anger (An), Bored (Bo), Excited (Ex), Hysteria (Hy)

The graphs in Figures 2 and 3 show the percentage of sentences marked up with an emotion different from the one that the synthesizer is trying to express in the case of the *"Hansel and Gretel"* tale and the intonated sentences.

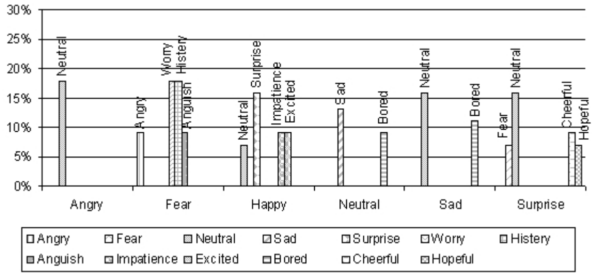


Figure 2. Percentage of sentences marked up with a wrong emotion in the Tale group by emotions.

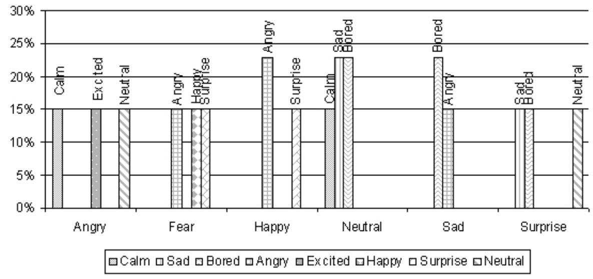


Figure 3. Percentage of sentences marked up with a wrong emotion in the Intonated Sentences group by emotions.

4.2 Force Choice Results

The graph in Figure 4 shows the percentage of sentences marked with the correct emotion. In each of the tests carried out with the two types of texts (tale, and intonated sentences), grouped by emotions. This figure seems to indicate that the emotion with a high success percentage are Sad (77%, 69%), Neutral (65%, 69%) and Fear (64%, 54%) in all the tests.

The graphs in Figures 5 and 6 show the percentage of sentences marked up with a different emotion from the one that the synthesizer is trying to express in the case of the tale and the intonated sentences.

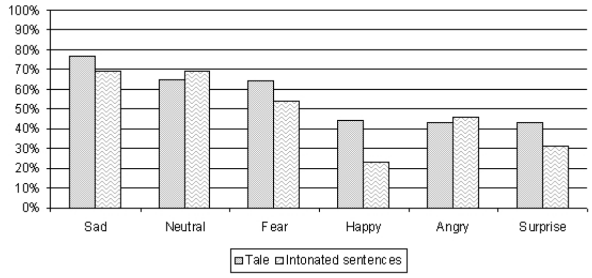


Figure 4. Percentage of sentences marked up with the correct emotion in the force choice tests grouped by emotions.

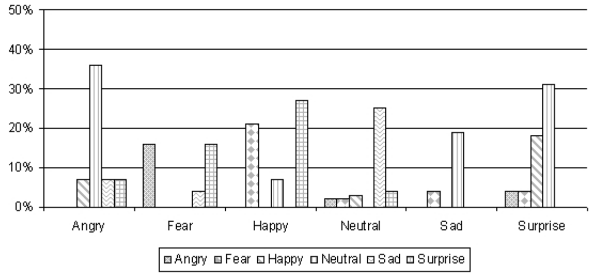


Figure 5. Percentage of sentences marked up with a wrong emotion group by emotions.

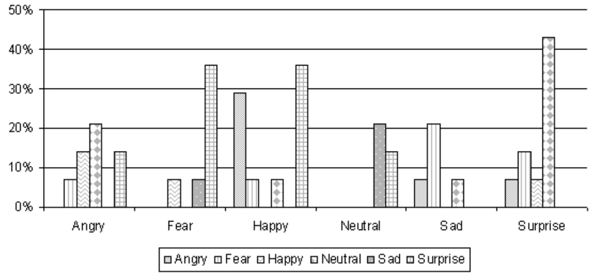


Figure 6. Percentage of sentences marked up with a wrong emotion group by emotions.

In Tables 7 and 8 the main confusions can be seen .

4.3 Conclusions of the Tests

4.3.1 Free Choice

In the case of intonated sentences Neutral is the emotion less recognized. That is because the voice base is very serious and has a low pitch base and pitch range so it confused mainly with the emotion which has these characteristics. Happy has more or less the same results in the two audio files, so we can conclude that the meaning of the text does not influence in this emotion. Surprise has better results in the case of the tale, so we can conclude that the meaning of the text influences in this emotion.

Angry is confused with the excited emotion. Excited is a type of anger, so we can consider that the angry sentences confused with ex-

	Angry	Fear	Happy	Neutral	Sad	Surprise
Angry				×		
Fear	×					×
Happy		×				×
Neutral					×	
Sad				×		
Surprise			×	×		

Table 7. Confusion between emotions in the tale in force choice test.

	Angry	Fear	Happy	Neutral	Sad	Surprise
Angry			×	×		×
Fear						×
Happy	×					×
Neutral					×	×
Sad		×				
Surprise		×		×		

Table 8. Confusion between emotions in the tale in force choice test.

cited are correct. I this way the percentage of correct angry sentences increases to 30%. Fear is confused with worry and hysteria emotions. These two emotions are types of fear so we can consider these sentences as correctly marked and increase the percentage of fear sentences correctly marked to 57%. Happy is confused with Surprise emotion. A surprise can be good or bad. In the case of good surprises the result is a happy emotion. The same occurs with the surprise sentences confussed with the sad emotion. Neutral sentences are confused with sad, calm and bored emotions. This indicates how the base voice is perceived by the evaluators.

If we compare the results of the tales and the intonated sentences in terms of confusion with other emotions, we can see that in both cases the following confusion are presented:

- Angry - Neutral
- Neutral - Sad.
- Surprise - Neutral.
- Happy - Surprise.
- Sad - Bored.

4.3.2 Forced choice

If we compare the percentage of sentences correctly recognized in the tale and the intonated sentences we can see that the percentage decreases in the case of happy and surprise. We can conclude that in these two emotions the meaning influence in a good way. The three emotions more recognized (Sad, Neutral and Fear) are common in the two cases and they are the same as the more recognized in the tale of the free choice test.

If we compare the results of the tales and the intonated sentences in terms of confusion with other emotions, we can see that in both cases the following confusion are presented:

- Angry - Neutral
- Neutral - Sad
- Surprise - Neutral
- Happy - Surprise
- Fear - Surprise

In the case of the sad sentences we can see that they are confused with neutral sentences only in the tale, which indicates that the meaning of the sentences plays a main role. The same occurs with the fear sentences which are confused with angry sentences only in the tale's test.

There are no sentences confused with the happy emotion.

4.3.3 General

If we compare the results of the two types of tests (Free choice and Force choice) we can conclude that the following confusion are present in both cases:

- Angry - Neutral.
- Neutral - Sad.
- Surprise - Neutral.
- Happy - Surprise.

5 Conclusions and Future Work

Expressive characters need an intuitive and simple interface which makes the interaction with the user easy. Communication through the voice is the best solution for this problem. Nowadays the voice generated by synthesizers sounds artificial and this is the main cause of rejection by the users. The success of expressive characters in the everyday life depends on the overcoming of this rejection. In order to obtain a lively synthesizer it is important to generate voice with different emotional states. The generation of emotional voice tries to get emotions clear enough to avoid confusion in the listener.

In this fist approach Sad and Neutral are highly recognized, around a 70% of sentences are recognized, Fear has around a 55% of sentences correctly recognized. Happy and surprise need to be improved because they have a low percentage of recognition, around 30%. These results confirm the ones obtained by [26], [24] and [13]; in general, emotions which can be considered negative are better recognized than emotions which can be considered positive. This is particulary true with the happy emotion, which is the worse recognized in the whole research.

If we compare this approach with a previous one [10], based on the FREETTS synthesizer, we can conclude that:

- As in the previous approach the results obtained in the case of the tale are better than those with the intonated sentences. Is in both cases meaning influences in a positive way the recognition of emotions.
- Surprise and Happy are the emotions less recognized in both approaches.
- Fear is better recognized in the present approach with a 55% of sentences well recognized, as opposed to the 50% of the previous approach.
- Sad is better recognized in the previous approach in which the percentage of sad sentences correctly tagged is around a 100% against the 70% of the previous approach.

We need to explore the new parameters that can be modified with FESTIVAL in order to improve the results. We have to explore the characteristics of the sad emotion in the previous approach in order to apply these characteristics to the previous approach and return to the 100% percentage of success.

There is much work that has to be done, and we are working on different approaches. In future versions we will consider a finer granularity for emotional units. We are considering the use of shallow parsing techniques to determine the different blocks of the sentences, and assigning different emotions to each of this blocks.

We also plan to use the knowledge acquired about the use of emotions in the generation of narrations. In this way, we expect to create more interesting stories, ready to be expressed with emotions.

ACKNOWLEDGEMENTS

This research is funded by the Spanish Ministry of Education and Science (TIN2006-14433-C02-01 project), Complutense University of Madrid and the G.D. of Universities and Research of the Community of Madrid (UCM-CAM-910494 research group grant).

REFERENCES

[1] K. Alter, E. Rank, S.A. Kotz, U. Toepel, M. Besson, A. Schirmer, and A.D. Friederici, 'Accentuation and emotions - two different systems?', in *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 138–142, Northern Ireland, (2000).

[2] F. Burkhardt, *Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren*, Ph.D. dissertation, TU Berlin, 2000.

[3] J. Cahn, 'The generation of affect in synthesized speech', *Journal of the American Voice I/O Society*, (July 1990).

[4] J.E. Cahn, 'Generation of affect in synthesized speech', in *Proceedings of the 1989 Conference of the American Voice I/O Society.*, pp. 251–256, (1989).

[5] W.N. Campbell, 'Databases of emotional speech.', in *ESCA Workshop on Speech and Emotion.*, pp. 34–37, Belfast, (2000).

[6] R. Cowie and R.R. Cornelius, 'Describing the emotional states that are expressed in speech', in *Speech Communication Special Issue on Speech an Emotion*, pp. 5–32, (2003).

[7] R. Cowie and R.R. Cornelius, 'Describing the emotional states that are expressed in speech', in *Speech Communication Special Issue on Speech and Emotion*, (2003).

[8] R. Cowie, E. Douglas-Cowie, and A. Romano, 'Changing emotional tone in dialogue and its prosodic correlates', in *Proc ESCA International Workshop on Dialogue and prosody*, Veldhoven, The Netherlands, (1999).

[9] G. Fairbanks and W. Pronovost, *An experimental study of the pitch characteristics of the voice during the expression of emotion.*, 87–104, Speech Monograph, 1939.

[10] Gervás P. Hervás R. Francisco, V., 'Analisis y síntesis de expresión emocional en cuentos leídos en voz alta', in *In Proceedings of Sociedad Española para el Procesamiento del Lenguaje Natural.*, volume 35, (2006).

[11] A. Gerrards-Hesse, K. Spies, and F. W. Hesse, 'Experimental inductions of emotional states and their effectiveness: A review.', *British Journal of Psychology.*, **85**, 55–78, (1994).

[12] P. Gervás, B. Díaz-Agudo, F. Peinado, and R. Hervás, 'Story plot generation based on CBR', in *12th Conference on Applications and Innovations in Intelligent Systems*, eds., Anne Macintosh, Richard Ellis, and Tony Allen, Cambridge, UK, (2004). Springer, WICS series.

[13] M. Guidetti, 'L'expression vocale des émotions: approche interculturelle et développementale.', in *L'Année Psychologique*, pp. 383–396, (1991).

[14] Lea Leinonen, Tapio Hiltunen, Ilkka Linnankoski, and Maija L. Laakso, 'Expression of emotional-motivational connotations with a one-word utterance', *J Acoust Soc Am*, **102**(3), 1853–1863, (September 1997).

[15] J.M. Montero, *Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano*, Ph.D. dissertation, Escuela técnica superior de ingenieros. Universidad Politécnica de Madrid., 2003.

[16] J.M. Montero, J. Gutiérrez-Ariola, S. Palazuelos, E. Enríquez, S. Aguilera, and J. M. Pardo, 'Emotional speech synthesis: From speech database to tts.', in *In Proceedings of the 5th International Conference of Spoken Language Processing*, volume 3, pp. 923–926, Sydney, Australia, (1998).

[17] S. J. L. Mozziconacci, *Speech Variability and Emotion:Production and Perception*, Ph.D. dissertation, Technical University Eindhoven, 1998.

[18] I.R. Murray and Arnott J.L., 'Implementation and testing of a system for producing emotion-by-rule in synthetic speech', *Speech Commun.*, **16**(4), 369–390, (1995).

[19] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotion.*, Cambridge University Press.

[20] J.A. Russell, 'A circumflex model of affect', *Journal of Personality and Social Psychology*, **39**, 1161–1178, (1980).

[21] K. R. Scherer, *Personality markers in speech*, Cambridge University Press, Cambridge, 1979.

[22] K.R. Scherer, *On the nature and function of emotion: A component process approach*, Scherer and K.R. and Ekman P and editors, Erlbaum, Hillsdale, NJ, 1984.

[23] Marc Schröder, 'Can emotions be synthesized without controlling voice quality?', *PHONUS*, **4**, 37–55, (1999).

[24] H. Siegwart, 'The differential perception of linguistic and emotional prosody: A neuropsychological study.', in *Colloque CERE (Coordination Européenne des Recherches sur l'Emotion)*, Paris, (1990).

[25] J. Stallo, *Simulating emotional speech for a talking head.*, Ph.D. dissertation, School of Computing, Curtin University of Technology., 2000.

[26] H.G Wallbott and K.R. Scherer, 'Cues and channels in emotion recognition.', *Journal of Personality and Social Psychology*, **51**, 690–699, (1986).

Kiwi: An environment for capturing the Perceptual Cues of an Application for an Assisting Conversational Agent

Jean-Paul Sansonnet, David Leray[1]

Abstract. This paper gives an overview of an architecture attempting to provide dialogical assistance between a user and an assisting conversational software agent. We will focus on the *perceptual issues* expressed by *ordinary novice users* when they need some assistance about an application. This paper proposes a methodology to capture in a structural symbolic model some kinds of perceptual phenomena that are part of the natural language requests expressed by ordinary users.

1 INTRODUCTION

1.1 The Function of Assistance

We are focusing on the problem of the "function of assistance" that can be provided by dialogical software agents to ordinary users. By *ordinary novice user*, we mean a non expert person using a software application. An example of such users could be the average Net surfer swapping between websites in order to use sporadically web-services (buying an airplane ticket…) or producing personal web-content (sharing photos, videos…). Inevitably, because of the complex nature of the new websites and commercial applications, these users will face difficulties in the achievement of their objectives. That situation can lead the users into a cognitive distress, with a significant negative impact for the application providers.

An interesting characteristic of this situation is the tendency of ordinary users (as opposed to computer skilled ones) to express their problems by using the natural language modality (see the "thinking aloud" effect [1]). Moreover, because these situations arise mainly because of the lack of knowledge about the software, this leads to linguistic difficulties such as: user-specific vocabulary, degraded spelling (for typed requests) or degraded prosody (for oral requests), bad categorizations etc., making this type of requests really difficult to process and to interpret.

Among those, the *categorization issue* is the more difficult to handle because the requests are likely to contain references to the entities of the application (objects, aggregates, behaviors …) expressed in terms significantly different from the software technical ontology. That is to say, someone talks about what she/he *perceives* with respect to his own knowledge and lexicon (e.g. the "Ozkan experiment" [2] can be seen as an example of such idiosyncrasies) thus leading to a "cognitive drift" between the user and the assisting system. In the following, we focus on the class of assisting requests that exhibit these phenomena: the so-called *perceptual requests*. The section 1.3 illustrates what is the cognitive drift on an example taken from a real website.

1.2 The Daft project

The problem of the perceptual requests is part of the Daft project at LIMSI-CNRS. A main goal of the Daft project [3], is to increase the level of genericity of assisting agents. The term 'genericity' is considered 1) at the linguistic level, along with J.F. Allen [4,5] pointing out the inherent specialization of current dialog-based systems; and 2) at the software level, following a middleware methodology, with the introduction of an intermediate representation, a *mediator*, between the application and the assisting tools.

Figure 1 illustrates the general architecture that implements the three functions an assisting agent must fulfill: *utterance analysis*: (the Dialogical Agent), *request solving* (the Rational Agent) and *presence and presentation* (the Embodied Agent) [6].

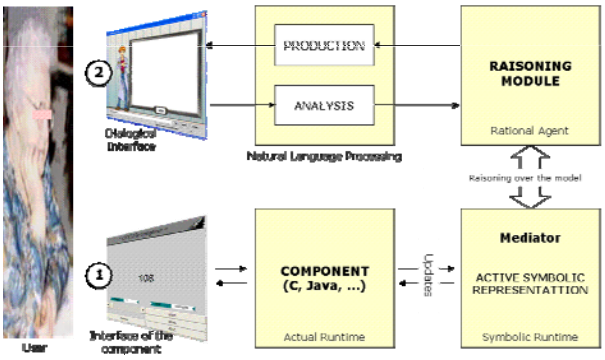
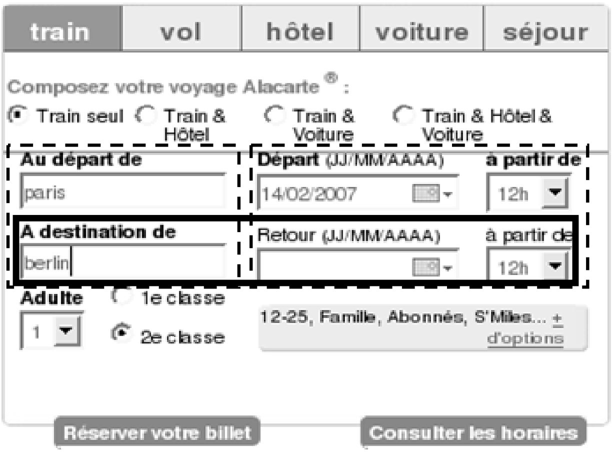


Figure 1. The Daft general architecture: when a novice user has a difficulty using the application interface (1), she/he prompts the Dialogical Agent (2) with a natural language request. Then the Analyzer builds a formal representation of the request. The Request is sent to the Rational Agent which must resolves it over the active symbolic representation of the application (the mediator). This produces a formal answer which is used by the Embodied Agent (2) for providing the user with a natural answer.

LIMSI-CNRS, France, email : {jps, leray}@limsi.fr

1.3 An example of semantic drift

The screen shot given in figure 2 shows a part of the web site of the French national railway travel agency (http://www.voyage-sncf.com). This area is dedicated to the useful information for the preparation of a private travel (this area is typically overcrowded with promotional offers and advertisements). First, suppose that a novice user wants to travel from Paris to Berlin: he has to inform the fields, especially those concerning the departure and return dates. Now suppose that he makes an error while typing the return date and he gives a month anterior to the current month but without modifying the year (e.g. 27/**12**/06 => 27/**01**/06 instead of 27/**01**/07): the travel application detects that the return date is in the past an makes an automatic correction. Then, the user doesn't understand why the month has (suddenly) been coerced!



Composez votre voyage Alacarte® :

● Train seul ○ Train & Hôtel ○ Train & Voiture ○ Train & Hôtel & Voiture

Au départ de: paris

Départ (JJ/MM/AAAA): 14/02/2007 à partir de: 12h

A destination de: berlin

Retour (JJ/MM/AAAA): à partir de: 12h

Adulte: 1

○ 1e classe ● 2e classe

12-25, Famille, Abonnés, S'Miles... + d'options

Figure 2. A view from the web site http://www.voyage-sncf.com. The internal structure of the site (i.e. as defined by the HTML code) is framed with a dashed rectangle whereas the user private perception of his area of interest is in bold.

When confronted with this behavior, the user can type the following request into a dialog box: «*pourquoi jpeux pas rentrer de berlin le 1er javier ??*» (TRANSLATION: "why is that I cn't return from berlin on jnary the first?" — Indeed, orthographic and syntactic errors are quite frequent, thus entailing a robust approach to the requests' analysis; this point will not be detailed further in the paper). Consequently the user is afforded to call the field « A destination de » (destination) by its content and in the same time to aggregate it with the field « retour » (return). In other words, the user considers that the row of the graphic components [« a destination de » + « retour » + « à partir de »] as an cognitively homogeneous aggregate dealing with the concept /information about the destination/. However, the internal structure of this web page area (i.e. as defined by the HTML code that is displayable in the navigator) does not correspond with such a semantic organization: instead, « au départ de » (departure) and « à destination de » (destination) are aggregated in the same structure. We can see that a group is aggregated along a *geographic* criterion whereas the other is aggregated along a *time* criterion.

An assisting agent which tries to understand the user's request and which exploits only the accessible information from the internal structure (i.e. the HTML code) cannot find the referred to group within the application because it has no actual implementation. We need then to introduce another representation, *a middleware*, between the assisting agent and the application which can deal with the perceptual view points of the users, especially novice ones. Hence, the key issue is the question of the acquisition of semantic knowledge about the structure and the functioning of the applications in the consumer field. Moreover, we want to develop a framework which is as generic as possible, i.e. that can handle new applications with little adaptation efforts.

2 THE KIWI FRAMEWORK

2.1 A methodology for dealing with perceptual requests

The problem of the perceptual requests has to be considered at three levels: the request level, the reasoning level and the modeling level.
- *The request level* is about the detection of the perceptual information embedded in the requests. For this purpose, we are currently building a corpus (up now, 11 000 requests have been registered in various assisting situations) that will be used for designing the ontology of the semantic classes characterizing the function of assistance.
- *The reasoning level* (rational agent) can be viewed as a collection of assisting heuristics based on the ontology defined above. For each heuristic, the assisting information has to be extracted from the model of the application.
- *The modeling level* is the symbolic representation of the runtime of the application. In order to map the user's concepts, expressed in his request, with the actual entities represented in the model we have to make available the annotation of perceptual cues.

This leads to two main questions: a) how can we represent these structures? and b) how can we provide them with pertinent semantic annotations (i.e. meta data dedicated to assistance)? For the representation of the structures, we use special components, named *aggregate*s that are attached to the model.
As for the semantic annotations to be attached to the aggregates, two sources are possible:
1) From external knowledge sources, when the application has already been released: UML schemes, Frequently Asked Questions, information attached to the source-code and/or Graphical User Interface (GUI) schemes, if they are available;
2) From the design stage, by involving directly the designer in the process of the validation of the assisting annotations. Roughly speaking, this approach consists in providing the designer with a WYSIWYG (What You See Is What You Get) [7] design environment where he is asked to carry on his usual design task and he is prompted to explicit some information about what he is currently doing (see figure 6)

Both these approaches have their own limitations: the first case assumes that all the required information has been made explicit and is computationally available. However we have to face here the lack of documentation created for assistance purposes. Moreover, to be efficient, this case requires a human operator committed to the task of validating the mapping between the semantic and the perceptual knowledge. The second case lacks the formal knowledge about the functional capabilities of the application and moreover makes the design process very hard. Both these approaches need a human operator that can bring its semantic expertise. This is the reason why we propose a software framework (called Kiwi) that integrates the two approaches and that is operable by a human designer during the designing process.

2.2 Purpose of the model

The objective of the Kiwi framework is to provide a computational environment dedicated to the *study* of the extraction and the construction of the so called *assisting symbolic models* from the applications.

There are several reasons why an assisting agent cannot directly browse the runtime of an application in order to resolve users' requests and therefore needs an intermediate form:
- The multiplicity and the diversity of the notations and the programming languages actually employed make it difficult to develop symbolic reasoning tools adapted for each version, so we need a model that can represent different kinds of applications in a uniform way;
- The runtime of an application is not easily introspectable, even if it is not mere binary code (like with the JavaBeans technology that provides get/set functions), so we need a symbolic representation that can be easily read and, when needed, modified;
- As seen above, the cognitive drift associated with novice users is also a key issue that requires: a) to collect meta information about the structure and the behavior of the application and b) to make it possible to achieve representational transformations (e.g. to compute various kinds of perceptual aggregates).

During the process of assistance, the symbolic model plays the role of the application for the assisting agent. This entails the three following requirements:
a) it must be maintained in a dynamic state that reflects in real time the current state of the application itself. That is, the model must evolve together with the current state of the application: it cannot be a simple XML-like static representation of the application at t_0.
b) it must reflect the formal representation of the application as designed by its programmer: such event triggers such function, such function can be activated only when such condition holds… This representation can be mainly extracted from the code of the application.
c) it must reflect the (possible) mental representations of the users: what are the tools available to achieve his objectives? (such button displays such abject, such list contains such information, …). This kind of mental representations, linked to human perceptual attributes, are mainly non accessible for the assisting agent (they cannot be computed from the formal representation) but they play a major role in the way users express their natural language requests. This is the reason why we need a mechanism to exhibit (to retrieve from the code) and to register (to certify and to annotate by the designer) the *pertinent* perceptual structures of a given application which are not necessarily explicit in the formal representation derived from the code.

About the users' requests: In order to attempt an experimental characterization of the Function of Assistance, we developed a first version of our framework and we registered, over a period of two years (June 05 – Sept 07), a corpus of ~11 000 requests. This corpus is made of three sub-corpora: 1/3 was registered with ordinary users in front of small applications [6]; 1/3 was built from requests thesaurus [7]; 1/3 was collected from the FAQ (Frequently Asked Questions) of two well-used text editing applications (Microsoft-Word and Latex). A first analysis of the part registered with ordinary users showed that it can be in turn divided into four sub corpora, each corresponding to a particular activity as shown in figure 3:

1) Control activity: sub corpus made of direct controls, to make the agent interact himself directly with the application software in which it is embedded;
1) Direct assistance activity: sub corpus gathering help requests explicitly made by the user;
3) Indirect assistance activity: sub corpus made of user's *judgments* concerning the application that are actually implying the fact the user is in need of assistance; it certainly requires the system to use pragmatics to detect the implicit meaning;
4) Chat activity: sub corpus with all other activities which are not in direct relation with the application and often oriented towards the agent itself.

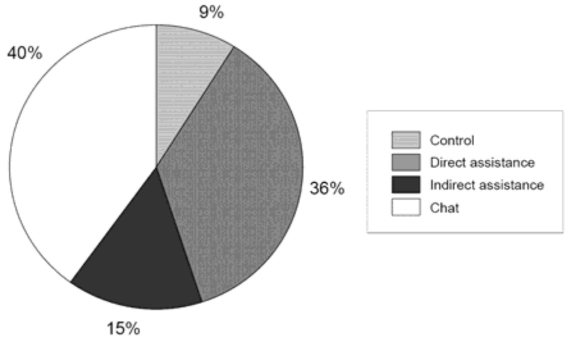


Figure 3. The four main conversational activities in the Daft corpus.

We can see that a large part of the corpus is related with the chat activity. This is because we used embodied virtual characters (the LEA technology — LIMSI Embodied Agents [8]) that afforded the users to interact with the system even in the absence of assisting needs. In the following, we will only consider the direct-assistance activity which represents 36% of the registered utterances. An exploration of the nominal groups of the direct-assistance activity requests shows that the mental representations of the users are expressed in two main categories:
- Those where the referential expressions are directed towards an entity accessible within the representation of the application. In this case, it is possible to build a formal referential expression that can be sent to a conventional reference handling operator in charge of the retrieving process;
- Those where the referential expressions have no direct counterpart within the representation of the application (as seen in example §1.3). Indeed, as we suppose that the user truly believes that the 'thing' he refers to is an actual entity of the application, we have to deal carefully with these references and we cannot just discard them as erroneous.

2.3 Structure of the assisting models

The structure of the assisting symbolic model of an application designed within the Kiwi framework contains typically:
1) A description of the internal structures of the application, that its organization in terms of its atomic components and of the links between these components;
2) A description of the functions that are available in terms of the operations that are allowed from the user viewpoint: it is not a description of the internals of a given function but rather its

'user-manual' description including the purpose, the preconditions and the resulting state.

As the model is based on a set of atomic components, the Kiwi framework provides a library of the handled components that is organized as an ontology: the basic characteristics of the components, the possible links between them and the natural language lexicon typically associated with them are described in the Kiwi ontology.

2.3.1. *The ontology of the components*

The Kiwi components are organized within a global type hierarchy. All the internal entities are considered as *concepts* in the Kiwi ontology. Contrary to more conventional implementations where events are considered as functions or results from function calls, here an event is represented in the same form as a widget (i.e. objects visible on screen); hence, we can attach semantic attributes to events and to widgets in a uniform way.

There are three families of components:

a) The graphic components*:* they are conventional widgets and the GUI events (punctual actions with a visible effect on screen)

b) The structural components*:* they are data structures (scalars, lists …) and the operations associated with these data structures;

c) The aggregated components: they are in turn, divided into two sub categories:

1) The visual aggregates: they group a set of widgets that exhibit a semantic homogeneity. They are typically synthesized (automatically computed) by the modeling agent either from the existing GUI or from some instructions provided by the human designer of the application.

2) The functional aggregates: they usually define a change of state in the application via the specification of four elements: a source, a destination, an effect and one (or several) pre-condition(s).

2.3.2. *Meta data of assistance*

In the code of the application, the components and the links between them can be automatically retrieved but then no particular semantics can be associated with them. This is the reason why the *meta data of assistance* (later referred to in this paper as the *metadata*) play a double part: first, they provide a lexicon in natural language that can be associated with the components and second, they provide a basis on which the library of the functionalities of the application can be developed. The metadata can take the form of a simple marker and/or a complete sentence (a kind of a gloze) that can be associated with any entity in the model: it can be a simple component or any arbitrary group of components. The metadata are provided either by the designer while he interacts with the Kiwi system during the design phase or from static knowledge sources (FAQ, textual documentation) under the supervision of a human operator using the same system.

The figure 4 illustrates a simple excerpt taken from a GUI. The Table 1 sums up the information taken from the Kiwi ontology that are needed to model this simple situation.

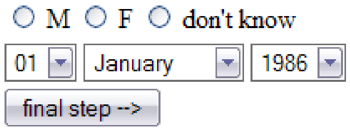


Figure 4: An excerpt of a GUI: a list of several possibilities is proposed to the user who is prompted to choose only one of them. A button is added for the user to signal to the system that his choice has been completed.

Table 1: Correspondence between the Kiwi ontology and the component synthesized from the excerpt of the GUI in figure 4.

Kiwi Ontology	**Component description**

SIZE part-of WIDGET
COLOR part-of WIDGET
RADIOBUTTON isa WIDGET
BOOLEAN isa DATASTRUCT
	COMP[
	ID=radiobutton#1
	TYPE = RADIOBUTTON,
	DISPLAY=SELECTED
	COLOR = GRAY,
	COMMENT= « choice of the male gender »
	]

RADIOBUTTON is-equal BOOLEAN
SHIFT isa ACTION
SHIFT part-of BOOLEAN
	COMP[
	TYPE=FUNC,
	SOURCE= radiobutton#1
	TARGET=bool#1
	ACTION=CLICK
	EFFECT=SHIFT
	]

2.4 Model synthesis methodology

Our methodology for the building the assisting models from the applications is divided into three main steps:
1) Acquisition of the structural model: this representation is extracted from the code of the application. It must contain the minimal information required to build a functional model;
2) Generation of the perceptual model: once the structural model built, it can be used by the modeling agent to synthesize the perceptual entities that are inherent to the graphical layout of the application but were not explicitly provided by the designer.
1) Acquisition of the semantic knowledge: finally, all the entities of the model must be annotated with metadata.

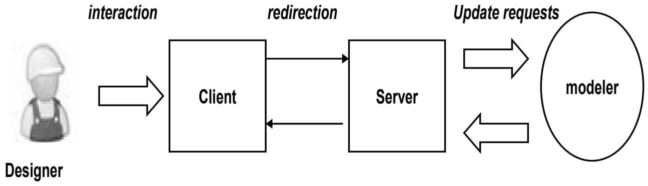


Figure 5. General architecture of the Kiwi framework.

The Kiwi framework has been developed to support and to study the feasibility of our methodology for the building of assisting models from the applications. The general architecture, given in figure 5, is decomposed into three main processing modules: the client, the server and the modeler (or modeling agent).
- The client, developed in JavaScript, supports the interactive web-based designing interface. This interface is divided into several frames, each one being associated with one of the viewpoints discussed before. The designing interface has two purposes: first it is used to display the information issuing from the assisting environment and second it is used to register the actions of the designer in order to build the model.
- The server, developed with a JSP servlet and the Tomcat technology, is just used to transfer the client information to and fro the modeler.
- The modeler, developed with the Wolfram Research symbolic environment Mathematica 5.2, is the core of the Kiwi architecture. The modeler makes handles the Kiwi ontology in order to display the information that are significant at a given time (e.g. what operations are associated with a given component). But the main role of the modeler is to build an assisting model of the application from the information provided explicitly or implicitly by the client. In the following section we give a complete example of this process.

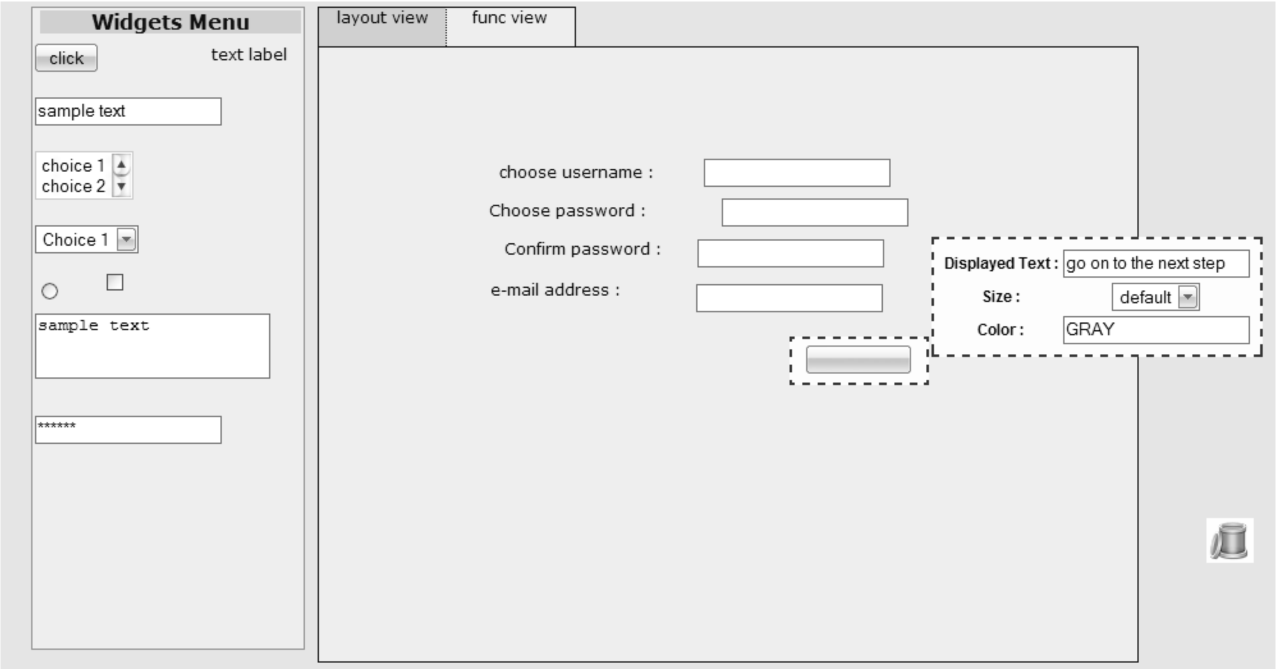
2.5 An example of model synthesis

When the designer is logged to the Kiwi interface, a first communication is performed between the client and the server in order to initialize the communication with the modeling module. The GUI of the client displays two separate developing viewpoints on the currently designed application: the graphical view and the functional view.
- To the graphical view (figure 6.a) is associated a menu which represents the complete set of widgets that are available. These widgets will act, in the model, as entry points to the application. The designer can chose a widget and can drag&drop it on the main display area which represents the main window of the future application. For the time being, we are restricting ourselves to applications with a single window.
- The passage to the functional view (figure 6.b) has the effect of blurring the view of the components. When the designer moves the mouse over a component, the list of the available action associated with the component is displayed. Then the designer needs only to move the chosen action towards the target component.

The first aspect of the communication with the server is mainly related to the consultation of the Kiwi ontology (which is part of the modeler module) in order to display the assisting information on the designer interface. For example, the menu that proposes various available widgets or the menu that proposes several actions for a given widget are automatically built from the information contained in the ontology. A client request (not to be confused with a novice user request) can achieve a filtering of the components in the ontology along their attributes: type, hierarchical position, or shape, color etc.
For example, a request like: SELECT(*type* = *widget*) just asks the modeler about components of type *type* and widget *widget*. The modeler replies with a list of the components matching this description. Then two situations can arise:
1) When the reply corresponds to a DOM object, it is displayed in the interface;
2) Else a graphic box is displayed with the reply as a label (see display of an action: figure 6.b)
In each case, the client has the responsibility of displaying correctly the components in the interface.

A second aspect of the communication is related to the organization of the components. Indeed before we can try to synthesize perceptual groups we need a base. This base can be provided by the designer: each time a new component is drag&dropped or a component is modified on the client window, a communication with the modeler is established in order to register the new item or the modification.

The third aspect of the communication is related to the validation and the semantic annotation of the entities that the modeler has automatically synthesized during the interaction with the designer. First a validation a posteriori by the designer is necessary because all the aggregates are not significant. The modeler sends the list of the synthesized structures on the client screen (figure 6.c-6.d): the designer can c) discard the entity as irrelevant or d) can affix to the structure an annotation in natural language that describes the semantics of the item. With this manual phase, we are sure that the aggregates that are produced in the assisting model are validated and annotated by the competent authority.



Displayed Text : go on to the next step

Size : default

Color : GRAY

Figure 6.a The Kiwi graphical interface

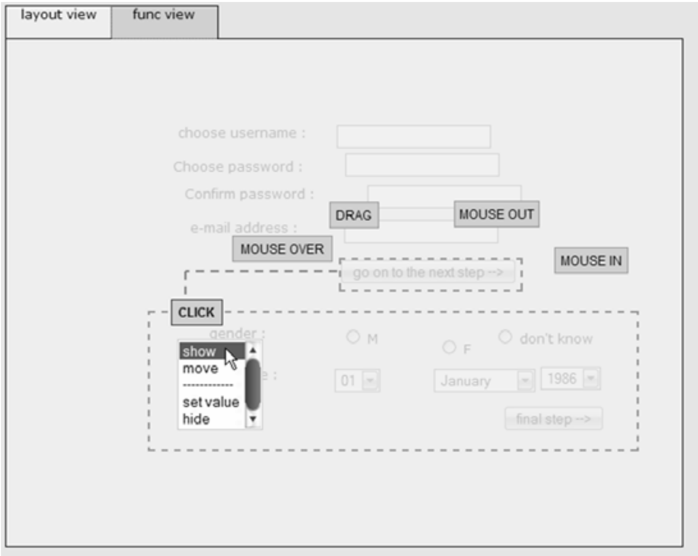
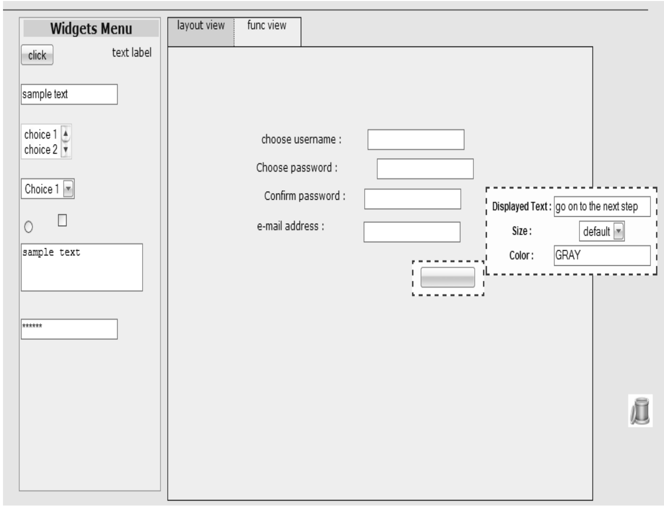


Figure 6.b Definition of a function



Displayed Text : go on to the next step

Size : default

Color : GRAY

Figure 6.c Annotation of a significant perceptual aggregate

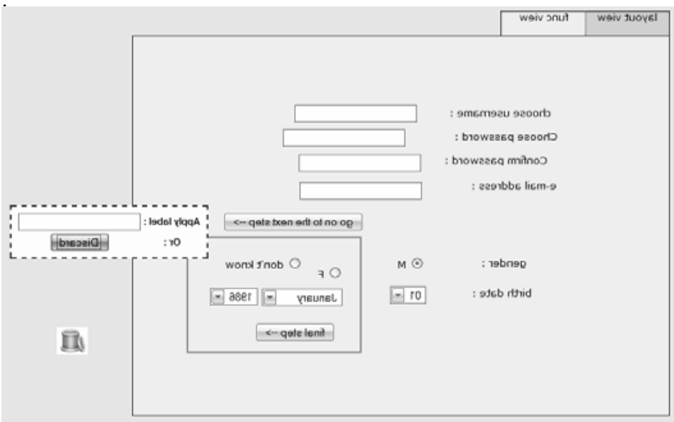
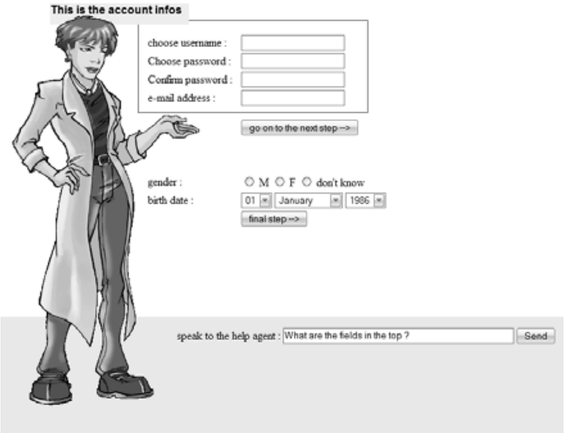


Figure 6.d Annotation of a **non**-significant perceptual aggregate



speak to the help agent : What are the fields in the top ?

Figure 6.e Final application with the assistant agent

3 EVALUATION

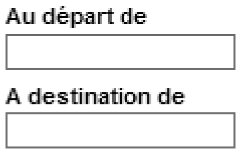
3.1 Methodology of evaluation

The process implemented in the Kiwi framework is quite complex and is based on several principles: first we try to build a symbolic representation of an application that can support effectively the Function of Assistance. On the other hand we try to integrate into this representation perceptual structures that can prevent the cognitive drift between the (novices) users and the implementation. This leads to consider the following questions:
1) Are we able, in general, to build a model so that a rational agent can reason upon it and provide users with relevant answers? We were able to establish in a previous work [9] that it is possible to extract the formal part of an assistant model from the source code of the application (code source in Java);
2) Is the algorithm of the modeling agent capable of exhibiting all, some, the relevant ... implicit perceptual structures? This key point will be discussed below while taking again the example detailed in § 1.3;
3) Are the annotations registered during the design process actually useful and can they significantly increase the quality of the Function of Assistance for an ordinary user? This very important issue will be submitted to the test in the next months, so we cannot discuss it here for lack of experimental data.

Now we want to build an assisting model from the HTML code of the application that is currently built on the web page of the Kiwi framework. Then we will try to evaluate the perceptual relevance of the aggregated structures the modeling agent will synthesize. We take as an example a copy of a part of the web page of the French voyage-sncf.com train travel reservation system.
First, the structural model can be acquired from the DOM structure (Document Object Model – is the standard ontology of the web page by the W3C consortium) of the area corresponding to the screen shot shown in table 2. A table is constructed where all the components visible on screen are listed. Le HTML tags of type 'container' like **`<div>`**, **`<table>`**, etc. are appended to the list only when their **`border`** attribute in the CSS description (Cascading Style Sheet) is specified in the HTML code. The client also registers the information related to the physical position of the components on the screen together with their size, their color etc. The table 2 shows the initial structure that is automatically extracted.

Table 2. On the left, is an excerpt of the HTML code of the web page voyage-sncf.com. On the right, is the formal structure extracted automatically by Kiwi, in order to build a simple structural assisting model.

HTML structure	Model structure

… <div id="od_train">
  <p class="input required"
    id="fi_ORIGIN_CITY">
    <**label** for="ORIGIN_CITY"
    title="départ">
      Leaving from
    </label>
    <**input** name="ORIGIN_CITY"
    id="ORIGIN_CITY"
      tabindex="1" type="text" />
  </p>
 	
  <p class="input required" id=
    "fi_DESTINATION_CITY">
    <**label** for="DESTINATION_CITY"
    title="destination ">
      Going to
    </label>
    <**input** name="DESTINATION_CITY"
      id="DESTINATION_CITY"
      tabindex="2" type="text" />
  </p>
</div> …
Note : for clarity, the tags <div> and <p> from the original HTML code have been omitted here because they do not support the 'border' attribute of the CSS.
	L = {
	    label#1,
	    ORIGIN_CITY,
	    label#2,
	    DESTINATION_CITY
	}
	
	Au départ de
	A destination de

3.2 Kiwi Aggregating algorithm

Some studies in psychology [10] have determined five criteria upon which are based the human visual perception system when it is in the process of categorizing perceptual aggregates: « proximity, similarity, good continuation, symmetry, closure ». The computational algorithms that have been proposed to implement these cognitive principles have mainly emphasized on the first three ones [11].
We use an algorithm similar to the Thorisson algorithm but it is applied to GUI objects where each one has a particular semantics that can influence the potential groupings. Our algorithm takes advantage of these semantic specificities.

```
Algorithm : Computation of the perceptual aggregates
begin
$V_0$ := {list of the components}
$V_1$ := computeNeighbors($V_0$)
i :=1
while $V_i$ != $V_{i-1}$ do
        foreach  $v_j$ in $V_i$ do
                  $v_j$ := applyHeuristic(CRITERION,vj)
$v_j$ := applyPattern(CRITERION,vj)
        endfor
endwhile
end
```

The computation of the neighbors **`computeNeighbors`** implements the *proximity* criterion; the application of the heuristics **`applyHeuristic`** implements the criterion of *similarity* and the application of the patterns **`applyPattern`** the criterion of *good continuity*. The criteria, passed as arguments to these functions, are those derived from the knowledge attached to the components: the type, the perceptual primitives like the color, the size etc. More especially, for the pattern, the idea is to try to find among a group of components a repetitive pattern with a higher priority given to some specific components (repetition of buttons, radio buttons, ticking boxes, hypertext links, …). Then two kinds of action are possible:
- *The fusion:* two or more components are considered as one and the same cognitive entity;
- *The aggregation:* two or more components are aggregated in a group.

3.3 Evaluation results

The table 3 shows the results obtained on the example. The structures that have been synthesized by our algorithm are compared with their counterpart in the DOM structure, when 1) the

DOM counterpart actually existed and 2) when they were considered relevant by the designer and finally 3) a semantic annotation was provided by the designer.

These results were obtained on a single case and they cannot be generalized without further experiments. However this simple case was taken as a part of a real consumer application and it exhibits some interesting phenomena: first, the fusion of components is considered relevant (indeed, one can see that fusion is present in the HTML code but **not according to the principles of human cognition**). Second, the aggregations of components are also considered relevant, as shown by the fact that it was possible for the designer to affix a semantic annotation for all the aggregations synthesized by the Kiwi algorithm. This is important because it has not been possible to retrieve these aggregated structures in the HTML code! Conversely some aggregated structures of the HTML code are not present in the list of the synthesized aggregates (this being the very source of the cognitive drift mentioned in the §1.3).

Table 3 Relation between the DOM structures and the model aggregates structures. The column ‘occurrences’ gives the number of the structures synthesized by the Kiwi algorithm. The column ‘HTML counterparts’ gives the number of the synthesized structures for which a HTML counterpart was retrieved, the column ‘DOM tag counterparts’ gives some examples of tags associated with these operations and finally the column ‘Annotated relevant structures’ gives the number of synthesized structures that were considered relevant by the designer and subsequently annotated.

Operations applied by Kiwi	**Synthesized occurrences**	**HTML counterparts**	**DOM tag counterparts**	**Annotated relevant structures**
Fusions of components	13	13	<p>, <span>, <label>	13
Aggregations of components	5	3	<div>, <ul>	5

4 CONCLUSION

To summarize, this paper is about the necessity of the integration of perceptual cues into the formal representation of an application in order to improve the efficiency of an assisting conversational agent by increasing the credibility of his answers thus increasing the global credibility of his behavior. The key point of our approach is to let a human operator to be involved in the definition of the assisting content, integrated in a formal representation of the application. This is important because perceptual features a) need to be exhibited and annotated in the assisting model and b) this must be achieved by the ordinary users not by the professional designers. Moreover, our approach tries to explore the involvement of ordinary users in the assistance design process by letting them dynamically access to the modeling information and add their personal semantic annotations, in a gradual process.

REFERENCES

[1] Wright, P. C. and Monk, A. F. 1990. The use of think-aloud evaluation methods in design. *SIGCHI Bull.* 23, 1 (Dec. 1990), 55-57. DOI= http://doi.acm.org.gate6.inist.fr/10.1145/122672.122685

[2] Ozkan, Nadine, Paris, Cecile, Balbo, Sandrine (1998): Understanding a Task Model: An Experiment. In: Johnson, Hilary, Nigay, Laurence, Roast, C. R. (ed.): Proceedings of the Thirteenth Conference of the British Computer Society Human Computer Interaction Specialist Group - People and Computers XIII. August 1-4, 1998, Sheffield, UK. p.123-137. see http://www.loria.fr/projets/asila/corpus_en_ligne.html#ozkan

[3] Sansonnet J.-P., “Composants dialogiques génériques: perspectives et méthodes pour une approche intégrée des outils assistants langagiers et de la programmation objet, Conférence", "Langages et Modèles à Objets", LMO 04, Lille (2004)

[4] Allen J.F., Byron D.K., Dzikosvska M.O., Fergusson G., Galescu L., and Stent A., « Towards conversational Human-Computer Interaction », AI magazine, 2001.

[5] Allen, J.F. et al, “The TRAINS Project: A Case Study in Defining a Conversational Planning Agent”, Journal of Experimental and Theoretical AI, 1995.

[6] Sansonnet J.-P., Leray D., Martin J.C., Architecture of a Framework for Generic Assisting Conversational Agents, IVA 2006, Marina Del Rey

[7] Steven J. Molinsky and Bill Bliss, 1994. Inventory of functions and conversation strategies, pages 177–187. Prentice Hall, January.

[8] Buisine, S., Abrilian, S., Martin, J.-C. (2003). Evaluation of Individual Multimodal Behavior of 2D Embodied Agents in Presentation Tasks. Proceedings of the Workshop "Embodied Conversational Characters as Individuals", 2003, Melbourne, Australia, in conjunction with the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'2003).

[9] D. Leray, J-P. Sansonnet, Librairie de Widgets Dialogiques pour un Agent Conversationnel Assistant, Short paper at IHM05, Toulouse, 27-30 sept 2005

[10] Wertheimer M., Untersuchungen zur Lehre von der Gestalt. Psychologische Forschung, 4, 301-50.Translation in W. D. Ellis (ed.), A Source Book of Gestalt Psychology. New York: H. B. J., 1938.

[11] Thorisson K.R., Simulated Perceptual Grouping: An Application to Human-Computer Interaction, Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society, Atlanta, Georgia, August 13-16, 876-881, 1994.

[12].Butler Lampson, *Bravo Manual* (in *Alto User's Handbook*, Xerox PARC, September 1979, pp 31-62)

[13] Maes P., Agents that reduce workload and information overload, Communications of the ACM, 37(7), 1994

[14] Lu, S., Paris, C., and Linden, K. V. 1999. Toward the Automatic Construction of Task Models from Object-Oriented Diagrams. In *Proceedings of the IFIP Tc2/Tc13 Wg2.7/Wg13.4 Seventh Working Conference on Engineering For Human-Computer interaction* (September 14 - 18, 1998). S. Chatty and P. Dewan, Eds. IFIP Conference Proceedings, vol. 150. Kluwer B.V., Deventer, The Netherlands, 169-189

[15] Moriyon, R., Szekely, P., and Neches, R. 1994. Automatic generation of help from interface design models. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating interdependence* (Boston, Massachusetts, United States, April 24 - 28, 1994). B. Adelson, S. Dumais, and J. Olson, Eds. CHI '94. ACM Press, New York, NY, 225-231. DOI= http://doi.acm.org/10.1145/191666.191751

[16] Aronsson, Lars (2002). Operation of a Large Scale, General Purpose Wiki Website: Experience from susning.nu's first nine months in service. 6th International ICCC/IFIP Conference on Electronic Publishing, November 8, 2002, Karlovy Vary, Czech Republic.

[17] Coombs, Michael J., Gibson, R., Alty, James L. (1982): *Learning a First Computer Language: Strategies for Making Sense*. In International Journal of Man-Machine Studies, 16 (4) p. 449-486

eDrama: Facilitating Online Role-play using Emotionally Expressive Characters

Kulwant Dhaliwal*, Marco Gillies[†], John O'Connor*Amanda Oldroyd[‡], Dale Robertson[‡], Li Zhang[☼]
*Hi8us Midlands, Unit F1, The Arch, 48-52 Floodgate Street, Birmingham, B5 5SL
[†]Department of Computer Science, University College London, Malet Place, London WC1E 6BT
[‡]BT Group CTO, Adastral Park, Ipswich, Suffolk, IP5 3RE
[☼]School of Computing and Technology, University of East London, Docklands Campus, 4-6 University Way, London E16 2RD

1 Introduction

This paper describes the results of a user study of a multi-user role-playing environment '*edrama*', which enables groups of people to converse online, in scenario driven virtual environments. Hi8us' *edrama* system is a 2D graphical environment in which users are represented by static cartoon like avatars.

An application has been developed to enable the integration of the existing *edrama* tool with several new software components to support avatars with emotionally expressive behaviours, rendered in a 3D environment.

In this paper we describe a user trial that demonstrates that the changes made improve the quality of social interaction and users' sense of presence.

2 Overview of *edrama*

In 1999, Hi8us Midlands began developing *edrama* — online multi-user role-play software that could be used for education or entertainment. This first major *edrama* project received substantial support from NESTA, the National Endowment for Science Technology and the Arts. *edrama* became one of Nesta's flagship projects.. In this software young people could interact online in a 2D flash based interface with others under the guidance of a director. The interface incorporated 2D static avatars and a text chat interface, with different photographic backgrounds as scenes to set the role-play. Over the years, edrama has been further developed by Hi8us Midlands and adapted for the delivery of commissions for a range of uses, such as Careers Advice and Creative Writing. The Dream Factory was a version of *edrama* developed for the University for Industry and piloted and tested Connexions-Direct advisers and young people at Skill City in Salford. The software continues to be developed as a 2D application, to be used online

The benefits of Hi8us' *edrama* 2D software include:

- Its use of drama to deliver almost any type of training, which is engaging and entertaining
- It's collaborative and multi-user, allowing people to learn together remotely, cutting out any geographical, social and cultural barriers
- One of the advantages of the 2D version is that anyone with an internet connection, on any platform, can access it – once loaded, it can even be used via a 56k modem and has been trialled in situations such as these
- Role-play in person is often an area that many shy away from due to inhibitions around performing, but *edrama* allows users to remain anonymous, allowing them to express themselves with out being seen — this is particularly useful with young people who may be afraid of expressing their views in front of their peers
- *edrama* can easily be customised, which is a real benefit for trainers wanting to create role-plays to make them more specifically relevant for their purposes or, change the scene backdrops available – in a few moments a photo taken on a digital camera can be transferred into the tool.
- edrama is chat with purpose, it builds on a popular pastime activity amongst users young and old, but gives a framework that allows it to be purposeful activity
- Facilitation of the role-plays is a crucial aspect of the tool, this ensures that the 'chat' is purposeful and assists users to respond to the given situations
- The text from each role-play is automatically saved, which means that there is a record of every session, which can be used for assessment purposes

Although the 2D version of *edrama* has been successfully used in a number of situations and continues to provide that capacity with further opportunities to implement it in the pipeline, it has the potential to benefit from additional features. One of the main reasons that Hi8us undertook the collaborative eDrama project, was to explore these areas and see how they might benefit the user experience and delivery of the tool. For example, can edrama, delivered as a 3D environment, allow users to be more thoroughly immersed? Or, could the role-plays be automatically facilitated and be made more engaging by the use of AI agents?

This paper describes an alternative version of the *edrama* software developed in collaboration with Hi8us Midlands, Maverick TV, Birmingham University and BT with the support of the PACCIT programme (People at the Centre of Communication and Information Technologies). Our collaboration aims to enrich the user-experience with emotionally responsive characters, including additional non-human characters within a 3D application. The addition of 3D capabilities include character and background scene rendering and enables real-time processing of animation to visually update the current emotional state of every character on screen.

In the following sections we will describe the three main components integrated into *edrama*, which support the addition of expressive characters. We describe the prototype actor application supporting 3D backgrounds and avatars and integration with existing *edrama* functionality. We then cover the two components that create the expressive characters – the improvisational AI bit-part character and Demeanour. We go on to outline the user study, including the scenarios used, the experiment setup and results.

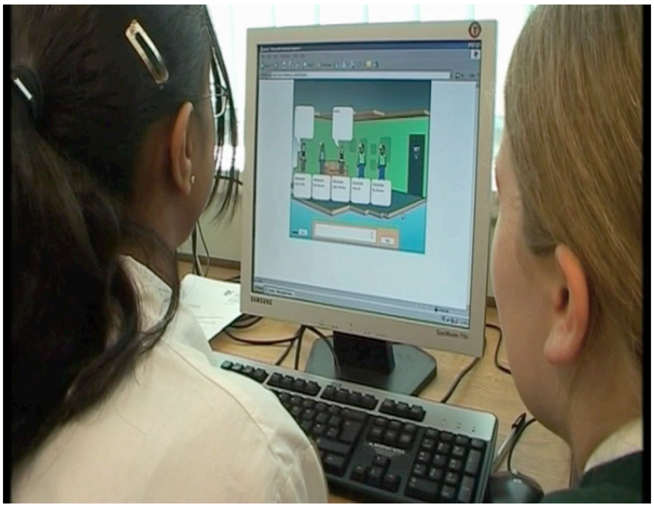


Figure 1 Children using Hi8us' 2D version of *edrama*

3 *edrama* application

The *edrama* software consists of two main user interfaces, an 'actor' client application that is used by the actors, and a 'director' client application which is available solely to the director. The director interface remains largely unchanged. It is a web-based interface which incorporates a number of tools to start role-play sessions, view the scene and avatars (in 2D), and monitor the conversation. The director can start, stop and change background scenes, and to talk to one or all of the participants using text chat. In contrast, the actor client has undergone significant developments to support the real-time rendering of expressive characters and is the focus of the user study.

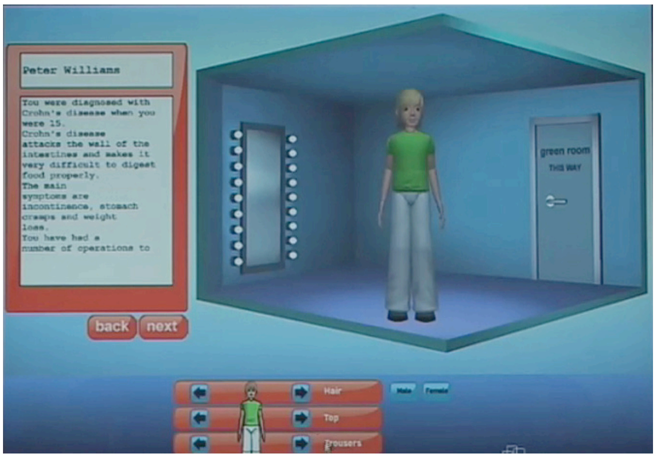
The 3D version of the *edrama* actor client is an MS Windows based application written using MFC. The application consists of two child windows; one houses the Flash* Player ActiveX control to enable Flash movies to be played within that window, the remaining is a TARA enabled window that displays the 3D visuals. Hi8us' *edrama* is a web-browser hosted Flash movie. This is noted because though the structure of the *edrama* client may have changed user interaction is still controlled through a Flash movie interface for consistency with Hi8us' versions.

TARA is an SDK developed by BT used to create real-time 3D enabled applications. The SDK provides a set of extensible components that are used to render geometry and effects using MS DirectX. The TARA SDK allows its core components to be replaced for more functional components tailored to meet a specific need. This provides a mechanism in which to integrate new technologies into TARA enabled applications, without the components of the SDK needing to know about them; in this case the Demeanour framework (Gillies et al, 2006). The creation of an alternative 3D system was an attempt to enrich the environment provided by eDrama to reinforce the emotional content of the role-play..

The flash movie developed for TARA, is the user interface that controls the flow of the application. It is a simplified version of the 2D web-based interface so that a move from one to another would require no learning on the part of the user; this also means that this prototype is compatible with the current Hi8us version therefore the two can be used in parallel. The flash movie is the client to the external server passing state related messages for each *edrama* client and capturing messages broadcast by the server. The flash interface communicates with the *edrama* application through a socket maintained by the application. Messages broadcast by the server are passed through this socket to be processed. The communication between interface and application is one way only, from interface to application as the application is responsible for reflecting the current state based upon users interactions.

To take part in an *edrama* session a user runs the actor client locally on their desktop. The user must first login, to select an available session and character to play. The login screen enables each actor to login with a unique ID. This ID can be anonymous – a numerical identifier for example or can be a username. In all cases this login does not require any personal details and is not referred to in the role-play. The login interface presents a number of options, including key stage (if required) , the role-play option and the characters available in the session. A total of 5 characters are available in each role-play.

Customisation of the character takes place in a virtual 'dressing room', available after login. This includes scenario details and customisation tools. The 3D window includes an interactive text panel which displays background information about the scenario and selected character, and also renders a default avatar in the 3D dressing room, animated with general waiting poses. Actual customisation is through the e-fit tool in the flash interface, which provides click through image based selection of gender, head, torso and leg options. The results are displayed in real time in the 3D dressing room window.



Peter Williams

You were diagnosed with Crohn's disease when you were 15.
Crohn's disease attacks the wall of the intestines and makes it very difficult to digest food properly.
The main symptoms are incontinence, stomach cramps and weight loss.
You have had a number of operations to

Figure 2 Avatar customisation

From this point the user moves into the multi-user sections. The first of these is the 'green room', which is a warm up space to meet other actors and the director. The 3D window displays the green room scenery and all logged in avatars. Text chat is displayed in speech bubbles above the avatars' heads. Text is input via the flash panel.

The director will enable the stage once all the characters have appeared in the green room and warmed up. The director signals the start of the role-play using an 'Action' command, which warns the actors of the scene change. A background scene from a library of options is displayed on all clients. This maybe updated at anytime during the session. The role-play is ended using a 'Cut' command from the director – at which point the application will close down all actor clients.

In both the green room and stage environments, each actor is given a set position on screen, resulting in a semi-circle of characters facing camera (the user's viewpoint). In this case the actor can see their own avatar in the 3rd person as part of the avatar group.

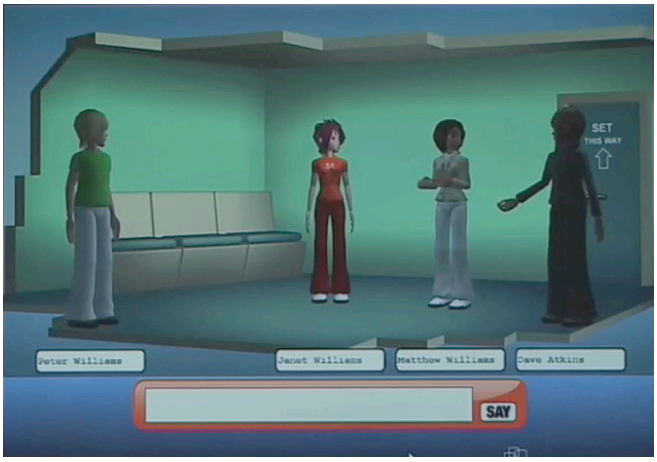


Figure 3 Four actors in the green room

This tableaux format provides the user a view of the whole role-play and avatar positions are consistent on each actor client. There is no ability to navigate the scene. This means that time is not taken up with actors trying to negotiate places on screen, or to have to arrange themselves so they can all be seen clearly. Additionally the actors can concentrate completely on talking to each other and watching the unfolding scenario on screen.

When the director speaks to the group or individual avatars a 2D director image overlays the window and text appears in a speech bubble. In this way the director can appear to the group or to single clients and give directions to assist the role-play.

When an actor types in text in the chat panel, the text appears in bubbles above the avatar. Each character is animated according to its emotional profile and to the text input of users during the session.

4 Emotionally expressive characters

Each scenario has a written description, or profile, of 5 characters who are able to participate. There is usually a main character or protagonist, who faces a conflict or issue, this character will have a counterpart who is the antagonist and takes an opposing view. The remaining characters will have specific relationships to these characters (parent, friend, enemy). This information is provided in the character background information. In many scenarios the basic character profiles have a similar pattern to provide the basis of a productive role-play. In Hi8us' versions of edrama this information can only inform the performance of the actor engaged in the role-play. In the 3D version described here it becomes influential in how the avatars are animated on screen.

Using a combination of character profiles and detected affective states from user's text it is possible to animate each character with expressive behaviour, without any direct user intervention via the *edrama* interface. This employs a combination of two technologies, affect detection in open-ended improvisational text (Zhang et al. 2006) and Demeanour framework (Gillies et al. 2006)

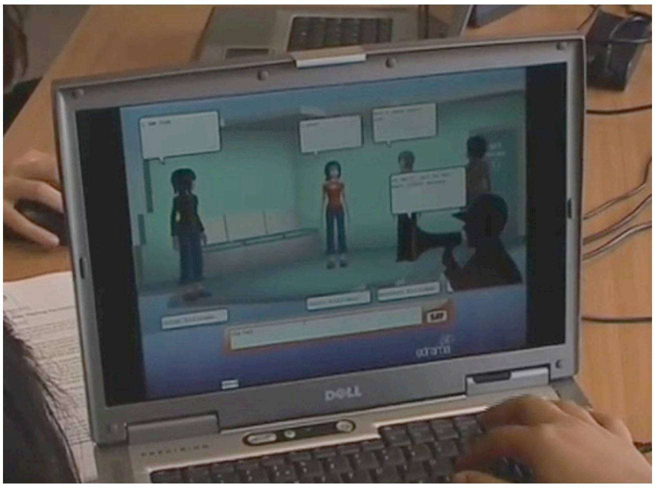


Figure 4 Director interacts with the actors

Affect detection in open-ended improvisational text

In *edrama*, the actors (users) are given a scenario within which to improvise, but are at liberty to be creative. There is also a human director, who constantly monitors the unfolding drama and can intervene by, for example, sending messages to actors, or by introducing and controlling a minor 'bit-part' character to interact with the main characters. This character will not have a major role in the drama, but might, for example, try to interact with a character who is not participating much in the drama or who is being ignored by the other characters. Alternatively, it might make comments intended to 'stir up' the emotions of those involved, or, by intervening, diffuse any inappropriate exchange developing. Additionally, the Director role was originally designed to be undertaken by Teachers, but it is now easily performed by pupils as well as teachers and this works quite well. It has also successfully been delivered by Careers Advisers, who have received no more than 30 minutes of training, to successfully perform the director role. However, within all sectors, commercial and otherwise, the need to cut costs in terms of staff time to deliver services is of great importance.

One research aim is thus partially to automate the directorial functions, which importantly involve affect detection. For instance, a director may intervene when emotions expressed or discussed by characters are not as expected. Hence we have developed an affect-detection module. The module identifies affect in characters' text input, and makes appropriate responses to help stimulate the improvisation. Within affect we include: basic and complex *emotions* such as anger and embarrassment; *meta-emotions* such as desiring to overcome anxiety; *moods* such as hostility; and *value judgments* (of goodness, etc.). Although merely detecting affect is limited compared to extracting full meaning, this is often enough for stimulating improvisation. The results of this affective analysis are then used to: (a) control an automated improvisational AI actor – EMMA (emotion, metaphor and affect) that operates a bit-part character in the improvisation; (b) drive the animations of the avatars in the user interface so that they react bodily in ways that is consistent with the affect that they are expressing, for instance by changing posture or facial expressions. The response generation component of EMMA uses this interpretation to build its behaviour driven mainly by EMMA's role in the improvisation and the affect expressed in the statement to which it is responding. The intention of EMMA's response is to hopefully stimulate the improvisation.

There has been only a limited amount of work directly comparable to our own, especially given our concentration on improvisation and open-ended language. However, *Facade* (Mateas, 2002) included shallow natural language processing for characters' open-ended utterances, but the detection of major emotions, rudeness and value judgements is not mentioned. Zhe and Boucouvalas (2002) demonstrated an emotion extraction module embedded in an Internet chatting environment. It uses a part-of-speech tagger and a syntactic chunker to detect the emotional words and to analyse emotion intensity for the first person (e.g. 'I' or 'we'). Unfortunately the emotion detection focuses only on emotional adjectives, and does not address deep issues such as figurative expression of emotion. Also, the concentration purely on first-person emotions is narrow. We might also mention work on general linguistic clues that could be used in practice for affect detection (Craggs & Wood, 2004).

Our work is distinctive in several respects. Our interest is not just in (a) the first-person, positive expression of affect case: the affective states or attitudes that a virtual character X implies that it itself has (or had or will have, etc.), but also in (b) affect that the character X implies it lacks, (c) affect that X implies that other characters have or lack, and (d) questions, commands, injunctions, etc. concerning affect. We aim also for the software to cope partially with the important case of communication of affect via metaphor (Fussell & Moss, 1998), and to push forward the theoretical study of such language, as part of our research on metaphor generally (see, e.g. Barnden et al., 2004).

Our affect detection module

The language in the textual 'speeches' created in *edrama* sessions severely challenges existing language-analysis tools if accurate semantic information is sought, even in the limited domain of restricted affect-detection. The language includes abbreviations, misspellings, slang, use of upper case and special punctuation (such as repeated exclamation marks) for affective emphasis, repetition of letters, syllables or words for emphasis, and open-ended interjective and onomatopoeic elements such as "hm", "ow" and "grrrr". To deal with the misspellings, abbreviations, letter repetitions, interjections and onomatopoeia, several types of pre-processing occur before the main aspects of detection of affect. We have reported our work on pre-processing modules to deal with these language phenomena in detail in Zhang et al. (2006).

Now we briefly introduce our work on the core aspects of affect detection. One useful pointer to affect is the use of imperative mood, especially when used without softeners such as 'please' or 'would you'. Strong emotions and/or rude attitudes are often expressed in this case. There are common imperative phrases we deal with explicitly, such as "shut up" and "mind your own business". They usually indicate strong negative emotions. But the phenomenon is more general. Detecting imperatives accurately in general is by itself an example of the non-trivial problems we face. Expression of the imperative mood in English is surprisingly various and ambiguity-prone, as illustrated below. We have used the syntactic output from the *Rasp* parser (Briscoe & Carroll, 2002) and semantic information in the form of the semantic profiles for the 1,000 most frequently used English words (Heise, 1965) to deal with certain types of imperatives. Briefly, the grammar of the 2002 version of the *Rasp* parser that we have used incorrectly recognised certain imperatives (such as "you shut up", "Dave bring me the menu" etc) as declaratives. We have made further analysis of the syntactic trees produced by *Rasp* by considering of the nature of the sentence subject, the form of the verb used, etc, in order to detect imperatives. We have also made an effort to deal with one special case of ambiguities: a subject + a verb (for which there is no difference at all between the base form and the past tense form) + "me" (e.g. 'Lisa hit/hurt me'.). The semantic information of the verb obtained by using Heise's (1965) semantic profiles, the conversation logs and other indicators implying imperatives help to find out if the input is an imperative or not.

In an initial stage of our work, affect detection was based purely on textual pattern-matching rules that looked for simple grammatical patterns or templates partially involving specific words or sets of specific alternative words. This continues to be a core aspect of our system but we have now added robust parsing and some semantic analysis, including but going beyond the handling of imperatives discussed above.

A rule-based Java framework called Jess is used to implement the pattern/template-matching rules in EMMA allowing the system to cope with more general wording. In the textual pattern-matching, particular keywords, phrases and fragmented sentences are found, but also certain partial sentence structures are extracted. This procedure possesses the robustness and flexibility to accept many ungrammatical fragmented sentences and to deal with the varied positions of sought-after phraseology in characters' utterances. The rules conjecture the character's emotions, evaluation dimension (negative or positive), politeness (rude or polite) and what response EMMA should make. The rule sets created for one scenario have a useful degree of applicability to other scenarios, though there will be a few changes in the related knowledge database according to the nature of specific scenarios.

However, it lacks other types of generality and can be fooled when the phrases are suitably embedded as subcomponents of other grammatical structures. In order to go beyond certain such limitations, sentence type information obtained from the *Rasp* parser has also been adopted in the pattern-matching rules. This information not only helps EMMA to detect affective states in the user's input (see the above discussion of imperatives), and to decide if the detected affective states should be counted (e.g. affects detected from conditional sentences won't be valued), but also helps EMMA to make appropriate responses. Additionally, the sentence type information can also help to avoid the activation of multiple rules, which could lead to multiple detected affect results for one user's input. Mostly, it will help to activate only the most suitable rule to obtain the speaker's affective state and EMMA's response to the human character.

Additionally, a reasonably good indicator that an inner state is being described is the use of 'I' (see also Craggs & Wood (2004)), especially in combination with the present or future tense (e.g. 'I'll scream', 'I hate/like you', and 'I need your help'). We especially process 'the first-person with a present-tense verb' statements using WordNet. When we fail to obtain the speaker's affective state in the current input by using Rasp and pattern matching, WordNet is used to find the synonyms of the original verb in the user's input. These synonyms are then refined by using Heise's (1965) semantic profiles in order to obtain a subset of close synonyms. The newly composed sentences with the verbs in the subset respectively replacing the original verb, have extended the matching possibilities in the pattern-matching rules to obtain user's affective state in the current input.

After the automatic detection of users' affective states, EMMA needs to make responses in her role to the human characters during the improvisation. We have also created responding regimes for the EMMA character. Most importantly, EMMA can adjust its response likelihood according to how confident EMMA is about what it has discerned in the utterance at hand.

Details of the work reported in this section can be found in Zhang et al. (2006). The brief summaries here of our previous implementations and their capabilities aim to remind readers.

The detected affective states in the user's text input and EMMA's responses to other characters have been encoded in an xml stream, which is sent to the server by EMMA. Then the server broadcasts the xml stream to all the clients so that the detected affective states information can be picked up by the animation engine to contribute to the production of 3D gestures and postures for the avatars. Now we will discuss the generation of emotional believable animation in detail in the following section.

The users avatars and emotional animation

The topics discussed in the *edrama* scenarios are often highly emotionally charged and this is reflected in the animation of the characters. Each participant in *edrama* has their own animated graphical character (avatar). In order for the characters to enhance the interaction the characters all have emotionally expressive animations. [sic] Garau *et al.* (2001) point out that avatars that do not exhibit appropriate emotional expression during emotionally charged conversation can be detrimental to an interaction. The problem with animated avatars is that they can be very complex to use if users have to directly control the avatars animation. Vilhjálmsson and Cassell (1998, 1999) have shown that users find controlling animated avatars difficult and their experience and interaction is improved if they use an avatar whose behaviour is controlled autonomously. We therefore have an autonomous model of affective animation for our avatars based on the affective states detected in users' text input. These detected affective states control the animation of the user avatars using Demeanour expressive animation framework (Gillies et al., 2006).

Demeanour makes it possible for our characters to express the affective states detected by EMMA. When EMMA detect an affective state in a user's text input, this is passed to the demeanour system attached to this user's character and a suitable emotional animation is produced. The animation system is based around a set of short animation clips, each of which is labelled with one or more affective states. Each clip only affects an individual part of the body (torso, legs, arms) and thus several clips can be easily combined at the same time. When a new affective state is received a new set of clips is chosen at random from the clips labelled with that state and these new clips are combined together to produce a new animation. Every few seconds the set of clips used is varied at random to produce a new animation, but one which has the same affective state as before. This allows us to produce varied behaviour over long time periods. The animation system also implements affective decay. Any affective state will eventually revert to a neutral state if it is not replaced by a new one.

Another feature of the animation system is that characters can produce affective responses to the states of other characters. If a character produces a strong affective state then other characters will also produce a milder response. Each character has a profile which specifies how it responds to the behaviour of each other character. This makes it possible to implement different responses for different characters. For example, two characters with a positive relationship may empathise with each other, when one is unhappy so is the other. On the other hand if two characters have a negative relationship then one might gloat at the other's unhappiness, and therefore display happiness.

Demeanour generates a number of output affective states, which are used to select the animation clips. Each output state is a weighted sum of a number of input factors. The primary input factor is the affective state as detected from the input text, this always has weight 1. The inputs also include the states of other characters, with lower weights. So for example, the output state "happiness" depends on the input "happiness", but also on the "happiness" and "sadness" values of other characters. The weights of the other characters states are contained in the character's profile. The profile consists of a separate set of weights for each other character in the scenario. This makes it possible to respond differently to each character. For example, if two characters, A and B have a poor relationship and A is angry with B, B might respond by being angry back. On the other hand if B's parents were angry then B might be sad or submissive. At any time each character has a single focus of social attention (which is itself another character), determined by the character's direction of gaze (which is itself determined by an animated gaze model). In order to generate animation the first step is to update the input states based on any text typed in. Next the states of the current focus of attention are fetched. These are multiplied by the weights of given by the profile specific to the focus of attention and added to the input emotion to produce the output state.

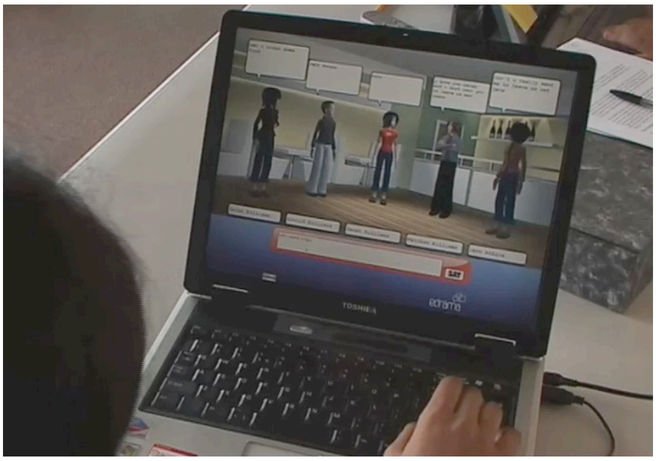


Figure 5 Children using the 3D version of *edrama*

Table 2. The Categories of questions used for the questionnaires

Category	Number of questions	Example Questions
Enjoyment	12	“ How much did you enjoy the roleplays?”
Difficulty	14	“I needed help to use edrama”
Presence	2	“I forgot I was at school when I was doing the role-play”
Co-Presence	8	“ Did you feel close to the group online?”
Quality of Social Interaction	6	“ Did you get to have your say?”
Own Avatar Appearance	3	“ I wanted the Avatar to look more like me”
Own Avatar Behaviour	5	“ My Avatar was expressive”
Other Avatars	10	“I was paying attention to other people’s Avatars”

5 The User Study

Our user study involved two trials of the prototype 3D *edrama* application. These were completed in July and October 2006, as new animation capabilities were added to the prototype system. This section describes the scenarios, the user study and results.

The scenarios

Three scenarios were used in the user testing; the first was entitled ‘Big Night Out’ which was delivered in the 2D version and served as a warm up; the other two were homophobic bullying and Crohn’s disease both of which are described below. In each case, introductory video produced by Maverick TV were shown to the trialists. These videos were case studies of both subject areas and featured interviews with either victims of bullying or sufferers of Crohn’s disease. This is additional information to help participants identify with the sensitive issues being explored in the scenarios. In these scenarios, Mr Dhanda (Homophobic Bullying) and Dave (Crohn’s Disease) are AI characters driven by EMMA.

Homophobic Bullying

In this scenario the character Dean (16 years old), captain of the footbal team, is confused about his sexuality. He has ended a relationship with a girlfriend because he thinks he may be gay and has told her this in confidence. Tiffany (ex-girlfriend) has told the whole school and now Dean is being bullied and concerned that his team mates on the football team will react badly. He thinks he may have to leave the team. The other characters are; Tiffany who is the ring leader of the bullying, and wants Dean to leave the footbal team, Rob (dean’s younger brother) and wants Dean to say he is not gay to stop the bullying, Lea (Dean’s older sister) who wants Dean to be proud of who he is and ignore the bullying, and Mr Dhanda (PE Teacher) who needs to confront Tiffany and stop the bullying.

Crohn’s Disease:

In this scenario the character Peter has had Crohn’s deisease since the age of 15. Crohn’s disease attacks the wall of the intestines and makes it very difficult to digest food properly. The character has the option to undergo surgery (ileostomy) which will have a major impact on his life. The task of the role-play is to discuss the pros and cons with friends and family and decided whether he should have the operation. The other characters are; Mum, who wants Peter to have the operation, Matthew (older brother) who is against the operation, Dad who is not able to face the situation, and David (the best friend) who mediates the discussion. The setting is a night out for an evening meal.

Procedures

There were 3 conditions in the user study:

1. Hi8us’ 2D version of *edrama* with no animation or affect detection
2. The 3D version of *edrama* with the bit part character but limited animation
3. The 3D version with the bit part character and full animation

In the version with limited animation the animations only occurred when an emotion was detected by the emotion detection system and there was only one animation per emotion. In the full animation condition animations were constantly being played and there were a variety of possible animations for each emotion.

The comparison between the 2D version and the 3D versions was performed within subjects while the comparison between the two 3D conditions was performed between subjects. The participants were therefore divided into two groups as show in the table.

Table 1. Experimental Conditions

Group	Condition 1	Condition 2
A	2D	3D with limited animation
B	2D	3D with full animation

The two groups were tested in different sessions. Group A used the 2D and 3D versions on different days while group B used them on the same day.

There were 10 participants per group. The participants were all female aged between 13 and 14 and pupils at Swanshurst School, a Specialist Science College in Billesley, Birmingham.

The participants were randomly assigned into groups of 4 and given a scenario and character. None of the participants knew who the other members of their group were. However, due to the proximity of the terminals, sometimes they were able to establish identities of fellow participants.

The participants were then asked to role play using the *edrama* system for 10-15 minutes per session, they undertook 3 sessions in the first trial but only 2 sessions in the second one. They had less time to undertake the 2D session in the second trial, due to the technical difficulties.

Results

Participants were asked to complete a questionnaire about their experience with the 2D *edrama* before using the 3D version and a second one after using the 3D version. The two questionnaires were mostly identical, but some minor changes were made to the questions to make them applicable, and 7 questions were added that were not applicable to the 2D version. The First questionnaire had 71 questions and the second had 78. All of the questions were 7 point Likert like scales.

The questions were divided into 7 categories, shown in table 2. For each participant the mean was taken of their answers to for the questions in each category and this was used as their score for that category. The mean was then taken for each condition for each category.

The first comparison was between the 2D condition and the two 3D conditions:

Table 2 Comparison of 2D and 3D conditions

Category	2D Mean	3D Mean	t-value
Group A			
enjoyment	5.075	4.875	-0.418
difficulty	2.042	2.07	0.087
presence	4.9	3.6	-3.00
co-presence	3.5	3.5	0
Social Dynamics	4.05	4.95	2.347
Avatar Appearance	3.266	3.7	0.979
Avatar Behaviour	3.3	2.9	-0.669
Other Avatars	4.483	4.03	-1.33
Group B			
enjoyment	4.5	5.516	1.53
difficulty	3.278	2.528	-1.494
presence	4.8	5.35	0.992
co-presence	3.614	4.137	1.016
Social Dynamics	3.633	5	2.306
Avatar Appearance	3.76	3.033	-1.568
Avatar Behaviour	3.85	4.96	1.863
Other Avatars	4.5	5.35	2.072

The main significant result that was obtained consistently between the groups was that the quality of social interaction improved with the 3D condition. Interestingly the participants reported evaluation of the avatars was not significantly different between conditions, except for the case of the other avatars in the full animation condition. This might be because the participants did the first questionnaire before doing the 3D version and so were not directly comparing the avatars in the two systems.

The second comparison was between the 3D conditions with limited and full animation:

Table 3 Comparison of limited animation and full animation conditions

Category	Group A	Group B	t-value
enjoyment	4.875	5.516	1.324
difficulty	2.071	2.528	1.186
presence	3.6	5.35	3.719
co-presence	3.5	4.137	1.481
Social Dynamics	4.95	5	0.092
Avatar Appearance	3.7	3.033	-1.862
Avatar Behaviour	2.9	4.96	4.616
Other Avatars	0.631	5.35	3.660

These results show a significant improvement of the evaluation of the behaviour of the participants own avatar and of the other avatars, demonstrating that realistic animation and emotionally expressive behaviour have a strong effect on people's evaluations of avatars. There was also a significant improvement in presence and a notable but non-significant improvement in co-presence showing that this has a real effect of the participants experience. Interestingly there was a reasonably strong result that the avatars appearance was considered worse in the full animation condition. This may be because participants concentrated less on the appearance when the characters' behaviour was more lively.

Following the testing the participants were invited to make comments about their experience in an open interview. Regarding the 3D developments feedback included, "I think it's good that it's 3D because you can sort of.. it gives you more of a vision on how everyone's acting.. sort of thing and how people can react.. because you can sort of see all the shadows and stuff it's more realistic and it gets you.. into it a bit more"

Regarding the animation one pupil observed, "They did move differently, like if I said something, erm, like lovingly towards someone they did an action kind of expressing that as well"

From the teachers perspective on the critical nature of the participants, "I think the feedback's been very positive, the children are very au fait with this kind of software, they are used to using these kinds of systems at home and they can give good quality feedback on it, they know what they want, they know what they want to have and how to use it and they know what's possible."

6 Conclusion/Discussion

edrama provides a platform for participants to engage in focused discussion around emotionally charged issues. This new prototype provides an opportunity for the developers to explore how emotional issues embedded in the scenarios, characters and dialogue can be represented visually without detracting from the learning situation.

The user trials demonstrate that the creation of a 3D animated version of *edrama* indicates a marked improvement on the role-playing experience using the *edrama* system. The 3D version of the system, with the automated bit part character, may contribute to improving the perceived quality of social interaction over and above the original 2D version. In addition to this, adding emotionally appropriate animations to the user avatars improves both the participants' evaluation of those characters and their sense of presence. There is great potential for the use of *edrama* in education in areas such as citizenship, PHSE and drama. Beyond the classroom *edrama* can be easily customised for use in professional training, where face to face training can be difficult or expensive, such as customer services training and e-learning in the workplace.

Our research shows that the application of expressive characters to online role-play contributes positively to an already engaging user experience. Future work could include the exploration of automated bit-part characters to fully develop a non-human director. Additionally tools to enable participants to replay the role-plays have been considered. These could enable further reflection and group discussion, allowing for comparisons of sessions between different groups of learners. Replays could even be altered to adjust the emotional states of each character and generate different online 'performances', which could create emotionally rich experiences for audiences as well as participants.

7 Acknowledgements

This work is a result of the LINK project eDrama - Enhancement of People, Technology and their Interaction: funded by PACCIT (People @ The Centre of Communication and Information Technologies).

The project partners would like to thank Swanshurst School, Billesley, Birmingham for their participation in this study.

8 References

Barnden, J.A., Glasbey, S.R., Lee, M.G. & Wallington, A.M. 2004. Varieties and Directions of Inter-domain Influence in Metaphor. *Metaphor and Symbol*, 19(1), pp.1-30.

Briscoe, E. & Carroll, J. 2002. Robust Accurate Statistical Annotation of General Text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Gran Canaria. 1499-1504.

Cassell, J. & Vilhjálmsson, H. 1999. "Fully Embodied Conversational Avatars: Making Communicative Behaviors Autonomous. *Autonomous Agents and Multi-Agent Systems* 2(1): 45-64.

Craggs, R. & Wood. M. 2004. A Two Dimensional Annotation Scheme for Emotion in Dialogue. In *Proceedings of AAAI Spring Symposium: Exploring Attitude and Affect in Text*.

Fussell, S. & Moss, M. 1998. Figurative Language in Emotional Communication. In S. R. Fussell and R. J. Kreuz (Eds.), *Social and Cognitive Approaches to Interpersonal Communication*. Lawrence Erlbaum. pp.113-142.

Garau, M., Slater,M., Bee, S. and Sasse, M.A. 2001. The impact of eye gaze on communication using humanoid avatars. *Proceedings of the SIG-CHI conference on Human factors in computing systems*, March 31 - April 5, 2001, Seattle, WA USA, 309-316.

Gillies, M. & Ballin, D. 2004. Integrating autonomous behavior and user control for believable agents, in the *Proceedings of the Third international joint conference on Autonomous Agents and Multi-Agent Systems* Columbia University, New York, July 2004

Heise, D. R. 1965. Semantic Differential Profiles for 1,000 Most Frequent English Words. *Psychological Monographs*. 70 8:(Whole 601).

Mateas, M. 2002. Ph.D. Thesis. Interactive Drama, Art and Artificial Intelligence. School of Computer Science, Carnegie Mellon University.

Vilhjálmsson, H. & Cassell, J. 1998. BodyChat: Autonomous Communicative Behaviors in Avatars. Proceedings of ACM Second International Conference on Autonomous Agents, May 9-13, Minneapolis, Minnesota.

Zhang, L., Barnden, J.A., Hendley, R.J. & Wallington, A.M. 2006. Exploitation in Affect Detection in Open-ended Improvisational Text. In *Proceedings of Workshop on Sentiment and* Subjectivity at COLING-ACL 2006, Sydney, July 2006.

Zhe, X. & Boucouvalas, A. C. 2002. Text-to-Emotion Engine for Real Time Internet Communication. In *Proceedings of International Symposium on Communication Systems*, Networks and DSPs, Staffordshire University, UK, pp.164-168.

Agent Personality Traits in Virtual Environments Based on Appraisal Theory Predictions

Lori Malatesta, George Caridakis, Amaryllis Raouzaiou and **Kostas Karpouzis** [1]

Abstract. The current work investigates issues of expressivity and personality traits for Embodied Conversational Agents in environments that allow for dynamic interactions with human users. Such environments are defined and modelled with the use of state of the art game engine technology. We focus on generating simple ECA behaviours, comprised of facial expressions and gestures in a well defined context of non-verbal interaction.

1 INTRODUCTION

Research in affective computing and virtual agents has still many challenges to take up. Enthusiastic reactions to the virtues of affective human-machine interactions have often been disproved by more in depth studies. Although it is clear such interactions are richer and allow for stronger human relations, what is not clear is in what situations one should use such types of interactions with computers [18].

In the case of the design of virtual agents, issues of believability and naturalness have to be addressed, along side with user expectations and the reported quality of interaction. To increase believability and life-likeness of an agent, she has to express emotion [4] and exhibit personality in a consistent manner [13], [16]. Several studies have shown the significance of cultural factors, personality and environment setting when designigning an agent [11]. These studies have also pointed out the importance of consistency in a virtual character. Traits are regarded as chronic propensities to get into corresponding emotional states and thus are a major source of emotional and behavioural consistency.

In our current work we focus on modelling affective virtual characters so as to depict different behaviours to similar situations depending on their personality traits and current moods. Our work is based on a collapsed version of the OCC model proposed by Ortony in [16]. The structure of the paper is as follows: Section 2 consists of a short literature review explaining the motivation of our current work, section 3 gives an account of the parameters used to manipulate expressivity, section 4 presents our modelling approach, the chosen context in terms of interaction scenarios, the application structure and the technologies adopted. We conclude with section 5 where the necessary next steps are identified in terms of extending and evaluating the model.

2 PERSONALITY, MOODS AND EMOTIONS

In psychology research the term affect is very broad, and has been used to cover a wide variety of experiences such as emotions, moods, and preferences. In contrast, the term emotion tends to be used to refer to fairly brief but intense experiences although it is also used in a broader sense. Finally mood or state describe low-intensity but more prolonged experiences [8].

It is common in personality and emotion literature to focus on general positive or negative moods and on the broad traits of positive/negative affectivity and extraversion/neuroticism. According to [7] extraversion concerns individual differences in the preference for social interaction and lively activity whereas neuroticism represents individual differences in proneness to unpleasant emotional experience. Traits of affectivity are often defined as stable individual differences in the tendencies to experience positive and negative mood states.

Nervertheless according to a detailed review of research on emotion and cognition by Rusting [19] very few studies in psychology have included measures of traits directly related to mood regulation (e.g. negative mood-regulation expectancies, meta-mood experience). There remain many gaps in the understanding of $Personality$ x $Mood$ interactions. Personality traits represent underlying propensities toward mood states, but do not necessarily always produce them (e.g. an individual high in neuroticism can be in a good mood at least for some of the time).

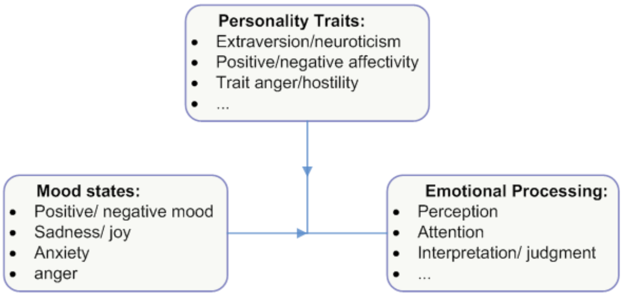


Figure 1. The mediation approach, in which mood-congruency effects depend on individual differences in emotional personality traits

In Rusting's review, it is acknowledged that among various emotional processing theories there is reasonable support for the moderator approach which claims that emotional processing depends interactively on personality and mood state (see figure 1). This gives us motivation to try and model an expressive character based on the prevailing personality traits and take under consideration a broad account for positive and negative mood states. It is stated that the broad dimensions of positive affectivity and extraversion, and negative affectivity and neuroticism, may represent similar underlying tendencies with respect to positive and negative mood experience, and they may therefore involve similar sensitivities to positive and negative

[1] Image, Video and Multimedia Systems Lab, National Technical University of Athens, Greece, email: lori@image.ntua.gr

emotional cues.

2.1 Emotion models based on appraisal

There exist several theories in psychology for modelling and representing the process of emotion elicitation [10], [17],[21], [20]. These models provide predictions for possible emotional states through the process of cognitive appraisal of stimuli. Various virtual human models have been put forward using these theories as foundations in conjunction with factor models of personality such as the five factor model [15]. To name a few, there is work on presentation strategies by affective virtual humans [2], the virtual human project based on a layered model of affect accounting for both mood and personality traits in a dialogue based agent interaction [9] and the multilayer personality model [12], [6]. The common denominator in this line of research is mutlimodality. Both speech and facial expressions are modelled as well as body gestures.

Our current research aims to initially scale down the problem and thus we currently choose to focus only on non-verbal human-to-agent interactions. We want to give virtual characters expressivity that makes sense in the context it is expressed. The context is provided through chosen scenarios put forward in a following section. Motivation for this approach was given by Ortony's recent simplification of the original OCC model [16].

2.2 A simplified version of the OCC Model of Appraisal

In this approach it is stated that believability is an application-dependent notion, strongly related to context. The simplification collapses the original 22 emotion types down to five distinct positive and five distinct negative reactions by taking under consideration the emotional states that make sense for a virtual character. The idea is to start simple in making the agent able to differentiate his expressions between positive and negative and then progressively develop more elaborate categories. An agent could have an identical positive expression in a situation where she is happy about obtaining a desired object or in a situation where she is happy because she feels proud when she has attained some goal. The expressivity would not change in such a coarse approach, only the context.

The main point of the OCC model is that the appraisal process taking place during an emotion elicitation event is either in terms of events, in terms of an agent's actions or in terms of objects (and attitudes towards them). As a first step to tackle the modelling problem we are only going to focus on events. This simplifies the agent's candidate emotion states to only ones related to events according to OCC such as joy, distress, hope, fear, relief etc.

According to this simplified version of the OCC model each emotion type is assocciated with a variety of possible reactions. It is considered that all emotions share the same set of $response\ tendencies$ and the differentiation lies in the extent each tendency participates in the state. Ortony defines three major types of emotion response tendencies: expressive, information - processing and coping (see figure 2).

In our current work there is yet to be an account of world knowledge representation for the virtual character and no information processing or coping functionality. Thus we are going to focus solely on the expressive reaction tendencies of each emotion state. The expressive reaction states are divided in three subcategories as depicted in figure 2.

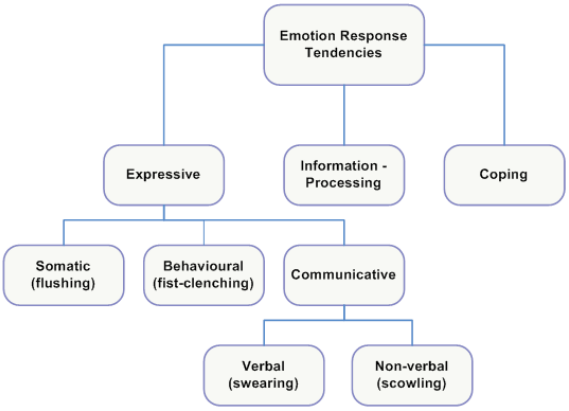


Figure 2. Emotion response-tendencies in the case of anger

A key issue is the mapping of emotional states of the character to behaviours and actions. Here is where personality and repsonse tendencies take over. Personality is the engine of behaviour. One tends to react in a certain way in a situation because she is that kind of person [16]. Thus personality is the key to character and behaviour consistency. The good news is that traits don't live in isolation. On the contrary they are strongly correlated and tend to cluster together. Upon this truth lie the various factor structures of personality.

While trying to keep the level of complexity of our model as low as possible we are going to account only for two personality traits: extraversion and neuroticism. These are only two traits of the five factor model that is comprised of openess, conscientiousness, extraversion, agreeableness and neuroticism [15]. Neuroticism is reported as the tendency to experience negative thoughts and extraversion as willingness to communicate and preference for social situations.

3 EXPRESSIVITY PARAMETERS

3.1 Facial expressions

Our group has previously focused on the animation of facial expressions based on the predictions of K. Scherer's component process model (CPM) theory [21], [14]. Component process model theory studies the emotion elicitation process and provides analytical facial deformation predictions based on the cognitive appraisal of the stimuli presented to the subject. These predictions were mapped to MPEG-4 facial animation parameters and videos of the evolution of the emotion expression were synthesised. This was a stand alone approach and the work produced is currently in the phase of evaluation through a rating tests and further expression synthesis. It is not yet obvious if such predictions can lead to realistic synthesis results and there remain a lot of issues to investigate. Keeping also in mind the fact that neither MPEG-4 facial animation parameters have a mapping in human models in virtual worlds as yet, our current approach caters for rudimentary positive and negative facial expressions with the intent to extend it for the MPEG-4 standard. The benefits of such an extension are several. The standard allows for flexible manipulation of objects and models in synthesised environments and facilitates both re-use and deep parametrisation of the produced animations.

3.2 Body expressivity

Our previous work on gesture expressivity from the frame by frame analysis of naturalistic video sequences [5] has six dimensions of expressivity to offer which we can manipulate in our current work. These dimensions have been designed for communicative behaviours only. Each dimension acts differently for each modality. For an arm gesture, expressivity works at the level of the phases of the gesture: for example the preparation phase, the stroke, the hold as well as on the way two gestures are co-articulated. The six dimensions of expressivity proposed:

- Overall activation
- Spatial extent
- Temporal
- Fluidity
- Power/Energy
- Repetitivity

These parameters were extracted both manually from the annotation of real data video corpus and automatically from the video corpus of acted data using image analysis techniques. They were then used through a copy-synthesis approach in synthesising similar behaviour in virtual humans (see figure 3).

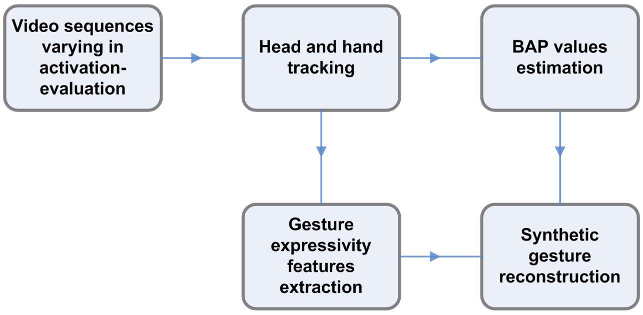

Video sequences varying in activation-evaluation
Head and hand tracking
BAP values estimation
Gesture expressivity features extraction
Synthetic gesture reconstruction

Figure 3. Synthetic Gesture Reconstruction

Overall activation was considered as the quantity of movement during a conversational turn in the video. Spatial extent was modeled by expanding or condensing the entire space taken up by a gesture. The temporal parameter determines the speed of movement for the participating body parts in a gesture and also signifies the duration of movements (e.g. quick versus sustained actions). Fluidity differentiates smooth/graceful from sudden/jerky actions and captures the continuity between movements.

In the synthetic gesture recunstruction phase both extracted expressivity parameters and MPEG-4 body animation parameters (BAPs) were used. There is currently no available mapping of the MPEG-4 standard parameters for a character in a virtual environment at the moment. Therefore in our synthesis approach we only rely in expressivity parameters at this phase.

In order to differentiate the two modelled personalities (extrovert with positive affectivity and neurotic with negative affectivity), we are going use the above expressivity parameters and adjust their intenisity in order to generate behaviours that make sense for the given personality traits and their given moods.

4 APPLICATION

4.1 Overview of modelling approach

At this point we are interested in modelling a virtual character's expressivity through simple interactive scenarios with human users. The idea is to provide a well defined context for non-verbal interaction between a human and an agent. The human user will be given a choice of actions and the agent will react affectively depending on the appraisal of the action by the user. It is not in the scope of the current paper to investigate issues of knowledge representation for the virtual character. Thus we adopt a simplified rule based solution for the agent's action/ event appraisal based on the personality traits attributed to the character. This means that the virtual characters' goals and intentions are defined in the domain of the application, in our case in the interaction scenarios. We use Finite State Machines in order to model each personality.

At this point one state machine is used for the extrovert/positive affectivity case and one for the neurotic/ negative affectivity. Having assumed that the extrovert is more prone to be in a good mood and to react to positive stimuli were as the neurotic personality is more prone to be in a bad mood and to pay more attention to negative stimuli we came up with the scenarios described in the following subsection.

In terms of finite state machine terminology in one *reaction* the machine maps a current state and an given input to a subsequent state and a specific output (see figure 4. In our case appraisal of stimuli serves as input and positive or negative expressions serve as outputs while state transision is attained between positive mood state and negative mood state as depicted in the state transision diagram. Appraisal of stimuli is modelled taking under consideration the constraints put forward for each personality. Probability functions are used to express a high likelihood for the extrovert personality to stay in a positive mood and react to positive emotions and similarly a low likelihood for the neurotic to move to a positive mood state etc.

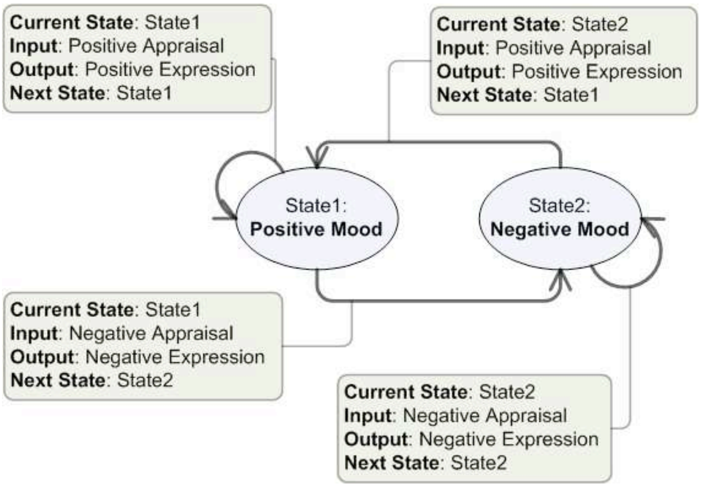

Current State: State1
Input: Positive Appraisal
Output: Positive Expression
Next State: State1
Current State: State2
Input: Positive Appraisal
Output: Positive Expression
Next State: State1
State1: Positive Mood
State2: Negative Mood
Current State: State1
Input: Negative Appraisal
Output: Negative Expression
Next State: State2
Current State: State2
Input: Negative Appraisal
Output: Negative Expression
Next State: State2

Figure 4. State transition diagram for a character's moods

4.2 Application scenarios

In a neutral virtual environment simulating a valley, the human user, originally situated in visual field of the agent, has a choice of actions to execute. Only one of the two virtual characters is present at a time. The stimuli produced by the user's action is appraised. Each agent will react differently to the user stimuli and consistently to her attributed personality and current mood. We have chosen the same 3D model for both personalities in order to counterbalance effects of appearance (a user could find one model more friendly judging only by her appealing appearance which would lead to confounding results on the perceived expressivity of the agent). The user can manually choose the affective state she depicts as she carries out one of the available actions. The choice is between neutral, happy or angry. In the first scenario the user can approach or move away from the agent $expressing$ one of the three affective states. In the second scenario the user apart from approaching, also has the choice to execute an action such as lighting a fire, planting a flower, throwing a rock etc. The scenarios are purposely chosen to be simple at this stage. The agent's appraisal of the events and consequent reactions/expressions are the fingerprint of her character/ mood state. The neurotic agent is more prone to interpret the approach of a neutral face approaching as negative stimuli and stay in a bad mood. The extrovert will react pleasantly and expressively in a similar case due to her tendency to stay in a good mood and to interpret stimuli in a positive manner.

4.3 Why use a game engine?

As previously mentioned, we are interested in modelling a virtual character's expressivity through simple interactive scenarios with human users. Game engines enable simplified, rapid development of the required interaction premises allowing one to stray from a formal game development approach. Through such an environment a user is given the chance of interacting with a virtual character in a realistic setting where she can move freely and invoke various actions.

4.3.1 Torque

Torque is a game engine by Garage Games [1]. It has been chosen as a platform among various others, mainly because of its flexibility/ease in quick virtual environment deployment, its vibrant developer community and its open source policy. It allows for the implementation of Finite State Machines in order to model the states of a virtual character. In our case we are interested in modelling both mood and emotional states. and introducing their interaction with the personality and current

Figure 5. Screen shot from the Torque game engine

5 CONCLUSIONS AND FUTURE WORK

As pointed out by Bartneck in his review of the OCC model for embodied characters [3], mapping emotion categories to available expressions should be based on strong theoretical foundations, that might not always be available. When not available the developer of the character is then forced to invent these mappings. Any such arbitrary mappings in our work will be empirically tested in user studies to follow. These studies will aim to measure the effects of expressivity of the virtual character, how they are perceived and how they are rated in terms of believability, naturalness and overall appeal.

Results from this formative evaluation will be used as feedback for further extension of the model. We are interested in expanding the covered emotion categories of the collapsed OCC model as well as accounting for a history function of the visited states allowing for appraisals that evolve over time and $remember$ previous states thus developing an attitude towards events.

Another interesting point worth investigating is the analysis of the human user's affective states during such interactions. It is common practice to use a game environment in order to collect such data and the specific context of interaction can provide vital information regarding the appraisal processes taking place, the temporal evolution of emotional episodes and the possible relationships of human and agent affective states.

REFERENCES

[1]

[2] E. Andr, M. Klesen, P. Gebhard, S. Allen, and T. Rist, 'Integrating models of personality and emotions into lifelike characters', 150–165, (2000).

[3] C. Bartneck. Integrating the occ model of emotions in embodied characters, 2002.

[4] J. Bates, 'The role of emotion in believable agents', *Communications of the ACM*, **37**(7), 122–125, (1997).

[5] G. Caridakis, A. Raouzaiou, K. Karpouzis, and S. Kollias, 'Synthesizing gesture expressivity based on real sequences'. Workshop on multimodal corpora: from multimodal behaviour theories to usable models, LREC 2006 Conference, Genoa, Italy, 24-26 May., (2006).

[6] A. Egges, S. Kshirsagar, and N. Magnenat-Thalmann, 'A model for personality and emotion simulation.', in *KES*, pp. 453–461, (2003).

[7] H.J. Eysenck and M. Eysenck, *Personality and Individual Differences: A Natural Science Approach*, New York, NY: Plenum Press, 1985.

[8] M. W. Eysenck and M. T. Keane, *Cognitive Psychology: A Student's Handbook*, Psychology Press (UK), August 2000.

[9] P. Gebhard, 'Alma: a layered model of affect', in *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pp. 29–36, New York, NY, USA, (2005). ACM Press.

[10] P.P. Jose I.J. Roseman, A.A. Antoniou, 'Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory', *Cognition & Emotion*, 241–277, (1996).

[11] K. Isbister and C. Nass, 'Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics', *Int. J. Hum.-Comput. Stud.*, **53**(2), 251–267, (2000).

[12] S. Kshirsagar, 'A multilayer personality model', in *SMARTGRAPH '02: Proceedings of the 2nd international symposium on Smart graphics*, pp. 107–115, New York, NY, USA, (2002). ACM Press.

[13] A. B. Loyall and J. Bates, 'Personality-rich believable agents that use language', in *AGENTS '97: Proceedings of the first international conference on Autonomous agents*, pp. 106–113, New York, NY, USA, (1997). ACM Press.

[14] L. Malatesta, A. Raouzaiou, K. Karpouzis, and S. Kollias, 'Mpeg-4 facial expression synthesis based on appraisal theory'. 3rd IFIP conference in Artificial Intelligence Applications and Innovations, (2006).

[15] R. R. McCrae and O. P. John, 'An introduction to the five-factor model and its applications', *Journal of Personality*, 175–215, (1992).

[16] A. Ortony, *On making believable emotional agents believable*, 189–211, Emotions In Humans And Artifacts, Cambridge, MA: MIT Press, 2001.

[17] A. Ortony, A. Collins, and G.L. Clore, *The Cognitive Structure of Emotions*, Cambridge University Press, 1988.

[18] *Affective Interactions, Towards a New Generation of Computer Interfaces.*, ed., A. Paiva, volume 1814 of *Lecture Notes in Computer Science*, Springer, 2000.

[19] C.L. Rusting, 'Personality, mood, and cognitive processing of emotional information: three conceptual frameworks', *Psychological Bulletin*, **124**, 165–196, (1998).

[20] D. Sander, D. Grandjean, and K. R. Scherer, 'A systems approach to appraisal mechanisms in emotion', *Neural Netw.*, **18**(4), 317–352, (2005).

[21] K. R. Scherer, 'The role of culture in emotion-antecedent appraisal', *Journal of Personality and Social Psychology*, **73**, 902–922, (1997).

An expressive ECA showing complex emotions

Elisabetta Bevacqua[1]
Maurizio Mancini[1]
Radoslaw Niewiadomski[2]
Catherine Pelachaud[1]

Abstract. Embodied Conversational Agents (ECAs) are a new paradigm of computer interface with a human-like aspect that allow users to interact with the machine through natural speech, gestures, facial expressions, and gaze. In this paper we present an head animation system for our ECA Greta and we focus on two of its aspects: the expressivity of movement and the computation of complex facial expressions. The system synchronises the nonverbal behaviours of the agent with the verbal stream of her speech; moreover it allows us to qualitatively modify the animation of the agent, that is to add expressivity to the agent's movements. Our model of facial expressions embeds not only the expressions of the set of basic emotions (e.g., anger, sadness, fear) but also different types of complex expressions like fake, inhibited, and masked expressions.

1 Introduction

Embodied Conversational Agents (ECAs) are a new paradigm of computer interface with a human-like aspect that are being used in an increasing number of applications for their ability to convey complex information through verbal and nonverbal behaviours like voice, intonation, gaze, gesture, facial expressions, etc. Their capabilities are useful in scenarios such as a presenter on the web, a pedagogical agent in tutoring systems, a companion in interactive settings in public places such as museums, or even a character in virtual storytelling systems. Our system provides control over the animation of a virtual agent head. It computes realistic behavior for the head movement (nods, shakes, direction changes, etc), gaze (looking at the interlocutor, looking away) and facial expression (performing actions like raising eyebrows, showing an emotion, closing eyelids, and so on). During conversation the agent moves her head according to what she is saying. Moreover eye movements are computed depending on the gaze intention. Since eyes and head are physically linked these two communicative modalities cannot be computed separately, so our system exhibits head and eye coordination to obtain a realistic gaze behaviour.

In this paper we present an ECA animation system, called Greta, focusing on two of its aspects: the expressivity of movement and the computation of complex facial expressions.

The *expressivity* of behaviour is "How" the information is communicated through the execution of some physical behaviour. Expressivity is an integral part of the communication process as it can provide information on the state of an agent, his current emotional state, mood, and personality [47]. Section 3 gives an overview of our head animation system architecture while Section 4 explains the implementation of the expressive animation computation.

There is a large amount of evidence in psychological research that human's repertoire of facial expressions is very large [13, 14, 34]. We call *complex facial expressions* the expressions that are different from the spontaneous facial displays of simple emotional states (e.g. display of anger or sadness). They can be displays of some combinations of emotions as well as expressions of emotions which are modified according to some social rules. It was shown [17, 24] that an expressed emotion does not always reveal a felt emotion. People may, for example, decide not to express the emotion they feel because of some socio-cultural norms called *display rules* [14]. When display rules are applied, a set of procedures of emotional displays management [42] is used. These procedures leads to different facial expressions [15].

It was proved that these facial expressions can be distinguished by humans (i.e. there are different facial signals) [16, 19] and have different role and meaning [14,34]. This is why we have introduced the Complex Facial Expression Computation module which is detailed in Section 5. In section 2 we discuss some of the previous works on ECAs focusing on gaze, head and facial expression models. Then we give a detailed explanation of our head animation system in sections 3, 4 and 5. Finally we conclude the paper in section 6.

2 State of the art

Overviews of recent ECA implementations have been described by Cassell et al. and Prendinger et al. [8, 38]. K. R. Thórisson developed a multi-layer multimodal architecture able to generate the animation of the virtual 2D agent 'Gandalf' during a conversation with a user [43]. Gandalf has been created to communicate with users also through head movements (nods) and gaze direction. 'Rea' [7] is a humanoid agent able to understand the user's behaviour and respond with appropriate speech, facial expressions, gaze, gestures and head movements.

A number of studies have underlined the importance of gaze and head behaviour in the communication process. Vertegaal et al. [45] found that gaze is an excellent predictor of conversational attention in multiparty situations and placed special consideration on eye contact in the design of video conference systems [46]. Peters et al. [31] proposed a model of attention and interest using gaze behaviour, defining the capabilities an ECA requires to be capable of starting, maintaining and ending a conversation. Head movements hold an impor-

[1] University of Paris8, IUT de Montreuil, 140 rue de la Nuovelle France, 93100, Montreuil, France,
email: e.bevacqua,m.mancini,c.pelachaud@iut.univ-paris8.fr

[2] Università degli Studi di Perugia, Dipartimento di Matematica e Informatica, Via Vanvitelli 1, 06123, Perugia, Italy,
email: radek@dipmat.unipg.it

tant role in conversation and researches have been done to determine their pattern in order to enrich ECAs with more believable head animation. Heylen analyzed head patterns to define their properties and functions [21] useful to implement ECAs behaviour.

In all of these systems the final animation is obtained by interpolating between pre-determined body and facial configurations. One of the novelty of our system is that the agent movements can be qualitatively modified (changing their amplitude, speed, fluidity, etc) applying some parameters to add expressivity to the ECA.

Most of animated agents are able to display a small number of emotions (e.g., [3, 10, 26, 43]). Only few works implement models of mixed emotional expressions. The existing solutions usually compute new expressions in which single parameters are obtained by “averaging” the values of the corresponding parameters of expressions of certain “basic” emotions. Among others, the model called “Emotion Disc” [41] uses bi-linear interpolation between two closest basic expressions and the neutral one. In the Emotion Disc six expressions are spread evenly around the disc, while the neutral expression is represented by the centre of the circle. The distance from the centre of the circle represents the intensity of expression. The spatial relations are used to establish the expression corresponding to any point of the Emotion Disc. In Tsapatsoulis et al. [44] two different approaches are used: the new expression can be derived from basic one by “scaling” it. The second approach uses interpolation between facial parameters values of two closest basic emotions. A similar model of facial expressions was proposed by Albrecht et al. [1].

Different approach was proposed by Duy Bui [5]. She introduced the set of fuzzy rules to determine the blending expressions of six basic emotions. In this approach a set of fuzzy rules is attributed to each pair of emotions. The fuzzy inference determines the degrees of muscles contractions of the final expression in function of the input emotions intensities. Blending expressions of six basic emotions are also used in [23].

Different types of facial expressions were considered by Rehm and André [39]. by In a study on deceptive agents, they show that users are able to differentiate between the agent displaying an expression of felt emotion versus an expression of fake emotion [39]. Prendinger et al. [37] implement a set of procedures called *social filter programs*. In a consequence their agent is able to modulate the intensity of the expression according to the social context.

Comparing with other models we introduce the diversification of facial expressions in relation to their meaning, role, and appearance. Thus, another novelty of our system is that our agent is able to express different types of facial expressions (like inhibited, masked or fake expressions). Moreover, following the psychological evidence [15] complex facial expressions are computed by composing whole facial areas of any facial expression. Thus the final expression is combination of facial areas of other expressions. Finally we can create complex facial expressions not only in a case of six basic emotions but for any expression that was described by the researchers (e.g., embarrassment [22] or contempt [13]).

3 The Greta head animation system

Greta is an Embodied Conversational Agent (ECA) that communicates through her face and gestures while talking to the user. The head animation system, topic of this paper, is a process that computes the low-level animation of the agent head. For example it has to precisely determine which horizontal angle the head should rotate in order to perform a head-shake, or to determine which facial points have to be moved to show a particular facial expression. Figure 1 shows the general architecture of the system. The input data of the system is a file with an high-level description of communicative acts that the agent aims to communicate. The input file follows the format of the Affective Presentation Markup Language APML [33] (see Figure 2 for an example of an APML input file). APML is an XML-based language whose tags represent communicative acts. In the example of Figure 2 the APML tags surrounding the text specify that the agent is going to *announce* something (line 5) while showing a *sad* emotional face (lines 6 and 14). The APML tags give information about the speaker’s goals of conversation. That is, the enclosed sentences could be translated into a facial expression and/or head movements and/or gaze change [35]. The animation corresponding to APML tags is computed by the *Head/Gaze/Face Computation* module, explained in detail in Section 4. In some cases, for some values of the *affect* tag (for instance a complex emotion), this module yields the generation of the facial expression to the *Complex Facial Expressions Computation* module, described in detail in Section 5.

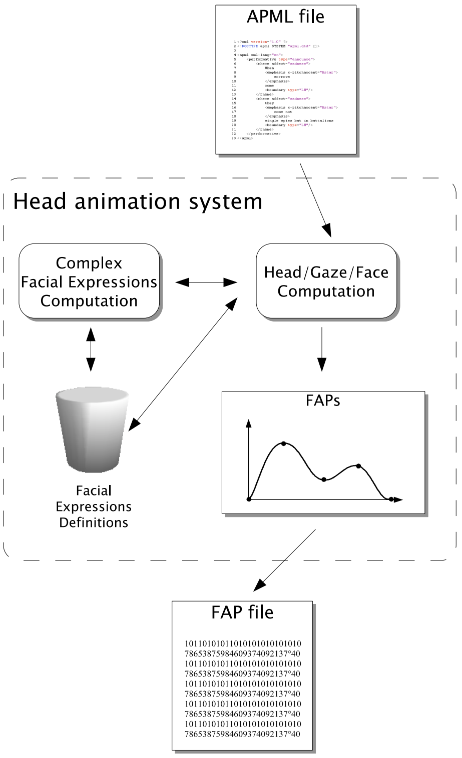
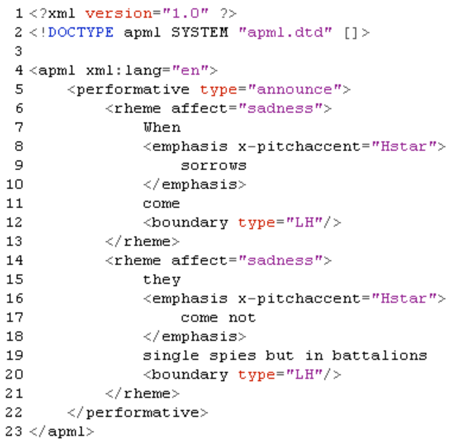


Figure 1. Low-level representation of the Greta’s face engine.

The output of the system is an animation file, that is a sequence of frames, and a wav file. In particular, our system produces an animation file following the MPEG4/FAP format [29, 32]. The standard defines some activation points on the agent’s face, called FAPs, and the way each FAP contributes to the deformation of the face area underneath it. A FAP file is a sequence of FAP frames, one frame for each time unit, and each FAP frame is a sequence of FAP values. Since this is a standard format, every talking head player implementing FAPs can playback the animation files generated by our engine.



```
<?xml version="1.0" ?>
<!DOCTYPE apml SYSTEM "apml.dtd" []>

<apml xml:lang="en">
    <performative type="announce">
        <rheme affect="sadness">
            When
            <emphasis x-pitchaccent="Hstar">
                sorrows
            </emphasis>
            come
            <boundary type="LH"/>
        </rheme>
        <rheme affect="sadness">
            they
            <emphasis x-pitchaccent="Hstar">
                come not
            </emphasis>
            single spies but in battalions
            <boundary type="LH"/>
        </rheme>
    </performative>
</apml>
```

Figure 2. Example of an APML input file.

4 Expressive computation of head/gaze/face

4.1 Expressivity

Many researchers (Wallbott and Scherer [47], Gallaher [18], Ball and Breese [2], Pollick [36]) have investigated human motion characteristics and encoded them into categories. Some authors refer to body motion using dual categories such as slow/fast, small/expansive, weak/energetic, small/large, unpleasant/pleasant. The expressivity of behaviour is "How" the information is communicated through the execution of some physical behaviour.

Greta is an expressive ECA, that is her animation can be qualitatively modified by a set of expressivity parameters affecting the physical characteristics of movements (like speed, width, strength, etc.). Starting from the results reported in [47] and [18], we have defined the expressivity by 6 dimensions:

- *Overall Activity* models the general amount of activity (e.g., passive/static or animated/engaged);
- *Spatial Extent* modifies the amplitude of movements (e.g., expanded versus contracted);
- *Temporal Extent* changes the duration of movements (e.g., quick versus sustained actions);
- *Fluidity* influences the smoothness and continuity of movement (e.g., smooth, graceful versus sudden, jerky);
- *Power* represents the dynamic properties of the movement (e.g., weak/relaxed versus strong/tense);
- *Repetitivity* models the tendency of the agent to replicate the same movement with short and close repetitions during time. Technical details on the implementation of these parameters can be found in [20].

Let us describe how each part of the *Head/Gaze/Face Computation* (see Figure 1) works.

4.2 Head model

The head model generates the animation of the head: a single movement corresponds to a change in head direction (up, down, left, etc.) while a composed movement is obtained by the repetition of a single movement (as in the case of head nod and shake). The quality of the head movement can be modified by varying the expressivity parameters, for example by increasing the *Spatial Extent* Greta's head movement will be wider. Variation in the *Temporal Extent* parameter changes the rotation speed: the smaller is such expressivity parameter the smaller is the rotation angle of the head. *Repetitivity* can cause one or more repetitions of the same movement; for example, it will increase the frequency of head nods/shakes.

Our agent follows the standard MPEG-4/FAP, so a head position is given by specifying the value of 3 FAPs, one for each axis, through a rotation vector:

$$RV = (HRx, HRy, HRz).$$

We define RV_{RP} the rotation vector that moves the head back to its *reference position*. A head movement HM is described by a sequence of keyframes where each keyframe is a couple (T, RV) containing a time label T and the rotation vector RV that specifies the head position at time T:

$$HM = ((T_0, RV_{RP}), (T_1, RV_1), ..., (T_{n-1}, RV_{n-1}), (T_n, RV_{RP})).$$

By default, a head movement starts and ends with the *reference position*, that is the first and last key frame correspond to the head position RV_{RP}. When two successive movements are computed we check if the first head movement needs to coarticulate into the next head movement or if it has time to go back to its reference position. The decision is based on the duration between successive head movements. If two head movements are too close to each other, the key frames to the *reference position* are deleted to avoid unnatural jerky movement. Let us consider two consecutive head movements:

$$HM_1 = ((T_{1_0}, RV_{RP}), (T_{1_1}, RV_{1_1}), (T_{1_2}, RV_{1_2}), (T_{1_3}, RV_{RP})),$$

$$HM_2 = ((T_{2_0}, RV_{RP}), (T_{2_1}, RV_{2_1}), (T_{2_2}, RV_{2_2}), (T_{2_3}, RV_{RP})).$$

For sake of simplicity, both movements perform rotations only around the x axis. Figure 3(a) shows the curve of the FAP HRx representing both movements HM_1 and HM_2. We calculate their temporal distance TD as:

$$TD = T_{2_1} - T_{1_2}.$$

If such a temporal distance is less than a given threshold, we consider both movements to be too close to each other and, in order to avoid jerky movements of the head, we delete the last key frame in HM_1 and the first key frame in HM_2 to obtain a smoother curve and then a better animation of the head. The new curve is shown in Figure 3(b). As explained in before the head movements can be modulated by the value of the expressivity parameters affecting the amplitude of their movement, as well as their speed and acceleration. Once all the key frames have been calculated they are interpolated to obtain the whole head movement. Further computation may be necessary to ensure correlation between head and eye movement (see Section 4.3.1).

4.3 Gaze model

The gaze model generates the animation of the eyes. It is based on statistical data obtained from the annotation of behaviour (smile, gaze direction, speaking turn, etc.) of dyads [30].

A belief network, embedded both types of information, is used to compute the next gaze direction. Personalized gaze behaviour is obtained by specifying temporal parameters of the belief network. Maximal and minimal time for mutual gaze, look at the other participant

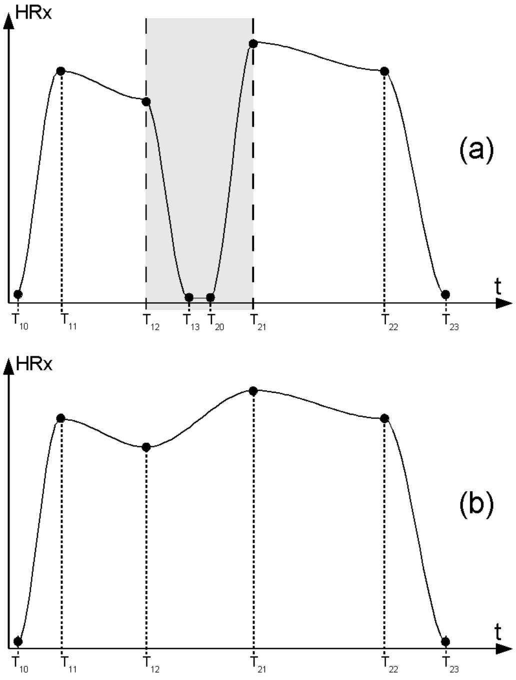


Figure 3. (a) Curves of two very close head rotations around axis x. The grey area shows the jerk in the head movement. (b) Key frames in $T1_3$ and in $T2_0$ have been deleted to obtain a smoother animation.

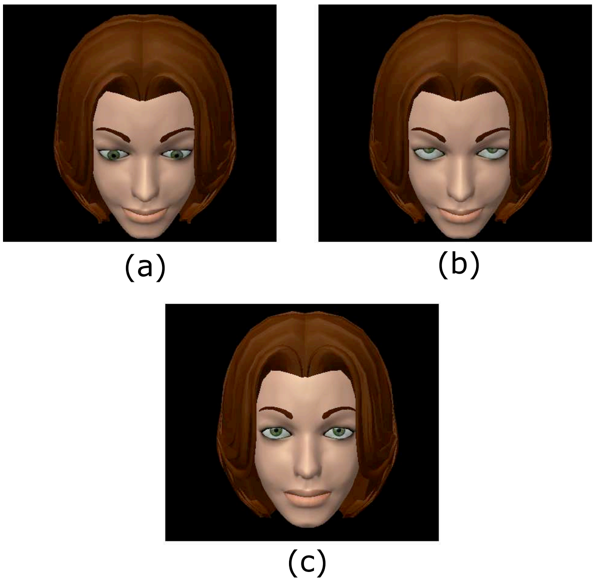


Figure 4. Example of an inconsistency between head and gaze. (a) Frame 1: head down and gaze away. (b) Frame 2: the head is still down but the eyes must perform a *LookAt* disappearing in the skull: inconsistency. (c) New Frame 2: the inconsistency is deleted forcing a rotation of the head.

and gaze away can be specified. This model computes the agent's gaze pattern as a temporal sequence of two possible states: *LookAt* and *LookAway*. *LookAt* means that the ECA gazes at the other participant (the user or an other agent in the virtual environment), whereas *LookAway* implies that the agent moves away her gaze. The result of the gaze model is a sequence of couples:

$$GAZE = ((t_0, S_0)...(t_n, S_n)),$$

where t_i and S_i are respectively the start time and the value of the i^{th} state ($S_i = 1$ means *LookAt* whereas $S_i = 0$ means *LookAway*). The gaze state *LookAt* corresponds to a precis direction while the gaze state *LookAway*is defined as negation of *LookAt*. In our algorithm the space is divided into 8 regions related to the user's head (up, up right, down, down left, etc.). Some communicative functions specifies the gaze should be direct to one of these regions; if no specification exists a region is chosen casually. Once a region is determined the exact eye direction is computed randomly. To ensure spatial coherency (the eyes do not move in every direction during a *LookAway*) a region is fixed for a certain duration.

4.3.1 Correlation between head and gaze movements

The result of the gaze model could be inconsistent with the animation of the head. Such inconsistency shows up when the directions of the head and of the gaze are too different causing an unnatural rotation of the eyes in the skull. Figure 4 shows such inconsistency. In Figure 4(a) the gaze of the agent is away (look down) and the head is down. The expression of sadness generates this gaze/head pattern. Figure 4(b) shows the next frame where the head is still down but the direction of the eyes changes because of a *LookAt*. Since the rotation of the head was quite wide, the iris of the eyes is no more visible creating an awkward animation. To remove all the inconsistencies between the gaze and the head movement we analyse the sequence $GAZE$ (deriving from the gaze model) and for each couple (t_i, S_i) we check the validity of the head position for each frame in the interval of time $[t_i, t_{i+1}]$, where t_{i+1} is the start time of the $(i+1)^{th}$ element of the sequence. A head position $RV = (HRx, HRy, HRz)$ (see Section 4.2) is *valid* if:

$$-th_x < HRx < th_x,$$

$$-th_y < HRy < th_y,$$

$$-th_z < HRz < th_z,$$

where th_x, th_y and th_z are respectively the threshold of the rotation around the axes x, y and z. When a *not-valid* position is found, the nearer key frames are modified (moved nearer to the *reference position*) and the interpolation recomputed to generate the new animation of the head. Figure 4(c) shows the same frame in Figure 4(b) where the inconsistency between the gaze and the head has been deleted. As we can see the head position has changed to allow the eyes to reach the direction defined by the *LookAt* and remain visible.

4.4 Face model

Depending on APML tags, the face model decides which facial expressions have to be performed by the agent. As explained in the introduction, a facial expression can be either a simple or a complex one. Simple expressions are directly retrieved from a static definition library (the *Facial Expressions Definitions* object in Figure 1). On the other hand, complex expressions are dynamically calculated by the *Complex Facial Expressions Computation* module which is presented in detail in section 5. In both cases, the simple or complex expressions are converted into a sequence of FAP values that are inserted into a data structure and will be interpolated afterwards.

As we explained before, our agent follows the standard MPEG-4/FAP, so a facial expression is specified by the value of the FAPs on the face. The first step to compute a facial animation is to define a sequence of keyframes. A keyframe is defined as a couple (T, FS) containing a time label T and facial shape FS that specifies the values of the FAPs of the face at time T. By default, each facial expression starts and ends with the *neutral expression* and it is characterized by four temporal parameters [25]:

- *attack*: is the time that, starting from the neutral face $FS_{neutral}$, the expression takes to reach its maximal intensity FS_1;
- *decay*: is the time during which the intensity of the expression lightly de-creases, usually to reach a stable value FS_2;
- *sustain*: is the time during which the expression is maintained, usually it represents the more visible part of the expression;
- *release*: is the time that the expression takes to return to the neutral expression $FS_{neutral}$.

A keyframe is computed for each temporal parameter and so, a facial expression animation FA can be defined as follows:

$$FA = ((T_{attack}, FS_1), (T_{decay}, FS_2),$$
$$(T_{sustain}, FS_2), (T_{release}, FS_{neutral})).$$

The final animation is obtained by interpolating between the resulting keyframes. Like for the head, when two consecutive facial expressions are computed we need to check their temporal distance. If such a distance is less than a given threshold, it means that the facial expressions are too close to each other and we need to delete the last keyframe of the first expression and the first keyframe of the second expression in order to avoid an abrupt return to the neutral face in between.

The facial animation depends also on the expressivity parameters. While computing the keyframes, the FAP values are modified according to the parameters. For example *Spatial extent* scales the FAP values of the expression; that is it changes the amplitude of the displacement of FAPs on the agent's face. *Temporal extent* increases (resp. decreases) the speed by which the expression appears: low (resp. high) values will make the expressions appear faster (resp. slower).

5 Complex Facial Expressions Computation

Our model of complex facial expressions is based on Paul Ekman's results [12–15]. We model complex facial expressions using a face partitioning approach. It means that different emotions are expressed on different areas of the face. More precisely, each facial expression is defined by a set of eight facial areas F$_i$, i= 1,..,8 (i.e., F$_1$ - brows, F$_2$ upper eyelids etc.). Then the complex facial expressions are composed of the facial areas of input expressions.

While analysing human facial expressions of emotions, Ekman distinguished between: modulating, falsifying, and qualifying an expression [15]. One modulates expressions by *de-intensifying* or *intensifying* them. For example, to intensify an expression one can change the intensity or duration of the expression. Falsifying a facial expression means to *simulate* it (to show a fake emotion), *neutralize* it (to show neutral face) or *mask* it. Masking occurs when one tries to hide "as much as possible" an expression by simulating another one. Finally, *qualification* means to add a fake expression (usually a smile) to a real one in order to express combination of both. In this case, the felt expression is not inhibited.

Using the model presented in this section we can generate the facial expressions of masking, as well as fake and inhibited expressions. The model generates different displays for these different types of expression. Complex facial expressions are obtained from the six basic emotions: anger, disgust, fear, joy, sadness, and surprise are described in the literature [13, 15]. Basing on it we have defined for each type of expression a set of fuzzy rules that describe its characteristic features in terms of facial areas. Each rule correspond to one basic emotion.

In the case of an input expression for which the complex facial expression is not defined explicitly by our rules (e.g. expression of contempt or disappointment) our algorithm chooses the most appropriate solution. This appropriateness is measured by analysing *visual resemblance* between expressions. For this purpose we introduced an innovative approach to compare two facial expressions. It is based on the notion of fuzzy similarity. In our approach any facial expression is described by a set of fuzzy sets. The main advantage of this approach is that slightly different expressions can be described by one label (like "joy" or "sadness"). Our algorithm compares two facial expressions attribute-after-attribute and then it composes single results into one value in the interval [0,1]. Finally, the values of similarity and the rules mentioned above are used to generate the complex facial expressions. Let us present in detail of our model.

5.1 Comparing Two Facial Expressions

The first step of the algorithm consists in establishing the degree of similarity between the input expression (i.e. the expression for which we want to find the complex facial expression) and the expressions of basic emotions. Let E_u and E_w be two emotions whose expressions we want to compare. Thus we want to establish the degree of similarity between *Exp(E_w)* and *Exp(E_u)*. In our approach each expression *Exp(E_i)* is associated with a set of fuzzy sets in terms of which all plausible expressions of emotion E$_i$ are defined. That is, for each numerical parameter (FAP) of an expression of emotion E$_i$ there is a fuzzy set that specifies a range of plausible values. Firstly, the value of similarity for each parameter (FAP) of *Exp(E_w)* and *Exp(E_u)* is established independently. The M-measure of resemblance *S*:

$$S(A, B) = \frac{(M(A \cap B))}{(M(A \cup B))}$$

where *A* and *B* are two fuzzy sets [4] is used in this case. Finally all values are combined by means of *Ordered Weighted Averaging (OWA)* operator (see [40] for detailed discussion).

5.2 Rules For Creation of Complex Facial Expressions

Several researchers have proposed a list of *deception clues* i.e. the features of expressions that are useful in distinguishing between fake and felt expressions [11, 12, 15]. At the moment, two of them are implemented in our model: *reliable features* and the *inhibition hypothesis*.

First of all humans are not able to control all their facial muscles. In a consequence expressions of felt emotions may be associated with specific facial features like: sadness brows [15] or orbicularis oculi activity in the case of joy [12]. Such *reliable features* lack in fake expressions as they are difficult to do voluntarily. For each basic emotion the features which are missing in fake expressions are known [12, 15].

On the other hand, people are not able to fully inhibit felt emotions. According to the *inhibition hypothesis* [12], the same elements of facial expressions which are difficult to show voluntarily in the case of unfelt emotions are also difficult to inhibit in the case of felt emotions. Finally, Ekman enumerates all facial areas that leak over the mask during the emotional displays management [15].

For each type of deception clues considered by us a separate set of rules has been developed. The first one - SFR$_{fake}$ - describes the features of a fake expression, while SFR$_{felt}$ - of a felt one.

In a case of the SFR$_{fake}$ the meaning of each rule is as follows: the more the input expression of E$_i$ is similar to the expression of E$_u$, the more possible is that facial areas of Exp(E$_i$) corresponding to reliable features of Exp(E$_u$) should not be used in the final expression. For example, in the case of sadness the following rule is applied: "the more the input expression is similar to sadness, the more possible is that the brows of the input expression should not be visible". Similarly, each rule of SFR$_{felt}$ describes the features which occur even in a covered expression of a felt emotion.

5.3 Generation of Complex Facial Expressions

Using our model different types of expression can be generated. Let us present the process of generation of a complex facial expression on the example of masking. Masking occurs when a felt emotion should not be displayed for some reason; it is preferred to display a different emotional expression. The expression of masking is composed from a fake expression that covers the expression of the real emotional state. Thus, both sets of rules SFR$_{felt}$ and of SFR$_{fake}$ should be applied in this case.

Let *B* be the set of the basic emotions (including neutral state) and Exp(E$_u$) be the expression corresponding to one of these emotions, E$_u$ $\in B$.

In the case of masking the input to the system consists in specifying two emotion labels: the felt one E$_i$ and the fake E$_j$. Both, E$_i$ and E$_j$ are specified in the APML input file.

In the *first step* our algorithm establishes the degrees of similarity between Exp(E$_i$), Exp(E$_j$) and all expressions of emotions that belongs to the set *B*. In a consequence we obtain two vectors [a$_k$] and [b$_k$], 1≤ a,b ≤ —B—, a$_k$, b$_k$ $\in [0, 1]$ of the degrees of similarity.

In the *second step* the deception clues for input expressions Exp(E$_i$), Exp(E$_j$) are established. For this purpose the sets of rules SFR$_{felt}$ and SFR$_{fake}$ are used. The vector [a$_k$] of felt expression E$_i$ is processed by SFR$_{felt}$, while the vector [b$_k$] of the fake expression E$_j$ is processed by SFR$_{fake}$. The SFR$_{felt}$ and SFR$_{fake}$ returns certain predictions about which parts of the face will (not) be visible in the masking expression.

The fake and felt parts of the final expression are considered separately. Finally, in the *last step* of the algorithm, for each facial area, the results of SFR$_{felt}$ and of SFR$_{fake}$ are composed in order to obtain the final expression. It is realized using another set of rules that takes as an input the outputs of precedent systems. The crisp output indicates the part of which expression (felt, fake or neutral) will be used in the final expression. The main task of this system is to resolve the eventual conflicts (i.e. the situation in which according to results of SFR$_{fake}$ and SFR$_{felt}$ different expressions should be shown in the same facial region). At the contrary, in the case in which neither felt nor fake emotion can be shown in a particular region of the face, the neutral expression is used instead.

Figure 5 presents the agent displaying the expression of a disappointment, that is masked by fake happiness. In the image on the right the parts of expression copied from the expression of disappointment are marked with blue and of happiness with red circles. We can notice that the absence of *orbicularis oculi* activity as indicator of fake happiness is visible on both images. Also the movement of brows can be observed, which is characteristic of disappointment. It is so because the expression of disappointment is very similar (according to the procedure described in section 5.1) to the expression of sadness. The facial areas F$_1$(forehead and brows) and F$_2$ (upper eyelids) cover the features of felt sadness that leak over the mask. As a consequence, they can be observed in inhibited sadness and thus they can be also observed in covered disappointment.

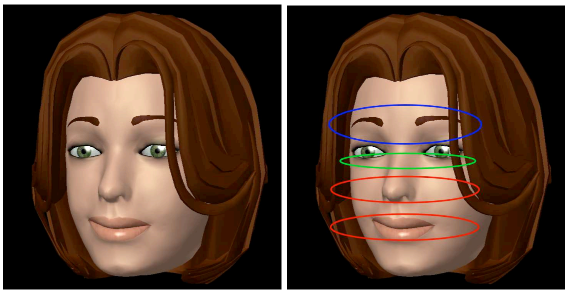


Figure 5. Example of a disappointment masked by a joy.

Similarly we can generate different complex facial expressions. Figure 6 shows two other examples of our algorithm's output. In the first row one can see: on the left the expression of contempt; on the right the same expression is inhibited. In the second row the expression of sadness is presented on the left, while the fake expression of sadness - on the right.

5.4 Evaluation

Complex facial expressions generated with our model were evaluated in a study based on the "copy-synthesis" method [6]. According to this approach the human behaviour is analysed and annotated by means of a high-level annotation schema. The animation of the agent is then obtained from the annotation. In one of such studies that is called EmoTV [9] different types of complex facial expressions were observed.

We generated a set of animations starting from two videos of the EmoTv video-corpus [9] that were annotated with different types of complex facial expressions. More precisely, four different animations were compared with each original video. The first two animations used simple facial expressions and body movements. Each of them displayed one of the two emotions indicated by the annotators. The two other animations used complex facial expressions that were created in two different ways: in the first one we used our model; in the second one the low-level annotation was used instead.

Then we evaluated the quality of the animations by asking subjects to compare them with the original videos.

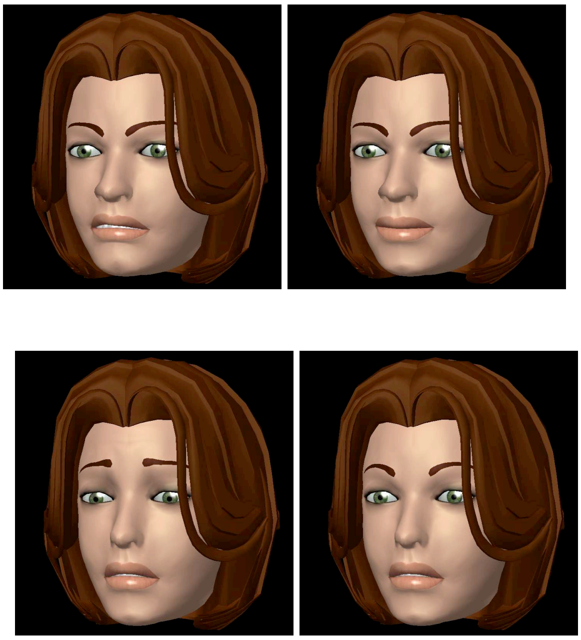


Figure 6. Examples of inhibited contempt (first row) and simulated sadness (second row).

The results are promising (see [6] for detailed results): The use of complex facial expressions created by our model has influenced the evaluation score significantly, especially in the case of animation in which facial expressions were easily observed. Animations created with our model obtained a satisfactory result in comparison with manually created animations of complex expressions. In one case (expression of masking) automatically generated expressions were evaluated even better than the manually defined complex expressions. In the second test the result was slightly worse, particularly in the no audio condition.

In another experiment we used different types of complex facial expressions in order to express different interpersonal relations between interlocutors (see [27] for details). We found different complex expressions generated using our model are recognized by humans and that these expressions comunicate different social signals [27].

6 Conclusions and Future

We have presented an expressive head animation system for ECAs. After giving a general overview of the system, we have focused on the implementation of two important aspects of behaviour: the expressivity of movement and the computation of complex facial expressions. Our head/gaze/face model generates facial expressions and coordinated head and gaze movements under the influence of some expressivity parameters.

Then we have described a model to compute complex facial expressions. Our model introduces the diversification of facial expressions. It builds different types of complex facial expressions. As a consequence, these different types of complex facial expressions can be distinguished by the user, because their appearance is different.

In the near future we are going to develop the head/gaze model to make the ECA pointing at objects in the environment with gaze. We will also integrate this model in a speaker/listener system for ECAs. We also plan to model other types of complex facial expressions and to implement other deception clues like *micro-expressions* and *time-related* deception clues. At the moment all expressions (basic and complex ones) are specified in the APML file. We aim at integrating our system with an Elicited-Emotion module which is responsible for the evaluation of an event and the emotion elicitation (see [28]).

REFERENCES

[1] I. Albrecht, M. Schroder, J. Haber, and H. Seidel, 'Mixed feelings: Expression of non-basic emo-tions in a muscle-based talking head', *Special issue of Journal of Virtual Reality on "Language, Speech & Gesture"*, (2005).

[2] G. Ball and J. Breese, 'Emotion and personality in a conversational agent', in *Embodied Conversational Characters*, eds., S. Prevost J. Cassell, J. Sullivan and E. Churchill, MITpress, Cambridge, MA, (2000).

[3] C. Becker, S. Kopp, and I. Wachsmuth, 'Simulating the emotion dynamics of a multimodal conversational agent', in *Affective Dialogue Systems*, eds., E. Andr, L. Dybkjr, W. Minker, and P. Heisterkamp, 154–165, Springer Verlag, (2004).

[4] B. Bouchon-Meunier, M. Rifqi, and S. Bothorel, 'Towards general measures of comparison of objects', *Fuzzy sets and systems*, **84**(2), 143153, (1996).

[5] T. Duy Bui, *Creating Emotions And Facial Expressions For Embodied Agents, PhD thesis*, University of Twente, Departament of Computer Science, Neslia Paniculata, Enschede, 2004.

[6] S. Buisine, S. Abrilian, R. Niewiadomski, J.-C. Martin, L. Devillers, and C. Pelachaud, 'Perception of blended emotions: From video corpus to expressive agent', in *The 6th International Conference on Intelligent Virtual Agents*, Marina del Rey, USA, (August 2006).

[7] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjlmsson, and H. Yan., 'Embodiment in conversational interfaces: Rea', in *Conference on Human Factors in Computing Systems*, Pittsburgh, PA, (April 15-20 1999).

[8] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, *Embodied Conversational Agents*, MIT Press, Cambridge, MA, 2000.

[9] L. Devillers, S. Abrilian, and J.-C. Martin, 'Representing real life emotions in audiovisual data with non basic emotional patterns and context features', in *Proceedings of First International Conference on Affective Computing & Intelligent Interaction*, pp. 519–526, Pekin, China, (2005).

[10] A. Egges, S. Kshirsagar, and N. Magnenat-Thalmann, 'Imparting individuality to virtual humans', in *First International Workshop on Virtual Reality Rehabilitation*, Lausanne, Switzerland, (2002).

[11] P. Ekman, *Non i volti della menzogna: gli indizi dell'inganno nei rapporti interpersonali, negli affari, nella politica, nei tribunali*, Giunti-Barbera, 1989.

[12] P. Ekman, 'Darwin, deception, and facial expression', *Annals of the New York Academy of Sciences*, **1000**, 205–221, (2003).

[13] P. Ekman, *The Face Revealed*, Weidenfeld & Nicolson, London, 2003.

[14] P. Ekman and W.V. Friesen, 'The repertoire of nonverbal behavior's: Categories, origins, usage and coding', *Semiotica*, **1**, 49–98, (1969).

[15] P. Ekman and W.V. Friesen, *Unmasking the Face. A guide to recognizing emotions from facial clues*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1975.

[16] M.G. Frank, P. Ekman, and W.V. Friesen, 'Behavioral markers and recognizability of the smile of enjoyment', in *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, eds., P. Ekman and E.L. Rosenberg, Oxford University Press, (1995).

[17] W.V. Friesen, *Cultural differences in facial expressions in a social situation: An experimental test of the concept of display rules*, University of California, 1972. Unpublished doctoral dissertation.

[18] P. E. Gallaher, 'Individual differences in nonverbal behavior: Dimensions of style', *Journal of Personality and Social Psychology*, **63**(1), 133–145, (1992).

[19] P. Gosselin, G. Kirouac, and F.Y. Dor, 'Components and recognition of facial expression in the communication of emotion by actors', in *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression*

Using the Facial Action Coding System (FACS), eds., P. Ekman and E.L. Rosenberg, 243–267, Oxford University Press, (1995).

[20] B. Hartmann, M. Mancini, and C. Pelachaud, 'Implementing expressive gesture synthesis for embodied conversational agents', in *The 6th International Workshop on Gesture in Human-Computer Interaction and Simulation*, VALORIA, University of Bretagne Sud, France, (2005).

[21] D. Heylen, 'Challenges ahead. head movements and other social acts in conversation', in *AISB 2005 - Social Presence Cues Symposium*, (2005).

[22] D. Keltner, 'Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame', *Journal of Personality and Social Psychology*, **68**, 441 – 454, (1992).

[23] J. Ktsyri, V. Klucharev, M. Frydrych, and M. Sams, 'Identification of synthetic and natural emotional facial expressions', in *ISCA Tutorial and Research Workshop on Audio Visual Speech Processing*, pp. 239–244, St. Jorioz, France, (2003).

[24] M. LaFrance and M. A. Hecht, 'Option or obligation to smile: The effects of power and gender and facial expression', in *The Social Context of Nonverbal Behavior (Studies in Emotion and Social Interaction)*, Cambridge University Press, (2005).

[25] M. Mancini, R. Bresin, and C. Pelachaud, 'From acoustic cues to an expressive agent', in *Gesture Workshop*, pp. 280–291, (2005).

[26] D.W. Massaro, M.M. Cohen, J. Beskow, S. Daniel, and R.A. Cole, 'Developing and evaluating conversational agents', in *First Workshop on Embodied Conversational Characters*, Lake Tahoe, CA, (1998).

[27] R. Niewiadomski, *A model of complex facial expressions in interpersonal relations for animated agents*, Ph.D. dissertation, University of Perugia, 2007.

[28] M. Ochs, R. Niewiadomski, C. Pelachaud, and D. Sadek, 'Intelligent expressions of emotions', in *Proceedings of First International Conference on Affective Computing & Intelligent Interaction*, Pekin, China, (2005).

[29] C. Pelachaud, 'Visual text-to-speech', in *MPEG4 Facial Animation - The standard, implementations and applications*, eds., Igor S. Pandzic and Robert Forcheimer, John Wiley & Sons, (2002).

[30] C. Pelachaud and M. Bilvi, 'Modelling gaze behavior for conversational agents', in *International Working Conference on Intelligent Virtual Agents*, Germany, (September 15-17 2003).

[31] C. Pelachaud, C. Peters, M. Mancini, E. Bevacqua, and I. Poggi, 'A model of attention and interest using gaze behavior', in *International Working Conference on Intelligent Virtual Agents*, Greece, (2005).

[32] E. Petajan, 'Facial animation coding, unofficial derivative of MPEG-4 standardization, work-in-progress', Technical report, Human Animation Working Group, VRML Consortium, (1997).

[33] I. Poggi, 'Mind markers', in *Gestures. Meaning and use.*, ed., N. Trigo M. Rector, I. Poggi, University Fernando Pessoa Press, Oporto, Portugal, (2003).

[34] I. Poggi, 'Interacting bodies and interacting minds', in *International Society for Gesture Studies - Interacting Bodies*, (2005).

[35] I. Poggi and C. Pelachaud, 'Performative facial expressions in animated faces', 155–188, (2000).

[36] F. E. Pollick, 'The features people use to recognize human movement style', in *Gesture-Based Communication in Human-Computer Interaction*, eds., Antonio Camurri and Gualtiero Volpe, number 2915 in LNAI, 10–19, Springer, (2004).

[37] H. Prendinger and M. Ishizuka, 'Social role awareness in animated agents', in *Proceedings of the fifth international conference on Autonomous agents*, pp. 270–277, Montreal, Quebec, Canada, (2001).

[38] *Life-Like Characters*, eds., H. Prendinger and M. Ishizuka, Cognitive Technologies, Springer, 2004.

[39] M. Rehm and E. Andr, 'Catch me if you can - exploring lying agents in social settings', in *International Conference on Autonomous Agents and Multiagent Systems*, pp. 937–944, (2005).

[40] M. Rifqi, *Mesures de comparaison, typicalit et classification d'objets flous : thorie et pratique*, Thse, 1996.

[41] Z. Ruttkay, H. Noot, and P. ten Hagen, 'Emotion disc and emotion squares: tools to explore the facial expression face', *Computer Graphics Forum*, **22**, 49–53, (March 2003).

[42] C. Saarni and Hannelore Weber, 'Emotional displays and dissemblance in childhood: Implications for self presentation', in *The Social Context of Nonverbal Behavior (Studies in Emotion and Social Interaction)*, Cambridge University Press, (2005).

[43] K. R. Thrisson, *Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills*, Ph.D. dissertation, MIT Media Laboratory, 1996.

[44] N. Tsapatsoulis, A. Raouzaiou, S. Kollias, R. Crowie, and E. Douglas-Cowie, 'Emotion recognition and synthesis based on mpeg-4 faps', in *MPEG4 Facial Animation - The standard, implementations and applications*, eds., Igor S. Pandzic and Robert Forcheimer, John Wiley & Sons, (2002).

[45] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt, 'Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes', in *Conference on Human factors in computing systems*, pp. 301–308, New York, NY, USA, (2001). ACM Press.

[46] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung, 'Gaze-2: conveying eye contact in group video conferencing using eye-controlled camera direction', in *Conference on Human factors in computing systems*, pp. 521–528, New York, NY, USA, (2003). ACM Press.

[47] H. G. Wallbott and K. R. Scherer, 'Cues and channels in emotion recognition', *Journal of Personality and Social Psychology*, **51**(4), 690–699, (1986).

4th International Symposium on Imitation in Animals and Artifacts

Imitation facilitates transmitting culture practices and ideas from generation to generation enabling humans, animals, and now robots, to learn skills others have already mastered. By avoiding the lengthy period of trial-and-error to accomplish new tasks, imitation is thus a very efficient learning method, and also a very intuitive way to program robots by teaching.

The mechanisms of imitation and social learning are not well-understood, and the connections to social interaction, communication, development, and learning are deep, as recent research from various disciplines has started to uncover. Comparison of imitation in animals and artifacts reveals that easy tasks for machines can be hard tasks for animals and vice-versa. However, computational complexity issues do not explain, by themselves, the existence or not of imitation behaviours in animals, and the integration of higher level cognitive capabilities like agent's goals, intentions and emotions, may play a fundamental role in explaining these differences.

This interdisciplinary workshop will bring together researchers from neuroscience, brain imaging, animal psychology, computer science and robotics to examine the latest advances to imitation, aiming to further advance our understanding of the underlying mechanisms. We hope that the workshop will contribute to the advance in research in imitation and a better integration between the several scientific disciplines.

The symposium will consist of invited talks, regular presentations and short presentations. It is our privilege to have three distinguished invited speakers: Marcel Brass from the Ghent University, Belgium, speaking on the neuronal mechanisms of imitation and Nicola McGuigan from the Heriot-Watt University, Scotland speaking on imitation in children.

We would like to acknowledge the financial support provided by euCognition – the European Network for the Advancement of Artificial Cognitive Systems and by the EU-Project RobotCub, to all the people who have contributed towards making this symposium happen, the programme review committee and, above all, the authors and participants.

José Santos-Victor, Manuel Lopes, Alexandre Bernardino (SymposiumCchairs)

Programme committee: Alexandre Bernardino, (IST, Portugal); Andrew Meltzoff, (U. Washington, USA); Aris Alissandraiks, (Hertfordshire, UK); Aude Billard, (EPFL, CH); Bart Jansen, (VUB, Belgium); Brian Scassellati, (Yale, USA); Chana Akins, (Kentucky, USA); C. L. Nehaniv, (Hertfordshire, UK); Frédéric Kaplan, (EPFL, Switzerland); Francisco Lacerda, (Stockholm, Sweden); Giorgio Metta, (Genova, IT); Harold Bekkering, (Nijmegen, NL); Heiko Wersing, (Honda R.I., Germany); Irene Pepperberg, (Harvard, USA); Jacqueline Nadel, (CNRS, France); Jochen J. Steil, (Bielefeld, Germany); Joanna Bryson, (Bath, UK); José Santos-Victor, (IST, Portugal); K. Dautenhahn, (Hertfordshire, UK); Ludwig Huber, (Vienna, Austria); Manuel Lopes, (IST, Portugal); Max Lungarella, (Tokyo, JP); Monica Nicolescu, (Nevada, USA); Nicola McGuigan, (St. Andrews, Scotland); Rui Prada, (IST, Portugal); Thomas R. Zentall, (Kentucky, USA); Tony Belpaeme (Plymouth, UK); Yukie Nagai, (Bielefeld, Germany); Yiannis Demiris,(Imperial College, UK).

euCognition
www.euCognition.org

RobotCub.org

Imitation and self/other distinction

Marcel Brass
Department of Experimental Psychology, Ghent University

There is converging evidence from different fields of cognitive neuroscience suggesting that the observation of an action leads to a direct activation of an internal motor representation in the observer. It has been argued that these shared representations form the basis for imitation, action understanding and mentalizing. However, if there is a shared representational system of perception and action, the question arises how we are able to distinguish between intentionally formed motor representations and externally triggered motor plans. I will first outline empirical evidence and theoretical accounts supporting the idea of shared representations. Then I will review neurological data as well as data from social psychology and cognitive neuroscience suggesting that self/other distinction is a crucial requirement of a shared representational system. Finally, I will present recent findings showing that the mechanisms involved in the control of shared representations share neural resources with social cognitive abilities such as action understanding and mentalizing. Taken together, these data point to the fundamental role of self/other distinction in social cognition.

Imitation of causally-opaque versus causally-transparent tool use by 3- and 5-year-old children

Nicola McGuigan
School of Life Sciences, Heriot Watt University
Andrew Whiten, Emma Flynn, and Victoria Horner
School of Psychology, University of St Andrews

We explored whether the tendency to imitate or emulate is influenced by the availability of causal information, or the amount of information available in a display. Three and five-year-old children were shown how to obtain a reward from either a clear or an opaque puzzle-box by a live or video model. Each demonstration involved two different types of actions. The first stage involved causally irrelevant actions and the second stage involved causally relevant actions. When presented with the clear box it could clearly be seen that the actions were irrelevant as the causal information was available. In contrast this information was not available with the opaque box, potentially making discrimination between irrelevant and relevant actions difficult. We predicted that the 3-year-olds would imitate with both boxes, whereas the greater cognitive sophistication and causal understanding of the 5-year-olds would allow them to switch between imitation and emulation depending on the availability of causal information. However, the results showed that both 3-and 5-year-old children imitated the irrelevant actions regardless of the availability of causal information following a live demonstration. In contrast the 3-year-olds employed a more emulative approach when the information available in the display was degraded via a video demonstration containing the puzzle box and the actions of the model only. The results indicated that the 5-year-olds were unaffected by the degraded information and continued to employ an imitative approach. We suggest that imitation is such an adaptive human strategy that it is often employed at the expense of efficiency.

From exploration to imitation: using learnt internal models to imitate others

Anthony Dearden and **Yiannis Demiris**[1]

Abstract. We present an architecture that enables asocial and social learning mechanisms to be combined in a unified framework on a robot. The robot learns two kinds of internal models by interacting with the environment with no *a priori* knowledge of its own motor system: internal object models are learnt about how its motor system and other objects appear in its sensor data; internal control models are learnt by babbling and represent how the robot controls objects. These asocially-learnt models of the robot's motor system are used to understand the actions of a human demonstrator on objects that they can both interact with. Knowledge acquired through self-exploration is therefore used as a bootstrapping mechanism to understand others and benefit from their knowledge.

1 Introduction

A robot, like humans and other animals, can learn new skills and knowledge both asocially, by interacting with its environment, and socially, by observing the actions of other agents [23, 20]. Interaction enables a robot to learn basic low-level models about its own motor system - for example, the appearance of its motor system and how it is controlled [1]. There is, however, a limit to what a robot can learn efficiently just from its own actions. To learn higher-level models, involving sequences of actions or the position of interesting objects for example, the role of other agents in the robot's environment becomes important. Social learning mechanisms such as imitation have been shown to be a powerful way to transfer knowledge from one agent to another [5, 22]. In robotics this has the particular advantage of relieving the user of the necessity of programming hard-coded knowledge, and instead allowing them to teach actions or movements by demonstration.

Many existing asocial and social models of learning in robotics are based, to varying degrees, on psychological or neuroscientific models of learning in animals, and in particular humans, e.g. [24, 18, 4]. The benefit of turning to the biological sciences for inspiration in robotic learning architectures is clear. Human infants are capable of effortlessly combining learning from both their own interactions, and the actions of a caregiver. Both asocial and social learning methods have previously been studied separately in robotics. In this paper, we present an architecture that enables these learning mechanisms to be combined in a unified framework. The underlying components of this architecture are internal models, internal structures or processes that replicate the behaviour of the robot's environment [11]. In this work we describe how the robot can learn two specific kinds of internal models: Internal Object Models (IOMs), which model the state of objects such as the robot's or a demonstrator's motor system, and Internal Control Models (ICMs), which model how the state of these objects can be controlled by the robot.

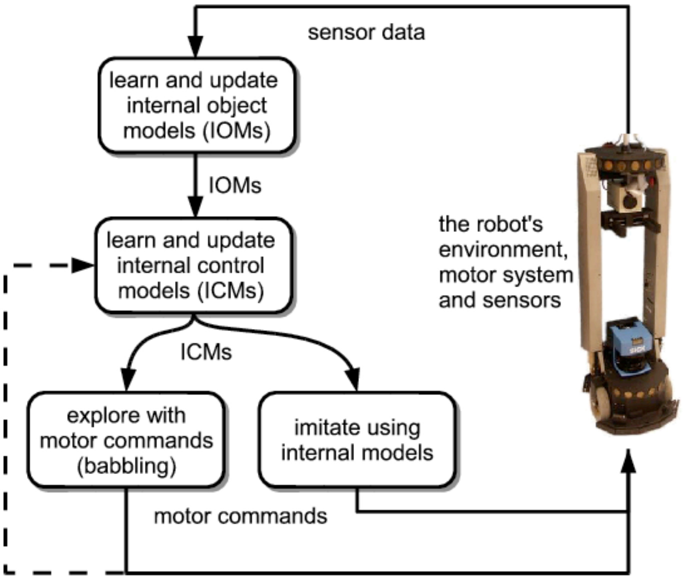

sensor data
learn and update internal object models (IOMs)
IOMs
learn and update internal control models (ICMs)
ICMs
explore with motor commands (babbling)
imitate using internal models
motor commands
the robot's environment, motor system and sensors

Figure 1. Overview of the learning software.

Drawing inspiration from motor babbling in infants [13], a system is presented that enables a robot to autonomously learn internal models with no *a priori* knowledge of its motor system or the external environment. Using the HAMMER architecture [5], the models that the robot learns of its own motor system are used to understand and imitate the actions of a demonstrator. Although learning is possible from observing movements, for example gestures, that do not involve interacting with objects, we are particularly interested in object manipulation.

Figure 1 shows an overview of the software components controlling the robot. Although the results are divided between the sections in this paper, each component runs simultaneously on the robot. Figure 2 shows the experimental setup. The robot used was an Activemedia Peoplebot, a mobile robot with a pan-tilt camera and a gripper.

[1] Department of Electrical and Electronic Engineering
BioART group, Imperial College London
E-mail: {anthony.dearden99, y.demiris}@imperial.ac.uk

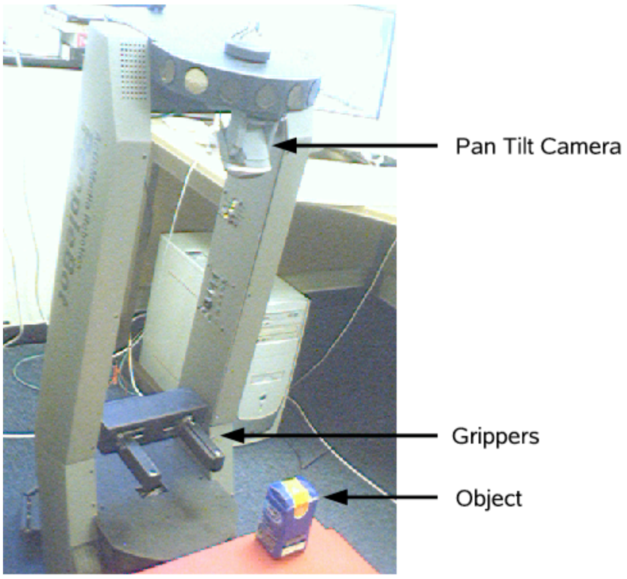


Figure 2. The experimental setup.

2 Discovering internal object models from visual data

Before a robot can learn *how* to control its environment, it needs to be able to *model* its environment. The robot's environment here is considered to consist of:

1. Its own motor system;
2. External, independent objects that its motor system can interact with;
3. The motor system of other agents.

IOMs are used by the robot to track and represent the state of these objects. There are clearly more properties that could be modelled, such as the position of walls, but these are not needed by a robot to imitate actions applied to objects.

In this work we are interested in vision-based robots - vision offers the richest information about the scene, despite the complexities involved in processing. A visual tracking system such as colour histogram-based tracking or even a full 3D tracking system could be used to find and track objects. The robot is much more autonomous, however, if it can discover objects for itself. Instead of being told about the appearance of objects, it would be able to learn about their appearance from the low-level vision data it receives. In [6, 14], visual knowledge acquired through experimentation and segmentation of motion history images is used at the image processing level to find interesting regions, which can be classified as objects. The focus in this work, however, is not currently on how new objects could be discovered and classified through interaction, but how they can be controlled and and used for imitation.

Algorithm 1 runs online to learn IOMs, with low-level input from the movement of pixel-level features in the scene tracked using the KLT optical flow algorithm [12]. Instead of calculating the optical flow for every point in the image, which would be inefficient and inaccurate, only corner features are tracked; these points are the easiest to track robustly. New points are automatically tracked and dropped as the robot's camera moves or new objects enter the scene.

Algorithm 1 Learning IOMs from optical flow data

```
• The input is a list of tracked optical flow points. Each point, p, is defined by its position and velocity in 2D space, {x,y,dx,dy}.
• The output is a list of objects. Each object is defined as the mean and covariance of its state, O = {X,Y,DX,DY}.
• If objects have previously been detected:
    – Given the previous state of the object, O[t-1], estimate its current state, O[t]. This prediction can be done using basic dynamic information, or if they have already been learnt, using a forward prediction from the internal models given the previous motor commands.
    – For each optical flow point, on each existing object, O[t], calculate the probability this point is part of that object - P(p | O[t]).
    – If P(p | O[t]) is greater than a threshold probability, pthresh, assign it to object O.
• Whilst there are unassigned points:
    – Create a new object $O_{new}$ using one unexplained point as a 'seed'.
    – Add other points for which P(p | $O_{new}$) is greater than the threshold probability, pthresh.
    – Update the mean and covariance of the object's state.
    – Repeat until all points are modelled, or no more points can be successfully modelled.
• Update the mean and covariance of each object's state with the new sensor data.
```

Algorithm 1 details how the IOMs are created and tracked by recursively clustering tracked points together. Unlike other clustering algorithms, such as K-means, the number of clusters does not need to be specified beforehand - this is important, because the robot should be capable of adapting to different numbers of objects. Instead, a probabilistic threshold of the variation in optical flow determines when points are added to or removed from IOMs - a value of 0.7 was found to work well.

The shape of objects can be estimated by fitting a convex hull to the clustered points, and by using the mean and the covariance of all optical flow points clustered to an object. The elements of the state vector of an IOM is defined by its position, size and shape. It is not just objects that can be tracked by this algorithm; the pan and tilt movement of the camera is tracked by clustering according to tracked points' velocities.

Clearly, objects cannot be detected unless they move. If the objects are part of the robot's own motor system, then it can discover them as it issues motor commands. If they are objects the robot could only interact with indirectly (such as the object in figure 3), then the robot has to either nudge into it, or be shown to it by a human teacher by shaking or waving the object.

Figure 4 shows the tracking of objects in an experiment. The robot's grippers are detected as soon as it starts to explore its motor system. The human hand and the object is detected when the human teacher moves. Figure 5 shows how the robot can also detect non-motor system objects by disturbing them with its own motor system.

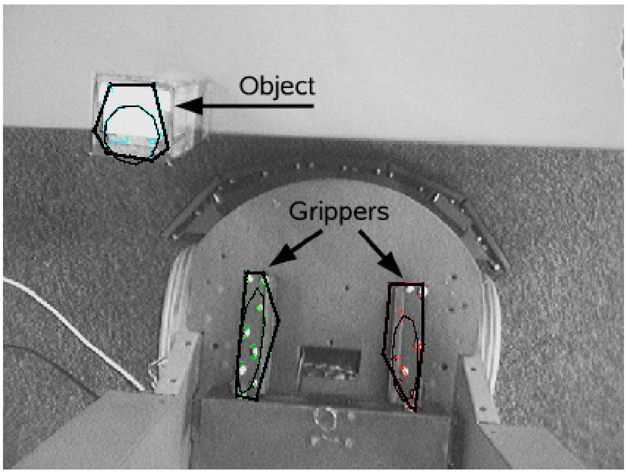


Figure 3. Moving image regions are clustered together; these regions are the robot's IOMs - internal models of where objects are in the scene. In this example, the grippers were moved by the robot, and the biscuit box object was shaken by a human demonstrator to make the robot aware of it. The thick black lines are the convex hull, and the thin ellipse shows represents the mean and covariance of the optical flow points' positions.

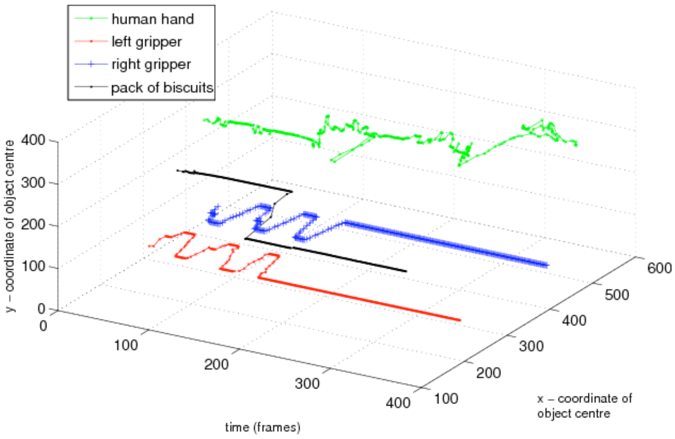


Figure 4. The movement of the IOMs in an experiment, as the grippers open and close and a human hand pushes a box of biscuits.

2.1 Classifying IOMs

A robot cannot imitate until it knows:

1. What it should imitate with;
2. Who to imitate;
3. What objects the imitation should involve;

This is equivalent to classifying objects in the environment according to how they can be controlled. The three kinds of IOMs are: *self* IOMs, objects that are part of the robot's own motor system and can be directly controlled; *demonstrator* IOMs, objects that are part of the demonstrator's motor system and cannot be controlled; and *shared* IOMs, objects that both the demonstrator and the robot can control indirectly. The imitation task considered here is for the robot to replicate, using its own motor system, the actions that the demonstrator takes on a shared object.

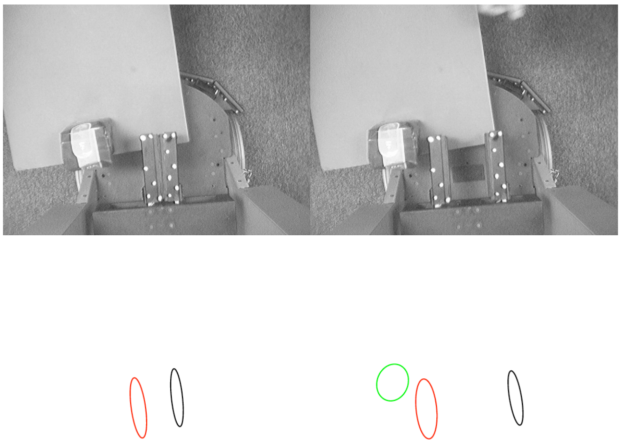


Figure 5. The robot can discover objects by moving them with its own motor system. The top images show frames from the robot 'babbling' in the environment. The bottom frames show the IOMs the robot has discovered before and after the movement.

The robot can learn to distinguish self IOMs from the other IOMs using the ICMs it has learnt for how to control IOMs. If a robot can directly control the state of an IOM, then it can classify it as its own motor system. Differentiating between active, *demonstrator* IOMs and passive, *shared* IOMs is more difficult because the robot can control neither. To solve this problem, the order in which objects are discovered is used. *Shared* IOMs do not move of their own accord, and therefore must be discovered by either being moved by the demonstrator or the robot. Therefore if an object is discovered close (less than 10 pixels) to the position of an existing object, it is classified as a *shared* IOM.

3 Internal control models

ICMs are used by a robot to model and learn how its motor command changes the state of IOMs. They are used as *forward models* to predict the consequences of its motor actions, or as *inverse models* to estimate the motor commands that will lead to a desired object state [1]. Coupling inverse and forward models gives a robot the ability to perform internal simulations of actions before physically executing them; through the *Simulation Theory* approach of the HAMMER architecture, these internal simulations can be used for action recognition and imitation [8, 17, 4].

A learnt ICM will not be able to completely accurately model a robot's motor system - errors will occur because of incorrect models, insufficient or noisy training data or the necessarily simplified internal representations of the model. The system that is being modelled may itself be stochastic. To overcome this uncertainty, it makes sense for an ICM to include information regarding not just its prediction, but how accurate it expects that prediction to be. This inaccuracy can be modelled by representing the internal model as a a joint probability distribution across the motor commands and and the state of elements of the robot's environment. The uncertainty in the model can be estimated from the variance of this distribution. Giving the robot information about the uncertainty of its internal models enables it to estimate how accurate, and therefore how useful, its internal models' predictions are - if multiple models are learnt, their predictive ability can be compared using the variance of their predictions. Section 5 shows how the robot can also use the variance in prediction to guide its exploration.

The basic elements of ICMs are the robot's motor commands and the state of the objects it has discovered - which are either part of its motor system or other objects. ICMs represent the causal structure of how these elements interact as a Bayesian network [19]. Bayesian networks are used in [7] to model how infants develop and test causal relationships. Here, we have taken this idea and applied it to the motor system of the robot. Figure 8 in section 4 shows an example of the Bayesian network structures that the robot learns. The motor commands and state of the IOMs are the random variables (nodes) in the Bayesian network, and the causal relationships between them are represented with arcs. The Bayesian network represents a learnt probability distribution across *N* possible motor commands, $M_{1:N}[t-d]$, the current states and previous states of the each object $S_x[t]$, $S_y[t]$, $S_{dx}[t]$, $S_{dy}[t]$, and the state of the proprioception feedback from the robot (e.g. gripper touch sensors). The variable *d* represents the delay between a motor command being issued and robot's state changing; in real robotic systems it cannot be assumed that the effect of a motor command will occur after just one time-step, so this is a parameter that the robot must model and learn. Table 1 shows the possible components of each internal model's Bayesian network. A benefit of using Bayesian networks to represent internal models is that their causal structure is understandable by a human. They can therefore be used to verify the correctness of what the robot is learning.

Random variable	Description
$M_{1:N}[t-d]$	Motor commands for *N* degrees of motor freedom, with different possible delays, *d*
$S_x[t]$, $S_y[t]$, $S_{dx}[t]$, $S_{dy}[t]$ ...	The state of each object - its position and velocity. For more complex objects, more statistical information can be calculated from its convex hull
$S_x[t-1]$, $S_y[t-1]$, $S_{dx}[t-1]$, $S_{dy}[t-1]$ ...	The state of each object at the previous time step
$P_1[t]$, $P_2[t]$	Proprioception information from other sensors, such as the touch sensors on the robot's grippers

Table 1. The variables the robot can use for its internal model. The robot has to learn Bayesian network structures and parameters using these variables as nodes on the network.

3.1 Learning through exploration

Practically any environment a robot works in will change, or have properties which cannot be modelled beforehand. Even if the environment is assumed to be completely predictable, endowing the robot with this knowledge may be beyond the abilities or desires of its programmer. A truly autonomous robot, therefore, needs to be able to learn and adapt its own internal models of its external environment. Unlike most machine learning situations, a robot has active control over the commands it sends to its as yet unknown motor system; this situation, where a learner has the ability to gather its own training data, is referred to as active learning [9]. Having the ability to interact with the system you are trying to model has the advantage that the data can be selected either to speed up the learning process, or to optimise the learnt model to be most useful for a particular task. The simplest way for a robot to learn about its environment through interaction is to issue random motor commands. This 'motor babbling' was used to learn internal models for a robot's grippers in [1]. A more sophisticated technique is to use an estimate of the ICM's prediction variance as function of motor command, $C\,(m,t)$. The actual motor command issued is the one expected to minimise this error. This technique was used to learn the control of a pan-tilt unit on both a real robot [2] and a camera in a football game simulation [3].

The decisions a robot makes about how to interact with the environment become more complex as more degrees of freedom (DOF) of the motor system or more exploration strategies are introduced. The robot has to decide what DOF or objects to learn about, not just what motor commands to send to its motor system. Instantly exploring all DOF at same time would take exponentially longer as the number of exploration possibilities increases. It would also lead to many more internal models having to be learnt simultaneously, which is computationally expensive. A developmental approach can be used to control how a robot explores its environment; more specifically, the robot needs to be able to decide on two things:

- When should the current exploration strategy be stopped?
- What should the next exploration strategy be?

We want the robot to realise when its current exploration strategy is not increasing the quality of the models it is learning. This information is available from the model learning system as the rate of change of the most accurate model's prediction variance, $C\,(m,t) - C\,(m,t-1)$. When this approaches zero, the robot knows the current exploration strategy is not improving the quality of the model. This is similar to using a 'meta-model' to estimate a predicting model's error to guide exploration [18].

The second question relates to what the robot should do next. The robot's goal is to learn models that explain how objects in its environment move. In the absence of any human intervention, the only cause of this can come from the robot's own interventions. In this situation, the robot can keep on exploring new degrees of freedom. Currently the degrees of freedom a robot explores are released in order of their distance from the vision system (camera movement, gripper movement then robot wheel movement).

3.2 Online learning of multiple internal models

ICMs consist of a structure, which represents how particular motor commands affect particular states of objects, and the parameters of the particular probability distribution being used for the model. Learning the parameters of a particular model is an online learning problem, with motor commands being the input data and IOMs' states being the output data. In the results here two types of distributions were used to represent the conditional probability distributions of the Bayesian network. For discrete motor commands such as the gripper controls, Gaussian distributions were used. The mean and the variance of the distribution are estimated recursively as:

$$\mu\,[t] = \frac{t}{t+1}\mu\,[t-1] + \frac{1}{t+1}S\,[t]$$

$$C\,[t] = \frac{t}{t+1}C\,[t-1] + \frac{1}{t+1}\left(S\,[t] - \mu\,[t]\right)^2$$

For continuous motor commands such as the robot's pan-tilt unit control the conditional probability distributions can be represented using the non-parametric LWR algorithm [15]. The results of previous trials are stored in memory and used to predict the consequence of future trials by performing linear regression on the set of data in

Algorithm 2 Learning multiple ICMs

- For the current motor command(s) being explored, multiple internal models are formed for the motor system. Table 1 shows the search space for possible model structures for a given motor command.
- At each timestep, the state of objects, $s_1 ... s_n$, in the scene is estimated by the vision system using algorithm 1.
- Each model predicts what it expects the states of the objects and interactions to be given the previous motor command. This is given as a Gaussian distribution: $P(S_1 ... S_n \mid M[t-d] = m) \sim N(\mu, C)$
- The likelihood of each model's prediction is calculated: $P(S_1 ... S_n = s_1 ... s_n \mid M[t-d] = m)$. This gives a metric for how well each candidate model is performing.
  - If processing or memory resources are limited, models with consistently low scores can be removed, as they are unable to predict accurately.
  - Objects which are moving in an unpredictable way, such as humans or objects they are interacting with, will have low likelihoods for all model predictions. This can be used by the robot to find objects which are not part of its motor system, which it may want to interact with.
- If the variance of the most accurate model's prediction converges, i.e. $C(m,t) - C(m,t-1) \approx 0$, then the robot's exploration of this motor command is not improving the accuracy of model. This is the cue to try a new exploration strategy.

memory, which is weighted according to its distance from the query point. Various other distribution types exist that can be learnt online but these methods were chosen principally for their quick convergence properties and ease of implementation [16].

The learnt structure of the Bayesian network represents which motor commands control which objects. The task of the robot is to search through the space of structures connecting every possible random variable to find the one that maximises the likelihood of the sensor data given the evidence, which here is the state of the objects given the sensor data. In this situation, learning the structure is simplified by the fact that the most recently observed change can be most likely explained by the most recent motor command issued. Furthermore, motor commands are always the parent node of the Bayesian network, as none of the other variables being modelled can influence it.

The online internal model learning system works by simultaneously training multiple possible internal model structures, and is described in algorithm 2. One difference between the models learnt here and those learnt by similar systems such as mixture of experts [10], is that there is no need for a responsibility estimator module to decide when each individual internal model should be used. Instead, as each model learns to estimate what the variance of its prediction is, $C(m,t)$, the 'responsible' model is chosen as the one with the smallest variance for a given prediction.

As multiple ICMs are trained, their prediction variance converges. In the experiments performed here, using models for estimating different delays in the motor-sensor system, the model which predicts most accurately is for the delay *d*=5 timesteps, equivalent to 0.33 seconds. This is reasonable given the latencies of the motor system and the lags which are present in the vision capture system. Figure 6 shows how this model's prediction varies as it is being learnt . The error bars on the graph show the variance in the prediction, $C(m,t)$.

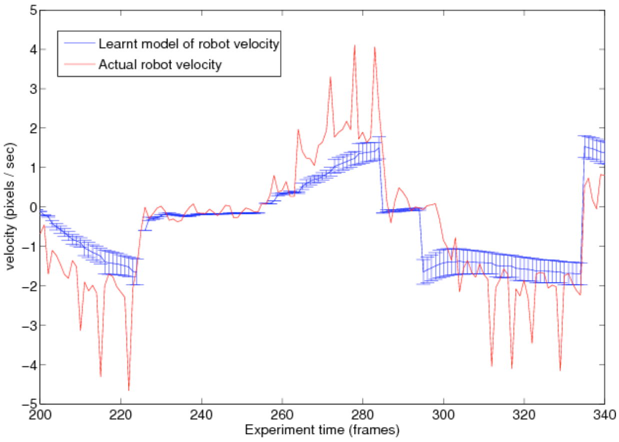


Figure 6. The robot learns online the mean and variance (shown with the error bars) of its velocity as it 'babbles' forwards and backwards. This is the prediction from the most accurate model, for which *d*=5. The large spikes in the actual data are because of dropped frames from the camera; the robot models this as noise.

Figure 7 compares two model structures being learnt for the *wheel velocity* motor command, which moves the robot forwards or backwards. Interestingly, the model it learns relates to how the motor command affects the position of objects in its environment: moving forward makes objects in front of it move closer. Figure 8 shows the structures of the internal models which the robot learns to be the most accurate for predicting the effects of its *gripper* and its *wheel velocity* motor commands.

This learning system is similar to the HAMMER architecture [5], used by the robot to perform imitation with learnt models in section 4, as it involves multiple competing internal models. The difference when learning is that the command fed to the motor system is not related to the models' predictions. Instead the predicted variance, $C(m,t)$, and its rate of change, $C(m,t) - C(m,t-1)$, is used by the active learning system to control how the robot interacts with the environment.

4 Imitating interactions using learnt internal models

The previous sections introduced the two types of internal models a robot learns from exploration: models of the objects in its environment, IOMs, and models of how to control them, ICMs. The HAMMER architecture presented here allows the robot to use these models to learn how to manipulate objects by observing the actions a demonstrator takes; we assume here that the robot has already learnt to classify IOMs, as discussed in section 3, so it knows the object to imitate (the demonstrator), the object to act with, and the *object* the action is performed on.

ICMs can be used directly as inverse models to imitate movements [1], but their usefulness is limited; they only model low-level motor commands and the sensory consequences over short time periods. The robot is unable to learn long term models from exploration because the motor commands it has available to explore with are all low-level commands: we are not assuming the existence of higher-level pre-programmed 'motor primitives' that control complex movements over multiple degrees of freedom.

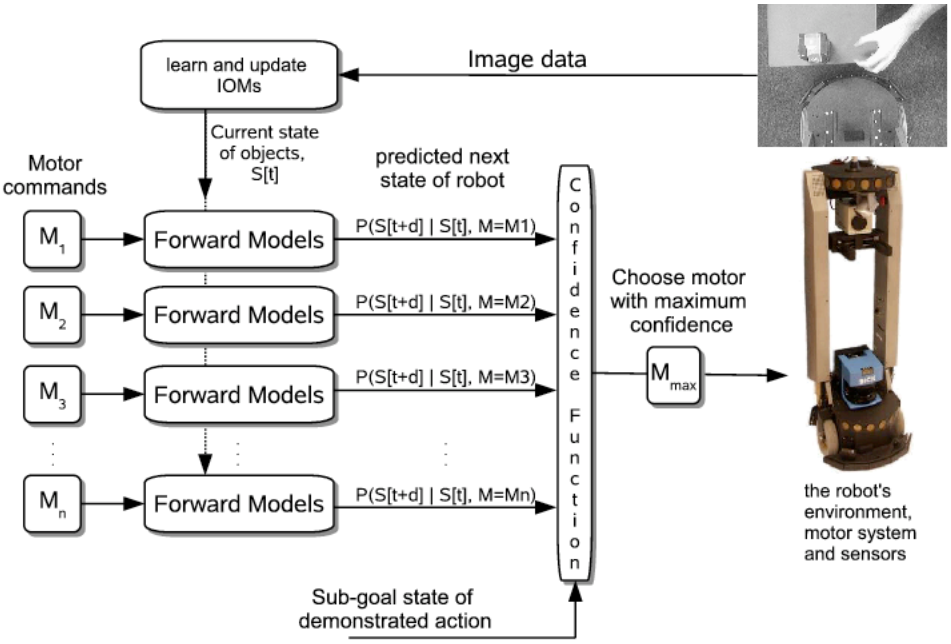

learn and update IOMs
Image data
Current state of objects, S[t]
Motor commands
predicted next state of robot
M1
M2
M3
Mn
Forward Models
P(S[t+d] | S[t], M=M1)
P(S[t+d] | S[t], M=M2)
P(S[t+d] | S[t], M=M3)
P(S[t+d] | S[t], M=Mn)
Confidence Function
Choose motor with maximum confidence
Mmax
the robot's environment, motor system and sensors
Sub-goal state of demonstrated action

Figure 9. The imitation architecture, using internal models learnt through exploration.

Despite being of limited use on their own, asocially learnt internal models provide the building blocks of the imitation architecture, shown in figure 9 . A generative approach to imitation is used: the internal models of the robot's motor system are used to understand the observed movements of a demonstrator by generating motor commands that will produce the closest match to this movement. The most important part of the system is the forward models, which predict how a motor command will change the state of objects.

These forward models are created from the learnt ICMs, and enable the robot to simulate numerous different motor commands. In the current set of experiments, the total number of commands is sufficiently small that each possible motor command can be simulated. In general, with limited computational resources and more degrees of freedom, this will not be the case. Future work will use the ICMs as inverse models to provide a smaller subset of relevant motor commands to simulate.

Internal models are learnt relative to the robots own visual system, so it has no way of directly understanding the actions it perceives others taking. Indeed, the robot's own motor system may not be capable of imitating the complex gestures and actions of a human motor system because of the different morphology. To overcome this 'correspondence problem', the observed action is represented, not using the states of the objects, but by the *difference* between the states of IOMs. This enables the interaction between the demonstrator and the shared object to be modelled in the same coordinate system as the interaction between the robot and the shared object.

The information about object interactions is a continuous stream of data. To perform the imitation at a more abstract level the sequence is split into sub-goals using peaks in the spatio-temporal curvature of the interaction distance between objects, as shown in figure 10. This technique is used in [21] to perform gesture recognition of humans by splitting the action into a sequence of movements. It is used here to find a sequence of interactions between objects; each element in the sequence is a sub-goal for the robot to imitate. By breaking a continuous stream of interaction data up into a set of key points in the interaction, the represented action and imitation is now independent of the specific timings involved in the movement - for most actions, it is the sequence of states in the movements that are important, not the time between the movements. Splitting a demonstration into a sequence also means it can easily be recognised if demonstrated again. Figure 11 shows screen-shots of the first three sub-goals extracted from an object interaction.

The confidence function's role is to assign a value to each possible motor command according to how close the robot estimates it will move it to the current sub-goal state. The confidence of each motor command, *m*, is calculated as:

$$confidence(m) = exp\left(-\left(abs\left(\widehat{S}_{self,m} - \widehat{S}_{shared,m}\right) - G_n\right)^2\right)$$

where G_n is desired interaction distance of the current sub-goal, and $abs\left(\widehat{S}_{self,m} - \widehat{S}_{shared,m}\right)$ is the predicted distance between the *self* IOM state and the *shared* IOM state. Confidences are higher for motor commands that make the robot's predicted motor system interaction with an object closest to the desired interaction. The confidences displayed in the graphs are normalised to sum to 1 at each time step for easy visualisation. To imitate a demonstrated sequence, the robot uses the motor command with the highest confidence.

The imitation process can be carried out entirely in simulation and visualised to the demonstrator. Figure 12 shows the simulated consequences of the robot imitating the first two sub goals of a demonstrated sequence. The simulation enables the intentions of the robot to be communicated to the demonstrator before executing them. The demonstrator can use this information to stop the robot performing an incorrect imitation, and potentially find out what is incorrect in the robot's knowledge. Future work will involve looking at how the demonstrator can become a more active element in the robot's development by adapting his actions according to visualisations of the

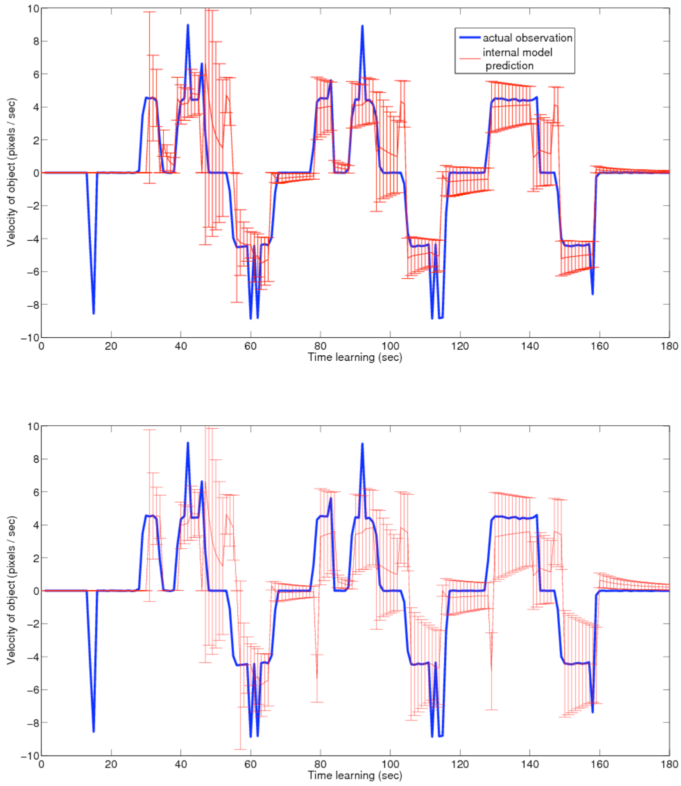


Figure 7. The predictions of two internal model structures for estimating the effect of the velocity motor command as the robot 'babbles' forwards and backwards. The top one can be seen to be the most accurate because it has the lowest estimated prediction variance, shown with the error bars. The structure of this model is shown in figure 8.

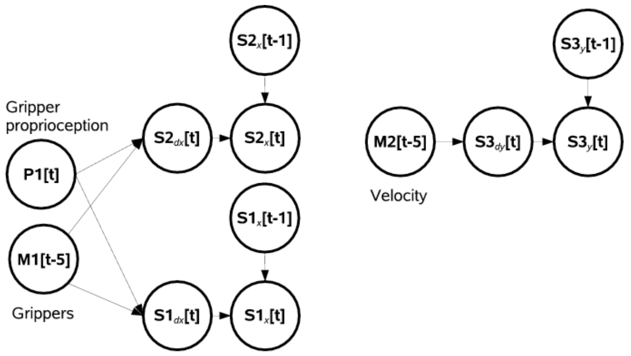


Figure 8. The most useful Bayesian network structures learnt for the gripper motor control (left) and the wheel velocity motor command (right). Both show that the motor commands affect the position of objects in the scene by changing their velocity. It has also learnt that the grippers' touch sensor can be used to predict how the grippers move.

robot's current knowledge. Figure 13 shows the confidence for multiple motor commands in simulation for the first two sub-goals: the robot moves forward, opens its gripper to touch the object, and then closes its gripper to move away.

The same architecture is used to make the real robot imitate an interaction with an object. Unlike the simulation, the state of the robot and the objects are not updated using the simulation, but with feedback from its vision system. Figure 14 shows the confidence of each motor command as the robot imitates the demonstrated interaction. Figure 15 shows screen-shots from an imitation.

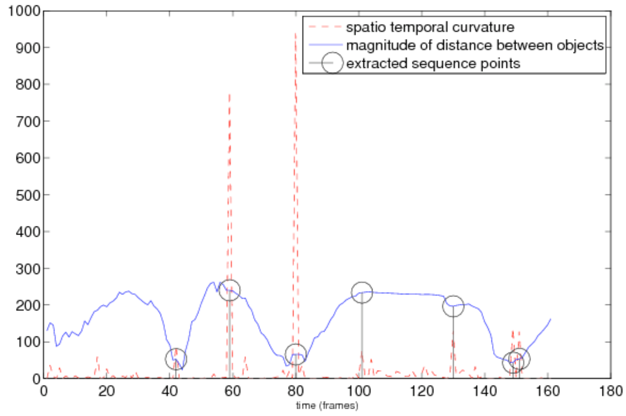


Figure 10. Extracting key points to imitate from an interaction sequence, shown in black circles. These points are extracted from peaks in the spatio-temporal curvature of the distance between the robot's motor system and the object it wishes to interact with.

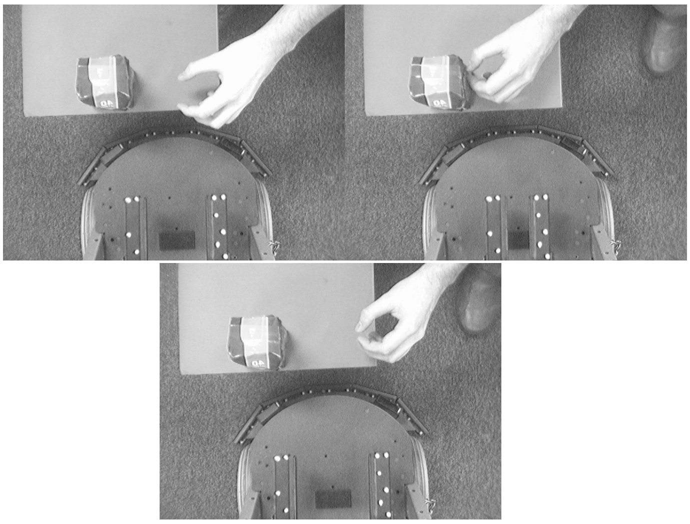


Figure 11. The first three sub-goals being imitated, extracted using the spatio-temporal curvature. Even though this action is occurring as the robot learns to control its gripper system, it is able to recognise it as an interesting action to imitate because neither the human hand nor the pack of biscuits can be accurately explained by its internal models.

In both the simulation and on the robot the observed interaction is successfully imitated. There are some interesting differences between the real system and the system simulated with the internal models. The real robot finishes the interaction in less time than the simulation. This is due to drift in the simulation, as errors in the internal models accumulate over time. When the gripper is fully open on the real robot, the *open gripper* command receives a lower confidence. As figure 8 shows, the ICMs had learnt during babbling that the gripper proprioception sensor data affected how the grippers move - when the gripper is fully open, the *open gripper* motor

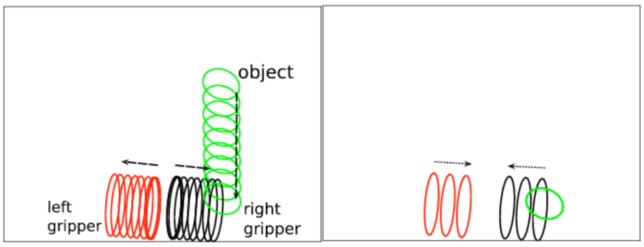


Figure 12. The simulated visualisation of the IOMs as the robot tries to touch the biscuits (left) and then moves away (right). The ellipses represent the means and covariances of the predicted objects' position, and the arrows show the direction of movement. Note that all aspects of this simulation, the appearance of the objects and their control with the motor system, are learnt from exploration. This is why the biscuits do not collide with the gripper; the robot has not learnt that objects can move when touched by other objects.

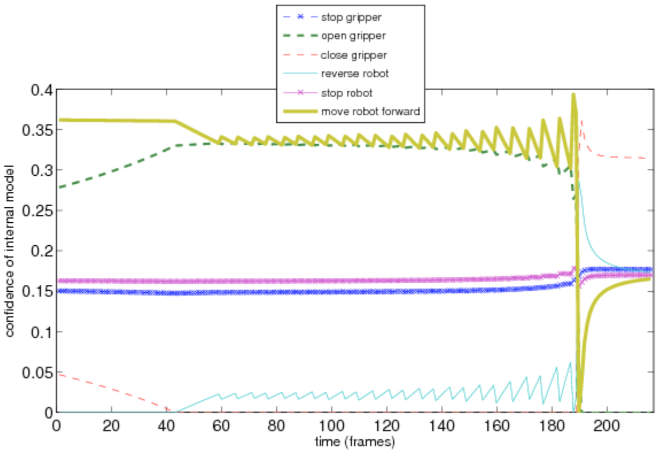


Figure 13. The progress of confidences of each learnt internal model in simulation as the robot tries to touch the object of interest. After this, the first goal state has been reached so the robot moves its grippers away to approach the next sub-goal.

command will not have any effect and will therefore not be useful in achieving the goal of moving the gripper closer to the object. This information is not available, however, in the simulation as the internal models do not currently learn *when* the proprioception information changes, just *how* to use it. The confidence values of the *open gripper* and *move forward* motor commands in the simulated imitation oscillate. This is because the simulation, unlike the robot, does not currently allow multiple motor commands to be issued simultaneously, so the two most appropriate motor commands end up being executed alternately.

5 Discussion

The purpose of both exploration and imitation presented in the experiments here is to enable a robot's knowledge and motor control ability to develop. So far, the process we have described is one-directional: the robot learns basic internal models and uses these to copy interactions on objects that both it and a human demonstrator can control. We are currently looking into the next stages of this teacher-imitator relationship, whereby imitation is not the final goal of the robot, but another process in its developmental repertoire that

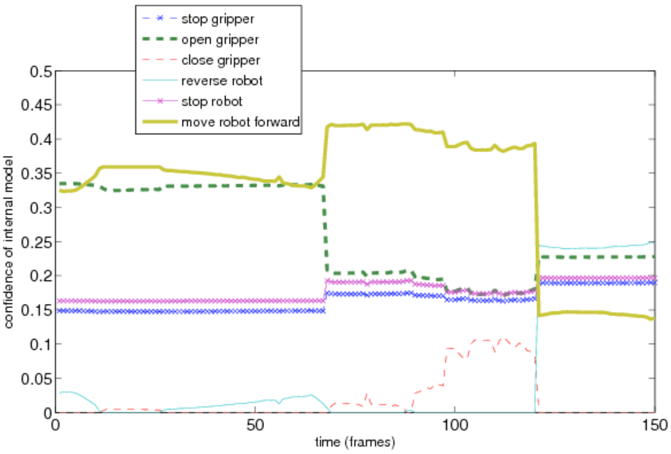


Figure 14. The progress of confidences of each motor command on the actual robot as the robot tries to imitate an interaction with the object.

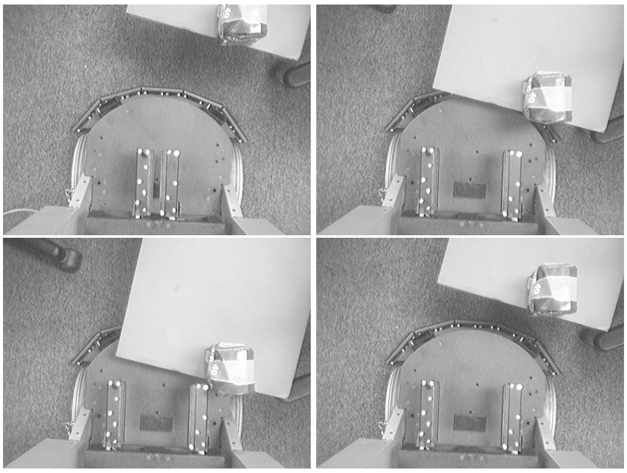


Figure 15. Frames 0, 50, 120 and 150 from the same imitation experiment as as figure 14.

is used to help it to learn.

Further results and experiments are currently being performed for more degrees of freedom in the robot's motor system, such as using its pan-tilt unit. With no *a priori* knowledge, the information available about the interactions is limited by the properties of objects the vision system can represent. Currently this is just the position and size of objects. This is why the only interaction the robot is currently capable of is object 'nudging'. Future work will involve investigating how the robot can attempt different interactions with the same objects so as to learn more detailed ways of interacting. This involves modelling more complex representations of objects and their interactions.

ACKNOWLEDGEMENTS

This work has been supported through a Doctoral Training Award from the UK's Engineering and Physical Sciences Research Council (EPSRC), and through a bursary from BBC Research. The authors would like to thank the members of the BioART team at Imperial College and Dilly Osbahr for their comments.

References

[1] A. Dearden and Y. Demiris. Learning forward models for robotics. In *Proceedings of IJCAI 2005*, pages 1440–1445, 2005.

[2] Anthony Dearden and Yiannis Demiris. Active learning of probabilistic forward models in visuo-motor developmen. In *Proceedings of the AISB*, pages 176–183, 2006.

[3] Anthony Dearden and Yiannis Demiris. Tracking football player movement from a single moving camera using particle filters. In *Proceedings of the 3rd European Conference on Visual Media Production (CVMP-2006)*, pages 29–37, 2006.

[4] Y. Demiris. Imitation, mirror neurons, and the learning of movement sequences. In *Proceedings of the International Conference on Neural Information Processing (ICONIP-2002)*, pages 111–115. IEEE Press, 2002.

[5] Y. Demiris and B. Khadhouri. Hierarchical attentive multiple models for execution and recognition (hammer). *Robotics and Autonomous Systems*, 54:361–369, 2006.

[6] Paul Fitzpatrick and Giorgio Metta. Grounding vision through experimental manipulation. *Philosophical Transactions of the Royal Society: Mathematical, Physical, and Engineering Sciences*, pages 2165–2185, 2005.

[7] A. Gopnik and A. N. Meltzoff. *Words, Thoughts, Theories*. MIT Press, 1st edition, 1998.

[8] R. M. Gordon. Simulation without introspection or inference from me to you. In M. Davies and T. Stone, editors, *Mental Simulation*, pages 53–67. Oxford: Blackwell, 1995.

[9] M. Hasenjager and H. Ritter. Active learning in neural networks. *New learning paradigms in soft computing*, pages 137–169, 2002.

[10] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

[11] M Jordan and D Rumelhart. Forward models: Supervised learning with a distal teacher. In *Cognitive Science*, volume 16, pages 307–354, 1992.

[12] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674–679, 1981.

[13] A. N. Meltzoff and M. K. Moore. Explaining facial imitation: A thoretetical model. *Early Development and Parenting*, (6):179–192, 6 1997.

[14] G. Metta and P. Fitzpatrick. Better vision through manipulation. In *Proceedings of 2nd International Workshop on Epigenetic Robotics*, pages 97–104, 2002.

[15] A. W. Moore, C.G. Atkenson, and S. A. Schaal. Memory-based learning for control. Technical report, 1995.

[16] Richard E. Neapolitan. *Bayesian Structure Learning*. Prentice Hall, 2004.

[17] S. Nichols and S. P. Stich. *Mindreading*. Oxford University Press, 2003.

[18] P. Oudeyer, F. Kaplan, V. Hafner, and A. Whyte. The playground experiment: Task-independent development of a curious robot. In *proceedings of the AAAI Spring Symposium Workshop on Developmental Robotics*, 2005.

[19] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[20] Jean Piaget. *The Child and Reality*. Viking Press Ltd., 111 edition, 1974.

[21] C. Rao and M. Shah. View-invariant representation ans learning of human action. In *Conference on Computer Vision and Pattern Recognition (CVPR'01)*. IEEE, 2001.

[22] S Schaal, A Ijspeert, and A Billard. Computational approaches to motor learning by imitation. *Phil. Trans. of the Royal Society of London B*, (358):537–547, 2003.

[23] Jessica A. Sommerville, Amanda L. Woodward, and Amy Needham. Action experience alters 3-month-old infants' perception of others' actions. *Cognition*, 2005.

[24] Robert W. White. Motivation reconsidered: The concept of competence. *Psychological Review*, 66(5), 1959.

Learning models of camera control for imitation in football matches

Anthony Dearden and **Yiannis Demiris**[1] and **Oliver Grau**[2]

Abstract. In this paper, we present ongoing work towards a system capable of learning from and imitating the movement of a trained cameraman and his director covering a football match. Useful features such as the pitch and the movement of players in the scene are detected using various computer vision techniques. In simulation, a robotic camera trains its own internal model for how it can affect these features. The movement of a real cameraman in an actual football game can be imitated by using this internal model.

1 Introduction

Imitation is a useful way to indirectly transfer knowledge from one agent to another by simply demonstrating the action to the imitator. In this paper, we investigate a particular scenario where this transfer of knowledge can be used to teach robotic cameras how to move in a football match. This scenario has useful applications in both simulation and real-world scenarios. In football computer games such as Pro Evolution Soccer, the movement of the camera during play and automated highlights is generated using pre-programmed control. The movement would be much more natural if it was imitating the movement of actual cameras during a football match. This would also save the programmer the effort of having to create the control algorithms. In actual football matches, up to 20 cameras can be used to provide coverage for a match, each requiring a human operator. Using robotic cameras, automated human-like camera control would give the broadcaster the ability to cover more matches or use more cameras viewpoints. Imitating not just camera movement, but also how the camera shots are selected by a director would enable the entire coverage process to be automated.

It would be advantageous if a robotic camera could be rapidly placed in a viewpoint, learn the effects that it has on a current location and then move accordingly based on the state of the game. To test the feasibility of this approach we implement a learning system on a football simulator, which learns to imitate the camera movements of a trained cameraman, by inverting the learnt effects that its own actions have on the visual field. We test the system on real data, provided by BBC Research, to demonstrate successful learning of the first step of the final system. Current work focuses on understanding the actions of groups of players, so the robotic camera can learn a model of the movement of human cameraman in terms of how the players are moving.

2 Extracting feature information from real and simulated football games

Before a robotic camera can understand the scene it is in, or is trying to imitate, it is first necessary to extract features from a scene which give the robot information about its own state and the state of other important features like the players. This section describes the image processing steps necessary to extract information about the movement of the camera and the position of players in the game. The same algorithms are applied to both the real and the simulated football match. The real data is taken from the feed from the central 'spotter' camera during the Everton vs Manchester City game in November 2005. The simulation was created using OpenGL to render a basic football game with players and line markings. The list of features which can be extracted from video sequences are shown in table 1.

Table 1. List of features that can be extracted from football video data and the information it provides.

Feature	Information provided
Pitch region information	Position of camera relative to pitch, change of camera shot by director
Skin region information	Close-up shots, crowd shots
Optical flow information	Approximate movement of camera
Player tracking	High-level state of game (e.g. who has possession)

2.1 Finding the pitch and player regions in a video

Figure 1 gives an overview of the computer vision process used to extract player regions. The basic idea behind the process is to subtract the distinctive green colour of the pitch to leave regions which are likely to be players. This idea has been used on numerous previous occasions e.g. [16].

The colour of the pitch is represented as a one- or two-dimensional histogram in HSV colour space. This histogram is back-projected onto each image and then, with a threshold applied, a binary pitch mask can be obtained. To estimate the entire pitch region, the pixels on the binary image are grouped into regions. By calculating the convex hull of the largest regions, the area in the image which the pitch covers can be calculated. Knowing the pitch regions in the image enables the tracking to be simplified by removing clutter from the crowd regions. The shape and position of the pitch region can also give information about the location on the pitch on which the camera is focused. As the colour of the pitch may drift over the duration of the match the histogram can be recursively updated by calculating the histogram of the pitch region excluding player regions.

[1] Department of Electrical and Electronic Engineering
BioART group, Imperial College London
E-mail: {anthony.dearden99, y.demiris}@imperial.ac.uk
[2] BBC Research,
Kingswood Warren, Tadworth, Surrey, KT20 6NP
oliver.grau@rd.bbc.co.uk

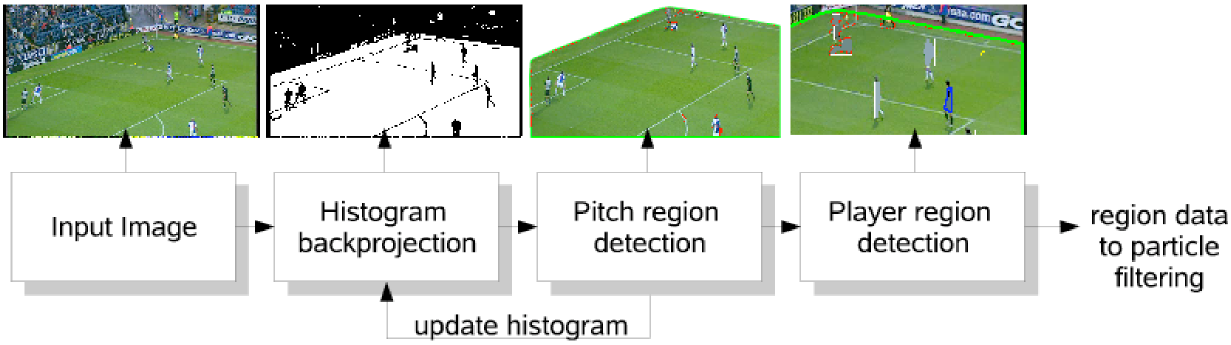


Figure 1. *Overview of the region extraction process*

Once the pitch region has been detected, player regions can be located by finding regions within the pitch region that do not correspond to the pitch colour. The regions can be filtered according to their area in the image, being separated into player regions and 'other' regions. These 'other' regions include noise and regions which are markings on the pitch.

The same technique for extracting the pitch regions can be used to detect regions of skin colour - this is a useful feature for detecting when a particular camera is doing a close-up shot on one of the players.

2.2 Tracking players

Many of the cameras being used to provide coverage for a football match have the sole purpose of tracking the action occurring in the game. Important information about the state of the game can be found from the position and movement of the players on each team; this is obviously an extremely useful feature for any robotic camera wishing to perform imitation.

Tracking footballers in video is made difficult by occlusions; other players or even the referee can obscure the information about a tracked player, as shown in figure 2; this is especially common during tackles, set-pieces and action in front of the goal. Overcoming the problem of occlusions can be done by fusing data from multiple camera sources, with the idea that the ambiguity will not be present from all angles [10, 9]. However, this adds to the complexity of the system; the goal here is to have a tracking system that can work directly from the image from a single moving camera tracking the action. Several techniques have been used previously to disambiguate player regions from a single camera source [7]. The first, also used here, is to apply morphological operators to erode close regions, hoping that they will split apart. Another method is to track the players using a graph representation, whereby the spatial relationship of players before a collision is stored so tracking can be continued when there is no longer an occlusion.

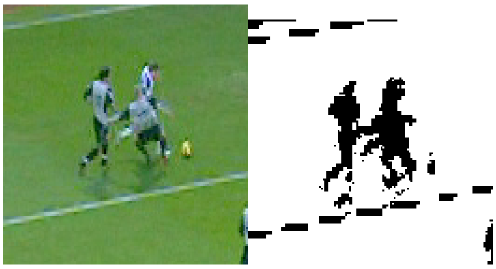


Figure 2. When players occlude each other, maintaining tracking can be difficult as the player region data (right) is ambiguous

To track players, here we use a particle filter. Particle filters have become extremely popular in recent years as a method for Bayesian tracking without the restrictive assumptions of linearity and Gaussian distributions of a Kalman filter [17, 1]. One aspect of particle filters which makes them especially useful in this situation is their ability to simultaneously maintain multiple hypotheses of the state of a tracked object. More details of the algorithm implementation and results can be found in [5].

Figure 3 shows sample frames from the sequences, together with the tracked positions of the two players. The particle filter is able to maintain tracking of both players, despite the occlusion occurring. As expected, when the occluding players separate again, the particles spread into multiple groups because of the increased uncertainty.

2.3 Estimating camera movement in a video

A useful source of information about the position of the camera in the scene comes from how it moves. To extract this information from a video sequence we use the KLT optical flow algorithm to track the movement of pixel-level features in the scene [12]. The pitch and player regions extracted above can be used to limit the points tracked to ones on the pitch; players will usually move independently of the camera. As the real camera moves across the scene, the low-level features leave the field of view, and new, untracked regions enter the scene. The algorithm continuously scans for new features to track and adds them to the list of points being tracked so that there is a continuous stream of tracked point features available. covering the entire image. The information from multiple points can also be combined by taking the average velocity of all points to give an overall metric of the camera's movement.

3 Imitating camera movement

Internal models are structures or processes that replicate the behaviour of an external process [11]. They have been hypothesised to

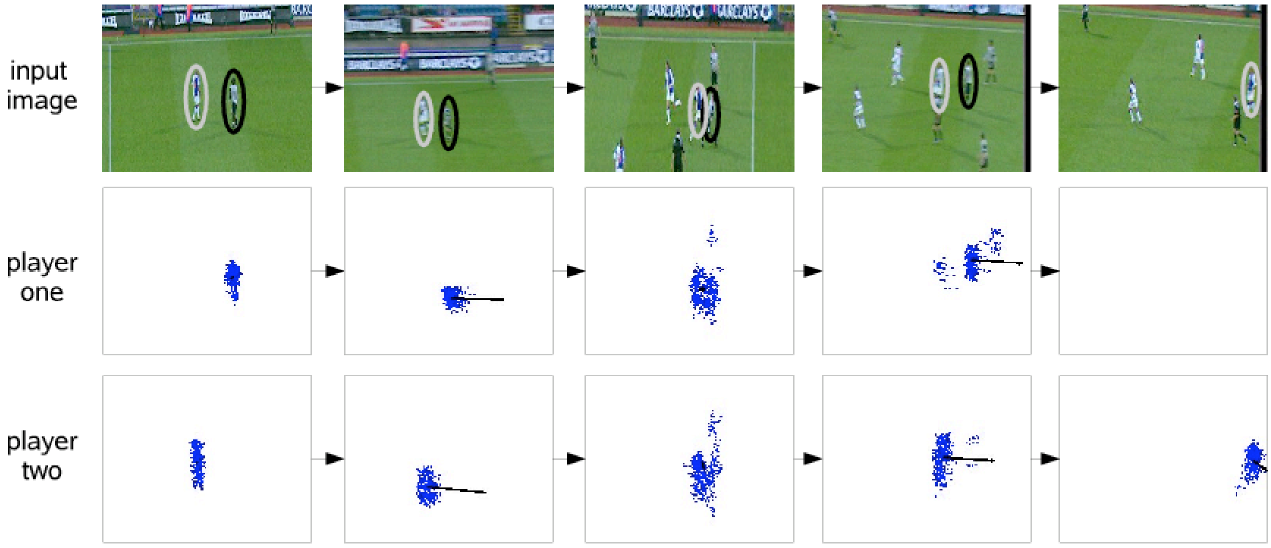


Figure 3. Frames from the tracking of two players. The last frame of player one is empty because the player has left the field of view. The black arrow in the particle represents the estimated velocity of the player. The players being tracked have been manually highlighted in the top images in black for player 1 and grey for player 2

exist in the the human central nervous system, for example to overcome the time delay of feedback from proprioception [18]. Giving a robot the ability to perform an internal simulation of external process enables it to simulate the effects of its actions internally before physically executing them. They enable a robot to predict the sensory consequences of its motor actions as *forward models,* or to estimate the motor commands that will lead to a desired state as *inverse models* [3]. They can be used for imitation by using a *simulation theory* approach [8, 13]. By using the internal models of its own motor system, a robot can understand and therefore imitate the actions it observes a demonstrator taking [6].

An inverse model could be programmed in using hard-coded software to track features on the pitch and thus estimate the position of the camera. The tracking problem in football is quite constrained, and unlike camera movement in other situations, there are reference points in the form of the pitch markings available that could be used. This approach is taken in [15]. A more generic solution would be to allow the robot to learn the internal model for itself through exploration. This would make the system applicable to other situations, and no effort is required by the programmer to come up with an algorithm for the inverse model.

In this work, the robot's actions are its pan, tilt and zoom commands; the camera is assumed to be stationary in the scene; a valid assumption for most cameras used in a football match. The sensory information it receives is provided by the computer vision features described in section 2, and listed in table 1. In this initial work we will just be focusing on using the optical flow information. The robotic camera needs to learn internal models which represent the effects its motor commands have on the optical flow data it receives back.

The internal models are represented either with radial basis functions or using the non parametric K-Nearest Neighbour (KNN) algorithm [2]. Radial basis functions had the benefit of being naturally smooth function approximators, whereas the KNN algorithm trains much faster[3], and allows the learnt forward model to be easily inverted and used as an inverse model to predict the motor command that can be used to recreate a particular movement. The KNN algorithm was implemented by storing the set of previous motor commands, the *pan* and tilt values, and the corresponding feature vectors, the optical flow velocity vector of image. To use a set as a forward model is a case of finding the K motor commands nearest to the one to be predicted for, each having a distance, *d* from the desired command. The corresponding K features for each of these commands can then be averaged to provide the predicted feature outputs. The average was weighted using a Gaussian kernel according to the distance, *d.* To use the KNN technique for an inverse model requires performing the process in reverse.

To train the internal model, the robot needs to execute multiple motor commands to produce a corresponding set of sensor data. In previous work [4], exploration of the motor-space with a camera was performed optimally so as to minimise the error in the internal model. As the only results currently available are on a simulated camera, the time taken for each camera to learn the internal model is less critical. Furthermore, only 2 degrees of motor freedom were involved. Therefore random motor commands were used to provide the training data.

The robotic camera uses the internal model it has learnt to imitate the movement of a trained camera man, and the optical flow features from the movement of the real camera man are given to the inverse model of the robotic camera. This will then output the motor commands the model expects will most likely recreate this movement in the robotic camera. The overview of this process is shown in figure 4. Selected screenshots for the simulated robotic camera imitating a real camera man are shown in figure 5. The left images are taken from the movement of the professional cameraman, and the right images show the simulated robotic camera's attempt to imitate the movement. Using only the optical flow features for imitation has the benefit of the robotic camera producing smooth, human like movement. Work is currently ongoing to make use of other features to ensure that absolute position information is used; as can be seen by the last frame, the imitating camera has drifted significantly from the camera it is imitating.

[3] training speed on a simulated camera is less of an issue than on an actual robot, where training time is limited

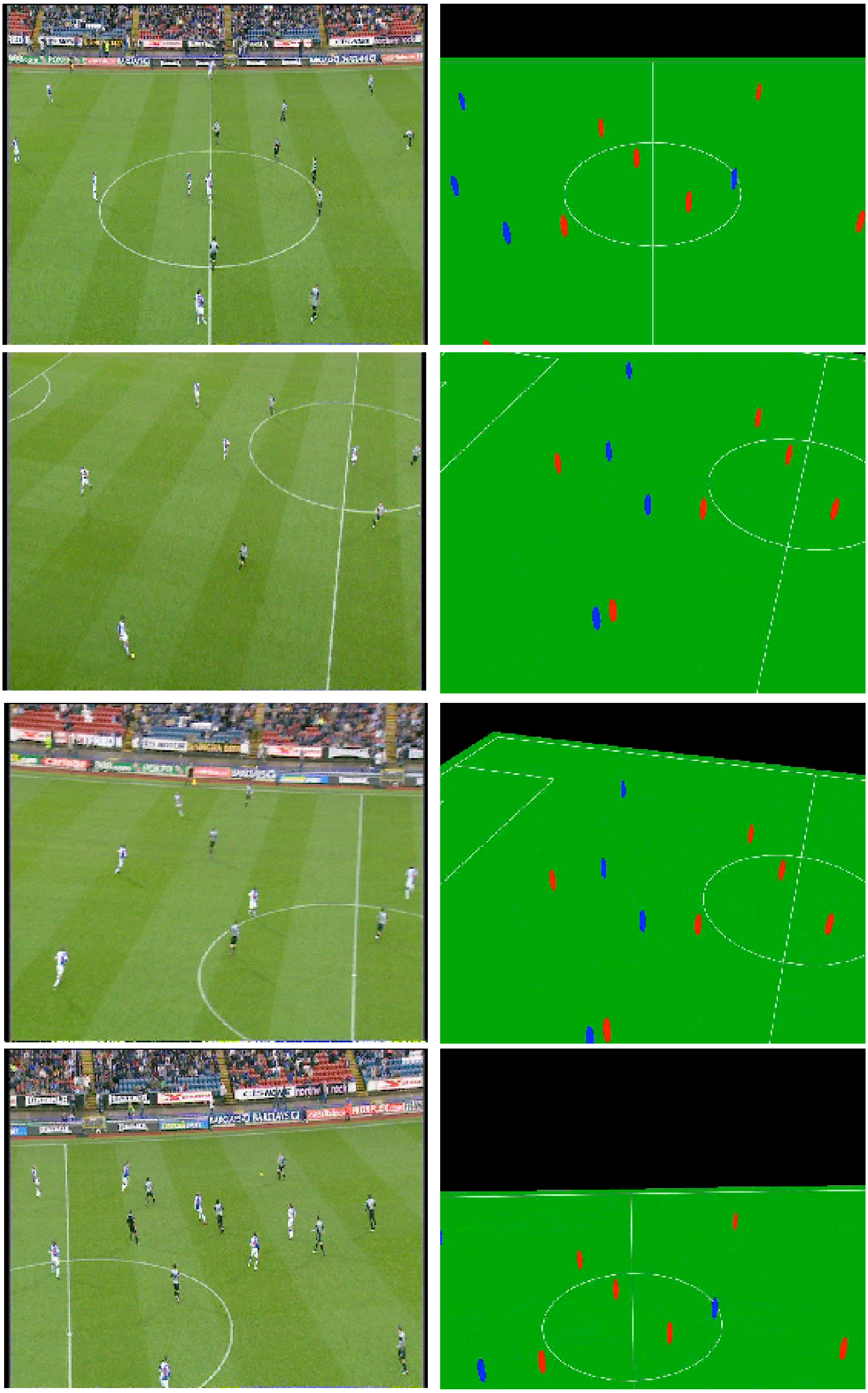


Figure 5. Frames 0, 100, 200 and 300 from the real football match and the imitating camera in simulation. The movement of the robotic camera is quite smooth and 'human-like'. However, as the movement is imitated using dynamic information, the absolute error in the robotic camera's position begins to accumulate.

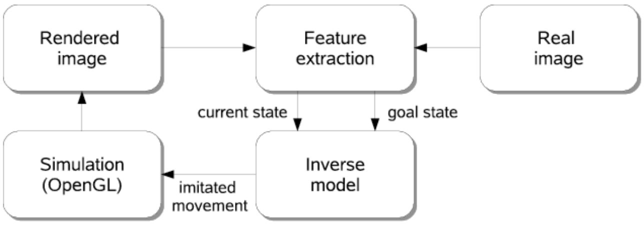


Figure 4. The imitation process using the learnt inverse model.

4 Discussion

The imitation the system performs so far is a mapping of one camera movement onto another. For imitation to be more general, the internal models need to be learnt at a level of abstraction at which they are applicable to any particular football match. It is intended that the robotic camera would be capable of tracking the action in a football match based on the actions taken by a professional cameraman with respect to the current state of the game. Much of the work on extracting information on the state of the game has been completed; the position of players on the pitch provides the most useful information for this. Work is currently ongoing to augment the structure of the inverse model so that camera movement is learnt as a function of player movement, I.E, given how players are currently moving, a robotic camera can move in the same way a human would move in the same situation.

Beyond the level of the movement of an individual camera, there is also the issue of how a human director switches between and sends requests to each camera. We are working to produce a system that can model and imitate this. The feature data that can currently be extracted provides useful information about the actions of the director. Figure 6, for example, shows how one of the features, the amount of pitch visible in the broadcast footage, varies over time. By detecting rapid changes in this value, it is easy to split the final footage into individual camera shots. A promising method for modelling these scene changes at a higher level is the use of dynamic Bayesian networks, such as hidden Markov models [14]. The switching between cameras given the state of the game can be modelled as sequence of hidden discrete states. The transition model for these states - i.e. how a human director switches between shots can be learnt using the low level features described in this work as the training data.

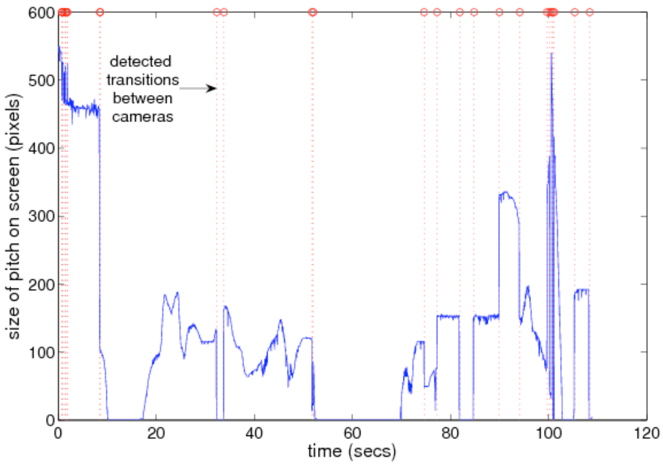


Figure 6. How the size of the pitch detected on screen varies over time. Rapid changes in this value can be used to detect when the director has switched between cameras

ACKNOWLEDGEMENTS

This work has been supported through a Doctoral Training Award from the UK's Engineering and Physical Sciences Research Council (EPSRC), and through a bursary from BBC Research. The authors would like to thank the members of the BioART team at Imperial College for their comments, and BBC Research and Match of the Day for access to the video footage.

References

[1] M. S. Arulampalam, S. Maskell, N. Gorden, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE transactions on signal processing*, 50(2), 2 2002.

[2] C. G Atkeson and S. Schaal. Memory-based neural networks for robot learning. *Neurocomputing*, (9):1–27.

[3] A. Dearden and Y. Demiris. Learning forward models for robotics. In *Proceedings of IJCAI 2005*, pages 1440–1445, 2005.

[4] Anthony Dearden and Yiannis Demiris. Active learning of probabilistic forward models in visuo-motor developmen. In *Proceedings of the AISB*, pages 176–183, 2006.

[5] Anthony Dearden and Yiannis Demiris. Tracking football player movement from a single moving camera using particle filters. In *Proceedings of the 3rd European Conference on Visual Media Production (CVMP-2006)*, pages 29–37, 2006.

[6] Y. Demiris. Imitation, mirror neurons, and the learning of movement sequences. In *Proceedings of the International Conference on Neural Information Processing (ICONIP-2002)*, pages 111–115. IEEE Press, 2002.

[7] P. Figueroa, N. Leite, R. M. L. Barros, I. Cohen, and G. Medioni. Tracking soccer players using the graph representation. In *Proceedings of the 17th international conference on pattern recognition (ICPR04)*. Los Alamitos, Calif.; IEEE Computer Society, 2004.

[8] R. M. Gordon. Simulation without introspection or inference from me to you. In M. Davies and T. Stone, editors, *Mental Simulation*, pages 53–67. Oxford: Blackwell, 1995.

[9] S. Iwase and H. Saito. Tracking soccer player using multiple views. In *IAPR workshop on machine vision applications*, 2002.

[10] S. Iwase and H. Saito. Tracking soccer players based on homography among multiple views. In *Visual Communications and Image Processing*, pages 283–293, 2003.

[11] M Jordan and D Rumelhart. Forward models: Supervised learning with a distal teacher. In *Cognitive Science*, volume 16, pages 307–354, 1992.

[12] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674–679, 1981.

[13] S. Nichols and S. P. Stich. *Mindreading*. Oxford University Press, 2003.

[14] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, 2 1989.

[15] G.A. Thomas. Real-time camera pose estimation for augmenting sports scens. In *Proceedings of the 3rd European Conference on Visual Media Production (C*, pages 10–19, 2006.

[16] O. Utsumi, K. Miura, I. Ide, S. Sakai, and H. Tanaka. An object detection method for describing soccer games from video. In *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*. IEEE, 2002.

[17] J. Vermaak, A. Doucet, and P. Perez. Maintaining multi-modality through mixture tracking. In *Ninth IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2003.

[18] D. M. Wolpert, R. C. Miall, and M. Kawato. Internal models in the cerebellum. *Trends in Cognitive Sciences*, 2(9):338–347, September 1998.

Imitating the Groove: Making Drum Machines more Human

Axel Tidemann[1] and **Yiannis Demiris** [2]

Abstract. Current music production software allows rapid programming of drum patterns, but programmed patterns often lack the *groove* that a human drummer will provide, both in terms of being rhythmically too rigid and having no variation for longer periods of time. We have implemented an artificial software drummer that learns drum patterns by extracting user specific variations played by a human drummer. The artificial drummer then builds up a library of patterns it can use in different musical contexts. The artificial drummer models the groove and the variations of the human drummer, enhancing the realism of the produced patterns.

1 Introduction

Our motivation for creating an artificial drummer was to combine the low-cost approach of programming drum parts through Digital Audio Workstations (DAWs, such as Pro Tools[3], Logic[4], Cubase[5], Digital Performer[6]) with the groove that a human drummer will provide. When producing music, recording the drums is a time-consuming and expensive process. The drums must be set up in a room with suitable acoustics and high quality microphones in order to produce good sounding drums. Subsequently, the drummer must play the actual part that is to be recorded. Most drummers do not play an entire song without any flaws, so the actual recording is also time-consuming. The current DAWs allow for cut-and-paste editing of the recorded audio, so a perfect take of a song is not required to produce a good result. This has drastically reduced the time required to record music in general, not only drums. But still the cost of recording drums is high, so for producers it is often more desirable to *program* the drums in the DAW. This approach is very low-cost, but it is often difficult to get a result similar to that of a real drummer. Programmed patterns have perfect timing and the velocity (i.e. how hard a note is played) of the beats is the same. A human drummer will always have small variations in both timing and velocity of each beat, which is often described as the *feel* or *groove* of the drummer. In addition, a human drummer will vary what he/she plays, such as adding an extra snare drum[7] beat or a fill when playing a certain pattern.

Programmed patterns can be altered to mimic these variations, but this requires the producer to manually change the velocity and timing of each beat, in addition to adding or removing beats to create variations. This can be very time-consuming, and requires musical knowledge of how to produce variations that will be perceived as those of a human drummer. Current DAWs have the ability to alter the beats by adding random noise, which might provide a more human-like feel to the drum tracks since the added noise will be perceived as human flaws. However, there is no guarantee that the result will sound more human-like, since the DAW itself has no understanding of what makes a drum pattern sound like it was played by a human. The research goal of this paper is to make an artificial drummer that is able to play patterns with feel and variation. This is realized by making the artificial drummer learn drum patterns from human drummers. The artificial drummer will *model* the variations that provide the feel of the drum pattern, which it can use to imitate the drumming style of the human drummer.

2 Background

The music software industry has created more complex samplers and synthesizers over the years as computers have become an important tool for musicians. To recreate the sound of a drumkit, a lot of effort has gone into recording huge libraries with gigabytes of samples (e.g. FXpansion BFD[8], Toontrack dfh[9], Reason Drum Kits[10], Native Instruments Battery[11]). The samples are then *layered* to simulate the dynamics experienced when playing real drums, i.e. that the pitch changes when playing soft or hard. Typically, when playing the snare drum in one of the aforementioned libraries, it will consist of a multitude of samples to achieve a more life-like response to playing dynamics.

These libraries are very sophisticated and sampled with meticulous precision, but they still need to be programmed. Even though these libraries come with software interfaces that are easy to program (most of them even come with rhythm pattern templates), there is still no substitution for a *real* drummer: the libraries themselves are merely tools for reproducing drum sounds, and the software interfaces have no intelligent way of generating human-like drum patterns. The templates will often be too rigorous and lifeless, something patterns programmed by the user also often suffer from (unless the user manually changes every note in the patterns generated, a very time-consuming process).

If the groove of a drummer could be *modeled*, a studio producer would have access to an artificial drummer that would be more lifelike than what is currently available. The artificial drummer would

[1] SOS Group, IDI, Norwegian University of Science and Technology. Email: tidemann@idi.ntnu.no

[2] BioART, ISN Group, Department of Electrical and Electronic Engineering, Imperial College London. Email: y.demiris@imperial.ac.uk

[3] http://www.digidesign.com

[4] http://www.apple.com/logicpro/

[5] http://www.steinberg.net/

[6] http://www.motu.com/

[7] A drumkit typically consists of a kick drum (which produces the low-frequency “thud” sound), a snare drum (a more high-pitched crackling sound) and a hihat (a high-frequent “tick” sound), see figure 3.

[8] http://www.fxpansion.com/index.php?page=30

[9] http://www.toontrack.com/superior.shtml

[10] http://www.propellerheads.se/products/refills/rdk/index.cfm?fuseaction=mainframe

[11] http://www.native-instruments.com/index.php?id=battery_us

be able to imitate a certain style of playing, specific to the drummer it has learned from. For producers, this would lower the cost of having life-like drums, and the producer could have the drummer of his choice to perform with the drummer's unique style. A professional drummer will have the opportunity to teach the artificial drummer his own unique style of playing, which he/she could later use in the studio or sell as a software plug-in.

We will now present a brief overview of research done in modeling the expressive performance of musicians. Saunders et al. [17] use string kernels to identify the playing style of pianists. The playing style is identified by looking at changes in beat-level tempo and beat-level loudness. However, imitating the style of the pianists was not attempted. Tobudic and Widmer also consider variations in tempo and dynamics as the two most important parameters of expressiveness. To learn the playing style of a pianist, they use first-order logic to describe how the pianist would play a certain classical piece, and then a clustering algorithm to group similar phrases together [19, 18]. They use the models to play back music in the style of given pianists, but some errors arise during playback. Tobudic and Widmer admit that these errors are due to the modeling approach (in fact, in [19] they claim it is "not feasible" to model the playing style of a pianist with the current data and training methods; the modeling approach was deemed too crude by the authors to be used as sufficiently accurate training data). Pachet's Continuator uses Markov models to create a system that allows real-time interactions with musicians [3, 5, 2], however his focus is more on replicating the tonal signature of a musician; the Markov model represents the probabilities that a certain note will follow another. A musician plays a phrase (i.e. a melody line), and the Continuator will then play another phrase which is a *continuation* of the phrase played by the musician (hence its name). Mantaras and Arcos use case-based-reasoning to generate expressive music performance by imitating certain expressive styles, such as joyful or sad [16, 15, 13, 12].

As far as the authors know, modeling the style of drummers is a novel approach to create an artificial drummer. The Haile drummer of Weinberg [23, 22] has some similarities, but there are some major points that separate it from our approach: first of all, it is a percussionist. Haile is a robot that plays a Native American Pow-wow drum, it uses only one arm and is far from being full-fledged drummer. In addition, it does not learn its patterns from human input, it has a database of rhythm patterns that are constructed by the designers of the system. Haile does imitate and modify patterns when interacting with human players, but it does not *learn* these patterns.

3 Architecture

We call our architecture "Software for Hierarchical Extraction and Imitation of drum patterns in a Learning Agent" (SHEILA). The following section will explain this architecture in more detail.

3.1 Input

Drum patterns are given as input to SHEILA. Ideally, the drum patterns would be extracted from audio files, however in this paper we have used MIDI[12] files as input to SHEILA. MIDI is a symbolic representation of musical information, and since it incorporates both timing and velocity information for each note played, it is very well suited for this application. SHEILA processes the MIDI file and learns the style of the human drummer.

Another advantage with representing the drum patterns using MIDI is that it is a tempo-less representation. Once SHEILA has learnt a pattern, it can be played back at a different tempo then when it was demonstrated, which gives the producer even greater flexibility.

3.2 Modeling

The system operates at two levels by modeling small and large scale variations, which will now be explained.

3.2.1 *Small-scale variations*

The *small-scale variations* arises as follows: when a drummer plays a specific pattern, he/she will play each beat of the pattern slightly different each time. The differences will occur in both timing and velocity. By calculating the mean and standard deviation of both the velocity and timing of each beat over similar patterns, the small-scale variations can be modeled using the Gaussian distribution. We investigated whether the Gaussian distribution was an appropriate model for the data by playing quarter-notes for about 8 minutes at 136 beats per minute (BPM), yielding 1109 samples. The histogram of the onset time and the velocity can be seen in figures 1 and 2 respectively, showing that the normal distribution is an appropriate model of the data.

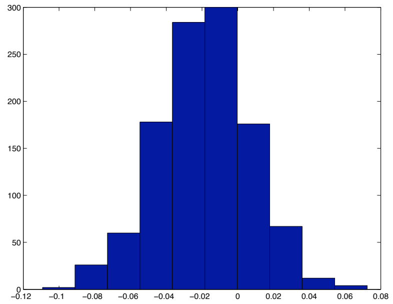


Figure 1. The histogram of the onset time after playing quarter notes for 8 minutes. The bars show distribution of the timing of the beats relative to the metronome.

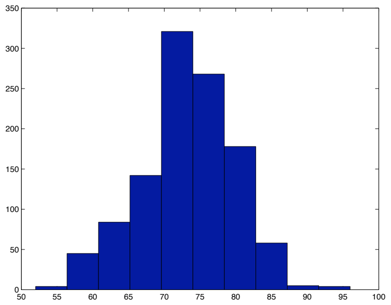


Figure 2. The histogram of the velocity after playing quarter notes for 8 minutes. Note that the histogram is not as nicely shaped as that of the onset time. This is most likely due to the velocity sensitivity in the pads that were used for gathering MIDI data, something that does not affect the onset time for each beat. The pads of the Roland SPD-S (see section 4 for description of the equipment) used in the experiment are rather small, and hitting towards the edge of the pad will affect the recorded velocity, even though the drummer might have hit the pad equally hard each time. Still, the histogram clearly shows the Gaussian bell-shaped curve for the samples gathered.

[12] Musical Instrument Digital Interface, a standard developed in the 1980s to enable communication between electronic music equipment.

3.2.2 Large-scale variations

Variations of the pattern itself, i.e. adding or removing beats are considered to be *large-scale variations*. Variations of a pattern is then stored along with the pattern it is a variation of, and based on a calculated probability, SHEILA will play back a variation of a certain pattern instead of the pattern itself. Exactly how this is done is elaborated on in the next section.

3.3 Training

To train SHEILA, the drum track of a song is given as input. In pop and rock music it is very common to divide a song into parts, such as a verse, chorus and a bridge. The song used in the experiments (see section 4) has the following structure: verse/chorus/verse/chorus/bridge, which is a common structure in pop and rock music. The point is that the drummer plays different patterns for the verse, chorus and bridge. We will now explain how SHEILA learns both large-scale variations of patterns and the small-scale variations of each pattern.

3.3.1 Learning large-scale variations

The occurrence of each of the patterns in the song is calculated (one pattern is then defined to be one measure, i.e. 4 quarter notes long). The patterns that are most frequently played are then considered to be *core patterns*. For instance, in a certain song the first core pattern C_1 occurs at measure 1. If the next core pattern C_2 appears at the 8th measure, the patterns that differ from C_1 between measure 1 and 8 are considered to be *large-scale variations* of C_1, named C_1V_x, where x is increasing with the number of variations of C_1. The ratio of variations of the core pattern (r_v) is calculated. This ratio will indicate how often a core pattern is to be varied when SHEILA will imitate the core pattern.

3.3.2 Learning small-scale variations

For each of the patterns (i.e. both core patterns and their variations), the mean (μ) and standard deviation (σ) of both the onset time and velocity is calculated, representing the *small-scale variations*. This is calculated the following way: the similar patterns are grouped together, and for each beat in the pattern, the mean and standard deviation for both velocity and onset time is calculated across the similar patterns. In order to calculate the mean and standard deviation of the onset time, a copy of all the patterns is quantized. Quantization means shifting each beat to the closest "correct" beat. If a beat was supposed to be on the "1", and it was slightly before or after, it is shifted to be exactly on the "1". The difference between the quantized pattern and the actual pattern is used to calculate the mean and standard deviation of the onset time for each beat. Each pattern (be it core or variation) will then have the normal distribution parameters assigned to each beat. An "ideal" (i.e. quantized and with no velocity information) version of this pattern is then stored in the SHEILA library, along with the mean and standard deviation of both onset time and velocity for each beat. A simplified outline of this procedure can be seen in algorithm 1. When imitating this pattern, the assigned parameters of the normal distribution will then be used to shift the beat forwards and backwards in time and to calculate the velocity. This will be explained further section 3.4.

3.3.3 Creating a library of the patterns

After processing the MIDI file, SHEILA will have built up a library of core patterns and their variations, see figure 3. SHEILA also stores which core patterns make up a song. This is simply an aid for the user of SHEILA; if the user knows the song the drum pattern was learned from, he will instantly know what kind of style the pattern was played in. In addition, SHEILA stores the name of the drummer playing this pattern. This is because it is very likely that different drummers will play the same pattern. SHEILA will model how each of them played the same pattern, and the name of the drummer can be presented to the user of SHEILA to further aid the user in indicating what kind of style the imitated drum patterns will be in.

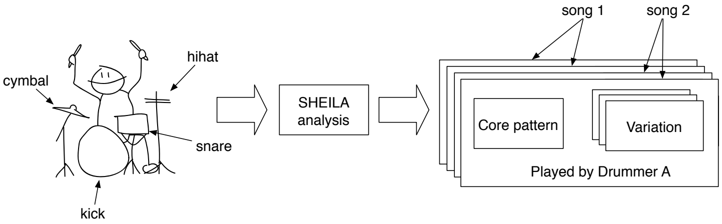


Figure 3. The learning process. Drum patterns are input to SHEILA, which analyzes the patterns and stores them in a library.

Algorithm 1 Training

```
count occurrence of each pattern in song
core patterns = most frequently played patterns
collect core patterns and their variations in groups
for all groups do
    calculate $\mu$, $\sigma$ of onset time and velocity for each beat across patterns (i.e. small-scale variations)
    store core pattern and variations (i.e. large-scale variations) along with $\mu$, $\sigma$ of each beat in SHEILA
end for
```

3.4 Imitation

This section describes how SHEILA can be used to imitate a given drum pattern in the style of a specific drummer.

3.4.1 Selection of playing style

If a producer wants SHEILA to play a certain pattern, he can write it down in a sequencer, export the pattern as a MIDI file and give it to SHEILA. If the pattern is recognized in the SHEILA library, it can then imitate the pattern in the style of the drummer that served as a teacher for the pattern. Indeed, if SHEILA recognized several drummers that played the same pattern, the producer will have the choice of selecting between the different drummers. The name of the song is also stored along with the drum patterns, allowing the producer to quickly have an idea of what the resulting pattern would sound like (presuming the producer knows the song). A good example is the pattern shown in figure 6. For many drummers, this is the first pattern learnt, and it is widely used in pop and rock music. If SHEILA had learnt the styles of all the major drummers in recorded music history, it would give the producer the choice of generating this pattern as played by Ringo Starr on "Help!" (the drummer of The Beatles, i.e. sloppy timing and simple variations) or Lars Ulrich on "Sad But True" (the drummer of Metallica, i.e. a rather "heavy" groove that is slightly behind the time, with typical heavy metal variations), among others. This is shown to the left in figure 4.

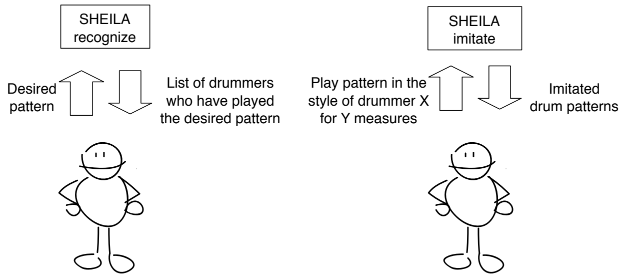

SHEILA recognize
Desired pattern
List of drummers who have played the desired pattern
SHEILA imitate
Play pattern in the style of drummer X for Y measures
Imitated drum patterns

Figure 4. Two steps that allows SHEILA to imitate a certain drummer. To the left the producer decides he wants SHEILA to play a specific pattern. He inputs this pattern in the MIDI format to SHEILA, which recognizes the pattern. Often several drummers will have played this pattern, and the output is a list of the drummers who can play this pattern and in which song it appeared. To the right shows the producer deciding which drummer should be imitated when generating the patterns, and he inputs this along with how many measures the pattern should be played for. SHEILA then imitates the style of the drummer specified, and outputs the imitated drum patterns back to the producer, ready to be used in the DAW of his choice.

3.4.2 Generation of patterns

Once the producer has decided which of the drummers in the SHEILA library he wants to use, he tells SHEILA to play the desired pattern in the style of drummer X for Y measures. At each measure, SHEILA decides whether to play the core pattern or one of the variations of the core pattern. The ratio of variations of a core pattern serves as the probability that a variation of the core pattern is played instead of the core pattern. The next step is to generate the actual beats that make up a pattern. When a pattern is to be generated, the onset time and velocity of each beat are calculated by generating random numbers from a Gaussian distribution, using the mean and standard deviation stored for each beat as parameters. This will yield slightly different patterns each time they are generated, but they will still sound similar, since the generation of patterns will come from a model of how the human drummer would play it. See algorithm 2 for a simplified description. The generated drum patterns are written to a MIDI file, which can later be imported into a DAW with high quality drum samples.

Algorithm 2 Imitation

```
present pattern p to be imitated to SHEILA
if p is known then
    make user select which drummer should be used for imitation
    of p, and for how many bars
    for the desired number of bars do
        if random number $< r_v$ then
            generate variation of p using the stored $\mu$, $\sigma$
        else
            generate p using the stored $\mu$, $\sigma$
        end if
    end for
end if
return generated patterns
```

3.5 Implementation

The SHEILA system was implemented in MatLab, using the MIDI Toolbox [10] to deal with MIDI file input/output. Propellerheads Reason 3.0 was used for recording MIDI signals and for generating sound from MIDI files, as explained in the following section.

4 Experimental setup

To acquire drum patterns, we used a Roland SPD-S which is a velocity sensitive drum pad that sends MIDI signals. Attached to the SPD-S was a Roland KD-8 kick drum trigger, along with a Pearl Eliminator kick drum pedal. A Roland FD-8 was used as a high hat controller. An Edirol UM-2EX MIDI-USB interface was used to connect the SPD-S to an Apple iMac, which ran Propellerheads Reason 3.0 as a sequencer, recording the MIDI signals. Reason was loaded with the Reason Drum Kits sample library to generate sound from the MIDI signals. The drummer would listen to his own playing using AKG K240 headphones connected to the iMac. The setup can be seen in figure 5.

Three drummers were told to play the same song, i.e. the same patterns for the verse, chorus and bridge, yielding three core patterns. If the verse is C_1, the chorus C_2 and the bridge C_3, then the structure of the song looks like this: verse (i.e. C_1) 8 measures, chorus (i.e. C_2) 8 measures, verse 8 measures, chorus 8 measures and finally the bridge (i.e. C_3) the last 8 measures. The drummer played along with a metronome to ensure that the tempo was kept constant. Each drummer would play in the tempo that felt most natural, so the tempo was varied around 100 beats per minute. After playing, the MIDI file was given as input to SHEILA. The pattern for the verse is shown in figure 7.

5 Results

This section is divided in three; the first two show how SHEILA models the drummers and how these models can be used to imitate the playing style of different drummers. The last section demonstrate listeners' ability to recognize which human drummer served as a teacher for the imitated patterns.

5.1 Modeling

Since all drummers played the same pattern, it is possible to see how SHEILA models each drummer differently. Figures 9-11 show the mean and standard deviation of the velocity for each beat when playing the pattern shown in figure 7 for drummers A, B and C respectively. Note that the scale along the Y axis is $[0 - 127]$, which is the range of the MIDI signal. The figures also show the mean and standard deviation of the onset time of each beat. The velocity bar is plotted on the mean onset time, which is why the velocity bars are not exactly on the beat. The standard deviation of the onset time is shown as the horizontal lines plotted at the base of each velocity bar (see figure 8 for a zoomed in plot with descriptive arrows that will help understand the plots). This is most clearly visible for drummer A (figure 9). Figures 12-14 more clearly show the mean and standard deviation of the onset time. The differences from 0 is how much the drummer is ahead or lagging behind the metronome. Between each quarter note beat (i.e. 1, 2, 3, 4) there are 100 ticks, divided in the range $[0 - 0.99]$. Since the data gathered is in the MIDI format, a tick is not a unit of time until the tempo has been decided. We present the results in ticks instead of another unit such as milliseconds, since

Figure 5. Playing drum patterns on the Roland SPD-S.

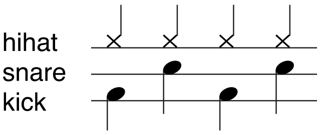


Figure 6. A simple and common drum pattern in pop and rock music.

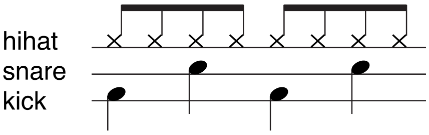


Figure 7. One of the patterns played by all the drummers in the experiments.

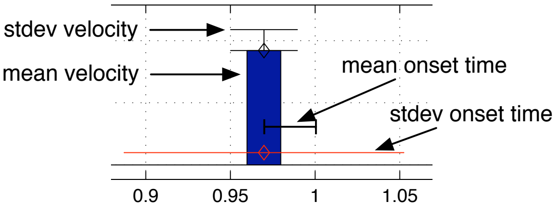


Figure 8. A zoomed in version of the third plot in figure 9. The arrows show how the mean and standard deviation of both the velocity and onset time is plotted. Note that the bar showing the mean onset time is *not* plotted on the figures, this is shown simply as the displacement from the nearest 8th note value (1 in this figure). These displacements are most easily seen in figure 9, for drummer B and C the displacements are smaller and are more easily observable on the onset time plots.

the ticks will accurately show the relative difference between each drummer, regardless of tempo. Drummer B has a mean onset time of -0.034 for the first kick drum beat (figure 13). This may not seem like a big difference, but these small variations are easy to pick up on when listening to a drum pattern. In fact, they are a crucial element to the groove of the pattern. MP3 files are available[13] that better illustrate these differences.

The figures clearly show how each drummer has his unique style. This is most easily seen on the hihat beats, as the accentuation is very different from drummer to drummer. Drummer B has a classic rock style of playing the pattern, with heavy accentuation on the quarter note beats (1, 2, 3, 4) and lighter notes on the off-beats (i.e. the *and* between the quarter notes), see figure 10. Figure 13 shows that he is constantly slightly ahead of time, which adds more aggressiveness to the playing style, and is also very common in rock music. Drummer A (figure 9) has a more even feel and is the drummer that varies most in timing (figure 12). This allows for a more relaxed feel, but will most likely sound rather sloppy when played at a high tempo.

Drummer C has the onset time mean closest to zero of all the drummers, see figure 14. Since he is both slightly ahead and behind the metronome it does not sound as tight as drummer B, which is constantly ahead of the beat. Instead, it has a more laidback feel that sounds more natural when played back at lower tempos.

It must be noted that describing the qualities of each of the drummers is inherently vague, but the graphs show that SHEILA successfully models the different styles of the drummers. Again we refer to the available MP3 samples.

5.2 Imitating

The models acquired for each drummer can now be used to *imitate* them. The imitation will be of both the small-scale variations (i.e. small changes in velocity and onset time in a pattern) and large-scale variations (varying the core pattern). To see how the large-scale variations are introduced, a simple experiment was done. After SHEILA had modeled each drummer playing the same song, SHEILA was used to imitate each drummer playing the same song all over again. Recall from section 4 that the structure of the song was playing verse/chorus/verse/chorus/bridge, each for 8 measures, and that the verse, chorus and bridge corresponded to C_1, C_2 and C_3 respectively. To imitate the same song, SHEILA was then told to play the same song structure (i.e. C_1 for 8 measures, C_2 for 8 measures, C_1 for 8 measures and so on). How the song was originally played along with the large-scale variations introduced when imitating the style for each drummer is shown in table 2.

Figures 15-17 show how the pattern in figure 7 was played back differently in terms of small-scale variations for each of the drummers. The figures show only one measure, over several measures these would be slightly different. They can be compared to figures 9-11, which show the mean and standard deviation of the velocity and onset time. Likewise, the onset time from the imitated pattern is shown in figures 18-20.

5.3 Evaluation by listeners

In order to examine how well SHEILA imitates the playing style of the three different drummers, we got 18 participants to compare the output of SHEILA to that of the original drummers. In order to make it harder to tell the drummers apart, the listeners heard 8 bars

[13] http://www.idi.ntnu.no/~tidemann/sheila/

of each drummer played at 120BPM, yielding 15 second samples of drumming. The same drumkit sample library was used to create identically sounding drumkits. The drummers originally recorded their drumming at different tempos (e.g. the tempo that felt most natural to them). Since the drumming was recorded in the MIDI format, it could be sped up without any distorted audio artifacts.

SHEILA then generated another 8 bars in the style of each drummer, played back at 120BPM. This included large-scale variations that were not present in the 15 second samples that the listeners would use to judge the imitation by. The evaluation was done as follows: the participants listened to the samples of the original drummers, and then the imitated patterns produced by SHEILA, which were presented in random order. The participants were free to listen to the samples in any order and as many times as they liked. The listeners completed the experiment by classifying each of the imitated drum patterns as being that of drummer A, B or C.

Table 1 shows that the listeners correctly classified which drummer served as a teacher for the imitated drum parts most of the time; the lowest classification rate being that of drummer C which was 83.3%.

Drummer	A	B	C
Classification	94.4%	88.9%	83.3%

Table 1. How often the imitated SHEILA output was correctly classified as being imitated from the corresponding human drummer.

6 Discussion and conclusion

We have implemented an artificial drummer that learns drum patterns from human drummers. In addition to simply learning the drum patterns themselves, the system *models* how a drummer would play a certain pattern, both in terms of small-scale variations in timing and velocity, and large-scale variations in terms of varying patterns. This has been demonstrated by letting three different drummers play the same song, and then showing how SHEILA models the different style of each drummer. Subsequently, we showed how SHEILA will play back the same song in a different way (in terms of large-scale variations), and also how the imitated pattern themselves are slightly different in terms of small-scale variations, but still in the *style* of the imitated drummer. By human evaluation, we have shown that the imitated drum patterns are often perceived as being similar to the originals. The work presented in this paper has demonstrated the core principle for using learning by imitation: namely to simply show the computer what you want it to do, and them make it imitate you.

Note that SHEILA need not be trained only on songs. For instance, to model how a certain drummer would play the pattern shown in figure 7, the drummer could play the pattern for a certain amount of measures, adding the large-scale variations the drummer would feel natural to play with this pattern. This would be a useful approach in terms of building up huge libraries of patterns and variations of these patterns, but this lacks the aspect of how the drummer played in order to fit the musical context. The advantage of training SHEILA based on patterns in a song is that the producer using SHEILA to generate drum patterns will instantly know which feel was on that track, and there would not be variations that will appear out of context.

The MIDI files used in this experiment was made by amateur drummers, since hiring professional drummers would be too expensive. The MIDI representation has the advantage of being tempo-less,

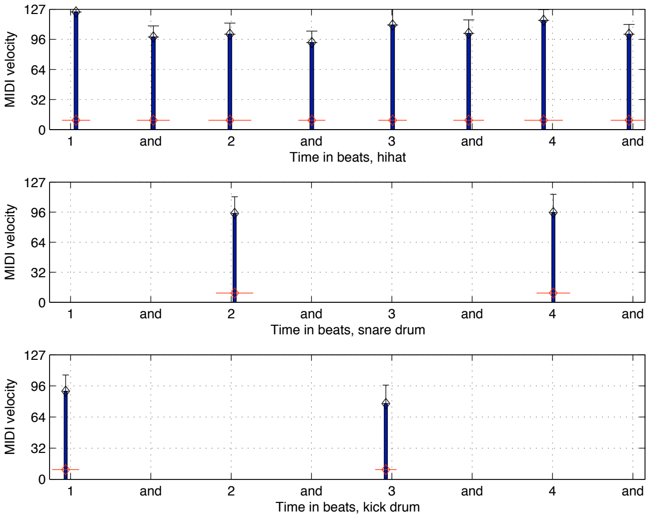


Figure 9. Velocity and onset time plot, drummer A. The hihat velocity is not varied to a great extent, but with more variance in the onset time gives the playing style a relaxed feel. Recall that the Y scale is $[0 - 127]$, which corresponds to the MIDI resolution. The X scale corresponds to the beats in the measure.

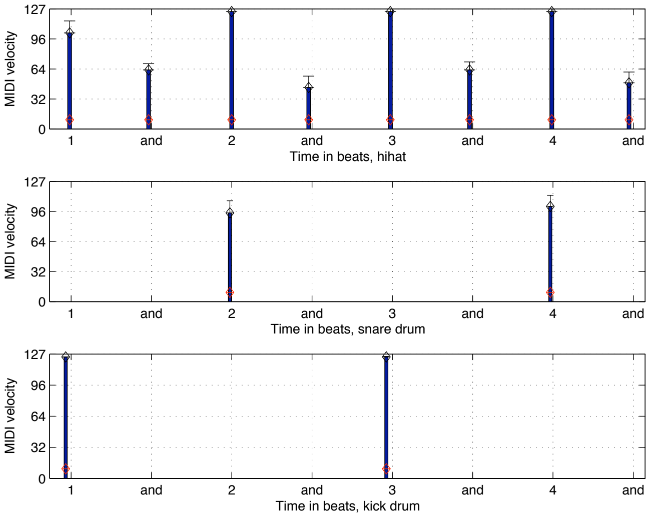


Figure 10. Velocity and onset time plot, drummer B. The hard accentuation on the downbeat is common for rock drummers.

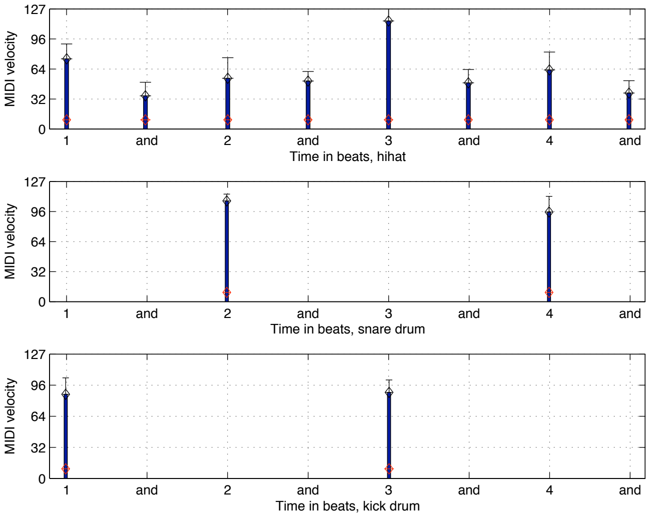


Figure 11. Velocity and onset time plot, drummer C. A more odd variation of velocity for the hihat, which creates a rather laidback feel.

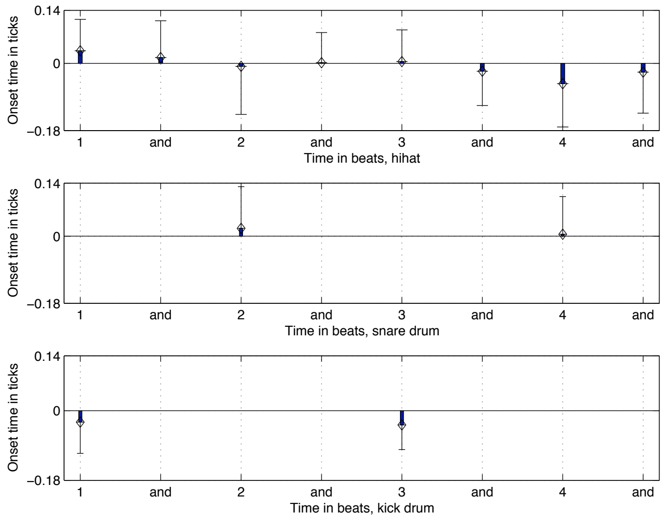


Figure 12. Onset time plot, drummer A. A rather big variance makes the groove feel less rigorous and more free and open, but this will most likely not sound very fluent when played back at high tempos. Recall that the onset time is measured in ticks between quarter notes, with range $[0 - 0.99]$.

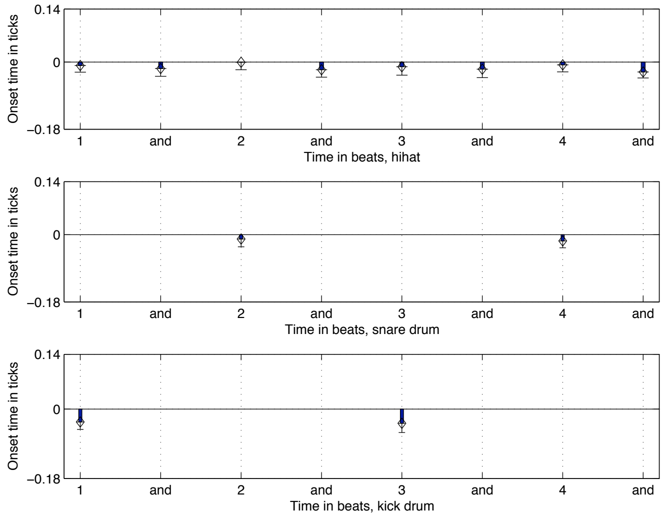


Figure 13. Onset time plot, drummer B. Constantly slightly ahead of the beat, which gives the groove a more aggressive feel.

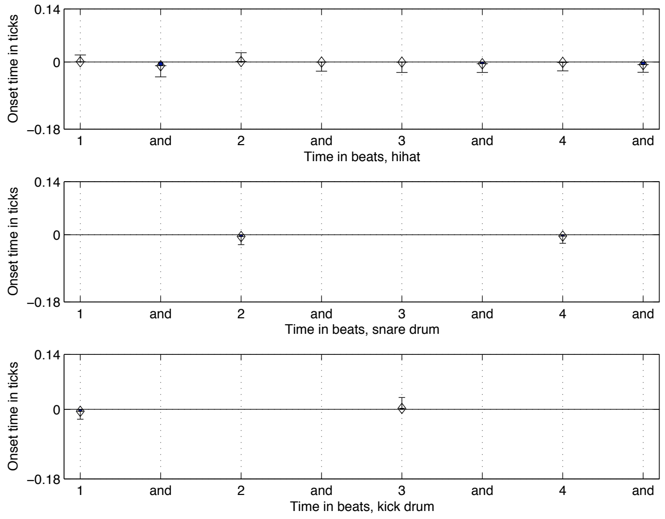


Figure 14. Onset time plot, drummer C. All the onset times are very close to the metronome, but the variations in being both before and after the beat makes this groove sound less tight than that of drummer B.

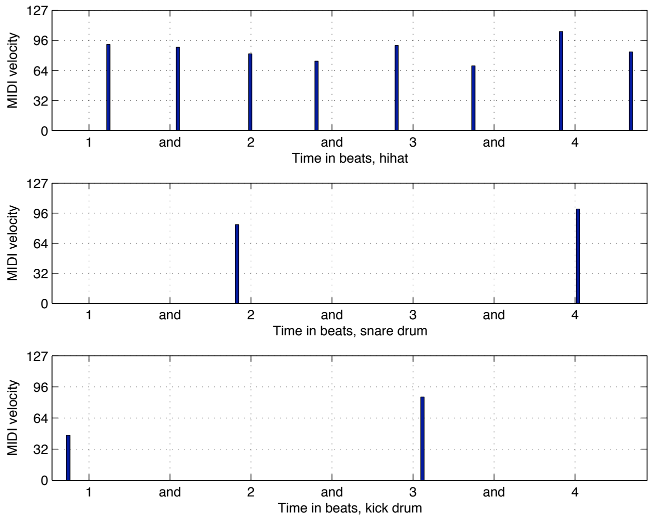


Figure 15. Imitated velocity and onset time plot, drummer A. Compare to figure 9 to see that the pattern deviates slightly from the mean and standard deviation.

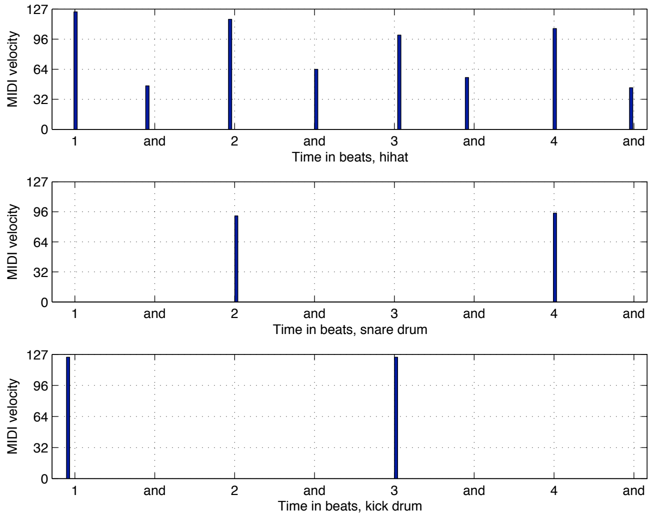


Figure 16. Imitated velocity plot, drummer B. The same “rock” feel is kept during the imitation (as can be seen in figure 10). Note how the hihat beat on the 3 is slightly behind the beat. This can be heard as a small flaw in the playing style, but will also add life to the resulting drum track.

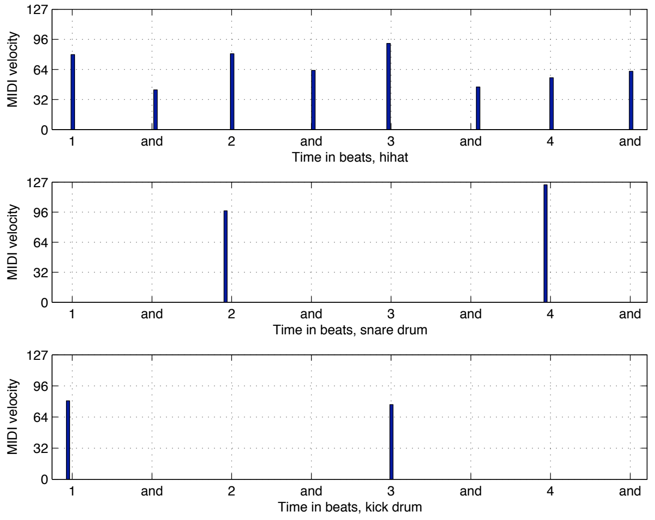


Figure 17. Imitated velocity plot, drummer C. The particular accentuated hihat beat on the 3 is present, albeit not so dominating (see figure 11 for reference). Timing is both ahead and behind the beat, as modeled.

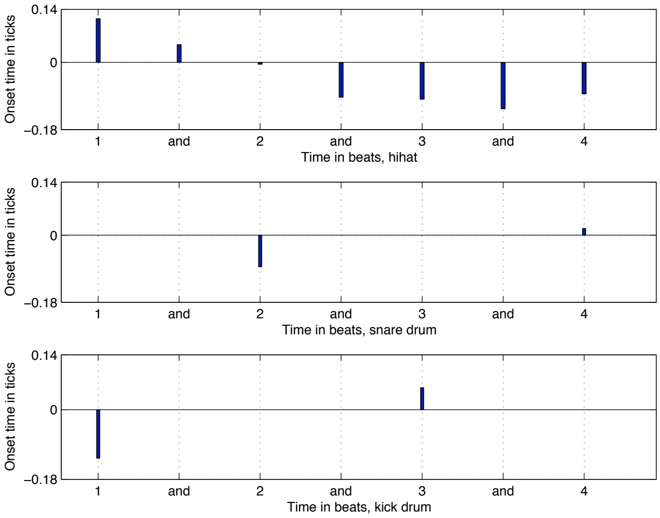


Figure 18. Imitated onset plot, drummer A. The plot complements figure 15, showing the timing with the different onset times which tend to be both ahead and behind the metronome beat.

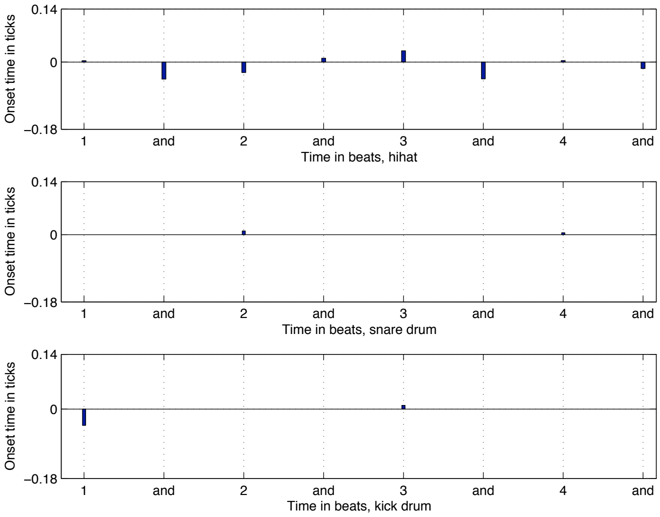


Figure 19. Imitated onset plot, drummer B. The beats are most of the time ahead of the metronome. The hihat beat on the 3 can more clearly be seen to be slightly behind the beat (as is also observable in figure 16).

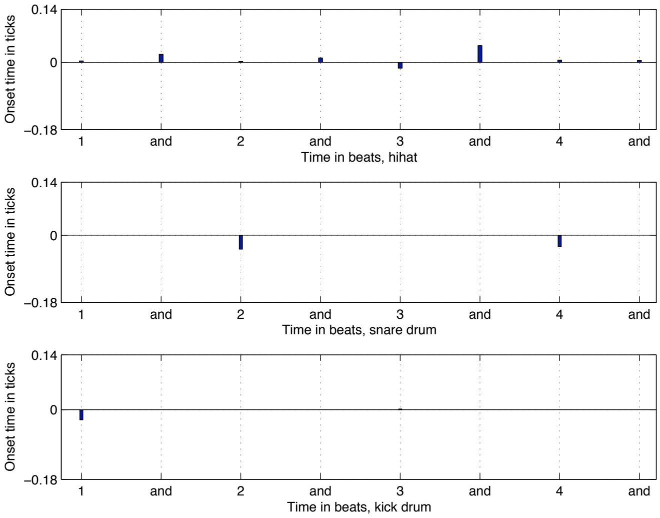


Figure 20. Imitated onset plot, drummer C. The mean was close to zero (as can be seen in figure 14); this plot clearly shows how the onset time of the beats varies both ahead and behind the beat over time.

but it can also yield drum patterns that would sound bad if played back at a tempo that is very different from when it was recorded. Another advantage of the MIDI representation is that it focuses solely on the playing style of the drummer. A drummer will often have a certain *sound* associated with him. This quality which is hard define formally is due to many factors; e.g. the brand of drums he/she is playing on, the producer, the genre of music, when it was recorded (i.e. drums recorded in the 80s sounds different from those in the 70s), to name a few. This further aids to develop the *signature* of the drummer, i.e. not just the patterns played but also the sonic qualities of the drumming. However, the results of this paper shows that human listeners are able to tell different drummers apart based only on the playing style of the drummer.

Table 2. How each drummer played the song in terms of core patterns and variations of core patterns. How each drummer originally played the song is shown to the left of each column dedicated to one drummer. How the imitated song differs from how it was originally played is shown in white text on a black background.

Drummer A		Drummer B		Drummer C	
Original	Imitated	Original	Imitated	Original	Imitated
C1	C1	C1	C1	C1	**C1V2**
C1	C1	C1	C1	C1	C1
C1	C1	C1	C1	C1	C1
C1	**C1V2**	C1V1	**C1V2**	C1	C1
C1	C1	C1	C1	C1	C1
C1	**C1V1**	C1	C1	C1V1	**C1**
C1V1	**C1V2**	C1	C1	C1	C1
C1	C1	C1	**C1V2**	C1	C1
C2	C2	C2	C2	C2	C2
C2	C2	C2V1	**C2**	C2V1	**C2**
C2	**C2V1**	C2	C2	C2	**C2V1**
C2	C2	C2	**C2V1**	C2	C2
C2V1	**C2**	C2	**C2V4**	C2V2	**C2V3**
C2	**C2V1**	C2V1	**C2**	C2V3	C2V3
C2	**C2V1**	C2	C2	C2V4	**C2**
C2	C2	C2V2	**C2**	C2V5	**C2V3**
C1	C1	C1	C1	C1	C1
C1	C1	C1	C1	C1V2	**C1**
C1V2	**C1**	C1	C1	C1V3	**C1**
C1	**C1V1**	C1V2	**C1**	C1	C1
C1	C1	C1	C1	C1	**C1V3**
C1	C1	C1	**C1V1**	C1	C1
C1	C1	C1	C1	C1V4	**C1**
C1V3	C1V3	C1V3	**C1**	C1	C1
C2	C2	C2	**C2V4**	C2	C2
C2	C2	C2	**C2V1**	C2	C2
C2V2	C2V2	C2	**C2V1**	C2	C2
C2	C2	C2	**C2V4**	C2	**C2V6**
C2V3	**C2**	C2V3	C2V3	C2V6	**C2**
C2	**C2V1**	C2	C2	C2	**C2V4**
C2V2	**C2**	C2	C2	C2V7	**C2V3**
C2V4	**C2**	C2V4	**C2V1**	C2V8	**C2**
C3	C3	C3	**C3V1**	C3	C3
C3	C3	C3	C3	C3	C3
C3	C3	C3	C3	C3	C3
C3	C3	C3	C3	C3	C3
C3	C3	C3V1	**C3**	C3	C3
C3	C3	C3	C3	C3	C3
C3V1	**C3**	C3	**C3V1**	C3	C3
C3	**C3V1**	C3	C3	C3V1	**C3**

7 Future work

One drawback of the system as it is currently implemented, is that it does not take musical context into account when modeling the different large-scale variations in a song. Very often, a drummer will make a large-scale variation in order to highlight dynamic parts in the song or in response to other instruments' melodies. This is often referred to as *breaks* or *fills*, and can be described as being big deviations from the core pattern, e.g. playing on the toms or doing a drum roll. Currently, breaks are modeled as mere variations of a core pattern, and can be played at any point during a song. A break will typically occur only at certain places, such as the measure leading up to the chorus or to highlight a specific section of the melody. These variations should be modeled on the basis of musical context, which would aid the modeling of the other patterns as well. The current implementation of SHEILA only looks at the pattern themselves, augmenting it with musical knowledge could allow for modeling *why* a drummer would play in a specific manner in response to the melody and the dynamics of a song, i.e. *understanding* how the drummer is being creative, as attempted by Widmer [24] and Pachet [4]. In addition, if the system could handle sound input instead of MIDI files, it would give easy access to vast amounts of training data. Such a system might be implemented according to Masataka and Satoru's approach to find melody lines in pop songs, also extracting the drum pattern [11] or using one of the systems described in [9].

In addition, we are interested in modeling the physical movements of the drummer as well. Drummers play differently, not just in terms of different patterns and styles, but also in the way they move their entire body when playing. By the use of motion tracking, we aim to be able to model the physical movements of the drummer playing, which would enable SHEILA to imitate the physical playing style of a specific drummer as well. This ability could be used in a more direct multi-modal interaction setting with other musicians, and opens up another interesting field of research, namely understanding how musicians interact when playing together [21]. Work in this direction would employ the concept of using multiple forward and inverse models [14] to control the robot as it learns to imitate, as done by Demiris [6, 7]. The idea of having a library of patterns was inspired from this multiple paired models approach, however the current implementation does not use forward or inverse models.

The ability to model the style of different drummers depends on the assumption that the drums were recorded using a metronome to keep the tempo constant. However, this is often an unnatural way of playing for drummers, as the tempo becomes too rigid and is not allowed to drift in tune with the dynamics of the song. Future implementations should enable SHEILA to imitate without the assumption that the drums were recorded with a metronome, such as the approach of Cemgil et al., who uses the Bayesian framework to quantize onset times without assuming the performance was recorded using a metronome [1]. Toivainen has implemented a system that allows tracking the tempo in real-time by using adaptive oscillators [20], Desain and Honing use a connectionist approach to real-time tracking of the tempo [8]. The latter approach would be necessary if the artificial drummer would be used in a live setting, as the tempo tends to drift more than when recording in a studio.

There are a lot of interesting directions for future research, and we believe that this paper is an important first step towards building an artificial drummer.

ACKNOWLEDGEMENTS

The authors would like to thank the BioART team (Anthony Dearden, Paschalis Veskos, Matthew Johnson, Tom Carlson, Balint Takacs and Simon Butler) and the reviewers for insightful comments and constructive criticism regarding the content of this paper.

REFERENCES

[1] Ali Taylan Cemgil, Peter Desain, and Bert Kappen, 'Rhythm quantization for transcription', *Computer Music Journal*, **24**(2), 60–76, (Summer 2000).

[2] François Pachet, 'Interacting with a musical learning system: The continuator', in *ICMAI '02: Proceedings of the Second International Conference on Music and Artificial Intelligence*, pp. 119–132, London, UK, (2002). Springer-Verlag.

[3] François Pachet, 'Playing with virtual musicians: The continuator in practice', *IEEE MultiMedia*, **9**(3), 77–82, (2002).

[4] François Pachet, *Creativity Studies and Musical Interaction*, Psychology Press, 2006.

[5] François Pachet, *Enhancing Individual Creativity with Interactive Musical Reflective Systems*, Psychology Press, 2006.

[6] Yiannis Demiris and Gillian Hayes, *Imitation in animals and artifacts*, chapter Imitation as a dual-route process featuring predictive and learning components: a biologically-plausible computational model, 327–361, MIT Press, Cambridge, 2002.

[7] Yiannis Demiris and Bassam Khadhouri, 'Hierarchical attentive multiple models for execution and recognition of actions', *Robotics and Autonomous Systems*, **54**, 361–369, (2006).

[8] Peter Desain and Henkjan Honing, 'The quantization of musical time: A connectionist approach', *Computer Muisc Journal*, **13**(2), (1989).

[9] Simon Dixon, 'Analysis of musical content in digital audio', *Computer Graphics and Multimedia: Applications, Problems, and Solutions*, 214–235, (2004).

[10] T. Eerola and P. Toiviainen, *MIDI Toolbox: MATLAB Tools for Music Research*, University of Jyvskyl, Kopijyv, Jyvskyl, Finland. Available at http://www.jyu.fi/musica/miditoolbox/, 2004.

[11] Masataka Goto and Satoru Hayamizu, 'A real-time music scene description system: Detecting melody and bass lines in audio signals', in *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, pp. 31–40, (August 1999).

[12] Josep Lluís Arcos and Ramon López de Mántaras. Combining AI techniques to perform expressive music by imitation, 2000.

[13] Josep Lluís Arcos and Ramon López de Mántaras, 'An interactive case-based reasoning approach for generating expressive music', *Applied Intelligence*, **14**(1), 115–129, (2001).

[14] Michael I. Jordan and David E. Rumelhart, 'Forward models: Supervised learning with a distal teacher', *Cognitive Science*, **16**, 307–354, (1992).

[15] Ramon López de Mántaras and Josep Lluís Arcos, 'The synthesis of expressive music: A challenging CBR application', in *ICCBR '01: Proceedings of the 4th International Conference on Case-Based Reasoning*, pp. 16–26, London, UK, (2001). Springer-Verlag.

[16] Ramon López de Mántaras and Josep Lluís Arcos, 'AI and music from composition to expressive performance', *AI Mag.*, **23**(3), 43–57, (2002).

[17] Craig Saunders, David R. Hardoon, John Shawe-Taylor, and Gerhard Widmer, 'Using string kernels to identify famous performers from their playing style.', in *ECML*, eds., Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, volume 3201 of *Lecture Notes in Computer Science*, pp. 384–395. Springer, (2004).

[18] A. Tobudic and G. Widmer, 'Relational IBL in music with a new structural similarity measure', volume 2835, pp. 365–382, (2003).

[19] Asmir Tobudic and Gerhard Widmer, 'Learning to play like the great pianists.', in *IJCAI*, eds., Leslie Pack Kaelbling and Alessandro Saffiotti, pp. 871–876. Professional Book Center, (2005).

[20] Petri Toivainen, 'An interactive MIDI accompanist', *Computer Music Journal*, **22**(4), 63–75, (Winter 1998).

[21] Gil Weinberg, 'Interconnected musical networks: Toward a theoretical framework', *Computer Music Journal*, **29**(2), 23–39, (2005).

[22] Gil Weinberg and Scott Driscoll, 'Robot-human interaction with an anthropomorphic percussionist', in *CHI 2006 Proceedings*, pp. 1229–1232, (April 2006).

[23] Gil Weinberg, Scott Driscoll, and Mitchell Perry, 'Musical interactions with a perceptual robotic percussionist', in *Proceedings of IEEE International Workshop on Robot and Human Interactive Communication*, (2005).

[24] Gerhard Widmer, 'Studying a creative act with computers: Music performance studies with automated discovery methods', *Musicae Scientiae*, **IX**(1), 11–30, (2005).

A unified framework for imitation-like behaviors

Francisco S. Melo and **Manuel Lopes** and **José Santos-Victor** and **Maria Isabel Ribeiro**[1]

Abstract. In this paper, we combine the formal methods from reinforcement learning with the paradigm of imitation learning. The extension of the reinforcement learning framework to integrate the information provided by an expert (demonstrator) has the important advantage of allowing a clear decrease of the time necessary to learn certain robotic tasks. Hence, learning by imitation can be interpreted as a mechanism for fast skill transfer. Another contribution of this paper consists in showing that our formalism is able to model different types of imitation-learning that are described in the biological literature. It thus unifies in the same abstract model what used to be addressed as separate behavioral patterns. We illustrate the application of these methods in simulation and with a real robot.

1 INTRODUCTION

In the early days of behavioral sciences, several processes used by animals to acquire new skills were often dismissed as “mere imitation”. As the knowledge of animal behavior, psychology and neurophysiology evolved, imitation has been promoted and is now considered a sophisticated cognitive capability that few species are capable of [1]. This change in the interpretation was accompanied by the discovery of several phenomena resulting in imitation-like behavior, *i.e.*, in a repetition of an observed pattern of behavior.

In social learning, a learner uses information provided by an *expert* to improve its own learning. For example, if the learner is able to observe the actions taken by a second subject, it can bias its exploration of the environment, improve its model of the world or even mimic parts of the other agent’s behavior. This process, generally dubbed as *imitation*, makes cultural transfer of knowledge fast and reliable—acquired knowledge enables fast learning. Cultural spreading becomes thus possible by a *Lamarckian* principle, where animals learn how to act by imitating others and having the same mannerisms as their peers. Through imitation, new discoveries are learnt by each individual very efficiently, simply by observation and behavior matching.

“Real” imitation occurs when a new action is added to the agent’s repertoire after having seen a demonstration. It is not enough to repeat an action after having seen it. In fact, this phenomenon can often be explained by reinforcement learning (or learning by trial-and-error). Although some social skill is usually developed when learning by trial-and-error, there is no real imitation (where *new* skills are acquired by simple observation). The concept of *imitation* is far from clear and led biologists to define several mechanisms to explain different types of *imitation-like* behaviors.

In this paper, we analyze several such imitation-like behaviors. We show how each can be modeled using a common formalism. This formalism borrows the fundamental concepts and methods from the reinforcement learning framework [2]. By considering different ways by which an expert can provide information to the learner, we feature different types of learning from observation and formalize each of the aforementioned behaviors in a reinforcement learning (RL) context.

We recall that RL addresses the problem of a decision-maker faced with a sequential decision problem and using evaluative feedback as a performance measure. The evaluative feedback provided to the decision-maker consists of a *reinforcement signal* that quantitatively evaluates the immediate performance of the decision-maker. To optimally complete the assigned task, the decision-maker must *learn* by *trial-and-error*: only sufficient exploration of its environment and actions ensure that the task is properly learnt. Therefore, in the standard RL formalism, the reinforcement signal is a fundamental element that completely describes the task to be learnt. If the agent knows how the reinforcement is assigned, it should be able to learn the task by trial-and-error (given enough time) and the information from an expert can, at most, speed up the learning process.

In real imitation as considered above, the learner should be able to acquire a new skill/learn a new task from the observations. However, and unlike the situation described in the previous paragraph, it generally should not be able to do this *without* the information provided by the expert. Considering everything stated so far, we could argue that, from a RL perspective, this corresponds to the *learning of the reinforcement function*.

Imitation has been proposed as a method to program the complex robotic systems existing today [3, 4, 5]. Programming highly-complex robots is a hard task *per se*; if a robot is capable of learning by observation and imitation, the task of programming it would be greatly simplified. To the extent of our knowledge, no systematic computational model has been proposed to formally describe imitation-like behaviors. The formalism proposed in this paper aims at fulfilling such gap. So far, the mainstream of the research in imitation aimed at individually clarifying/modeling several fundamental mechanisms individually: body correspondence [6, 7], imitation metrics [8], view-point correspondence [9] and task representation [10].

In this paper, we propose a formalism to address learning from observations. In this formalism, several types of information provided by an expert are integrated in a RL framework in different ways. We consider different assumptions on the information provided and on the way this information is integrated in the learning process, and show that this results in different imitation-like behaviors. It is our belief that the formal approach in this paper contributes to disambiguate several important concepts and clarify several issues arising in the literature on learning by imitation.

The paper is organized as follows. Section 2 reviews the main concepts in imitation learning. We describe several models of imitation in biological and artificial systems, as well as some computational problems arising in the context of imitation. Section 3 describes the framework of RL and introduces the fundamental notation. We pro-

[1] Institute for Systems and Robotics,Instituto Superior Técnico, Lisboa, Portugal.*E-mail*: {fmelo, macl, jasv, mir}@isr.ist.utl.pt

ceed in Section 4 by analyzing several methods to use expert information to speed learning. We show these methods to fall within specific classes of imitation-like behavior. To do this we describe how imitation and reinforcement can be combined and describe two simple methods to achieve this. We illustrate some of the methods in the paper in Section 5 and conclude the paper in Section 6.

2 IMITATION LEARNING

Several different mechanisms can result in a imitation-like behavior. One agent may perform an action after having seen it, but the mechanisms leading to it may be very different. Even when asking someone to imitate a hand movement, the results may vary substantially depending on the individual in question [11, 12]. From the study of imitation in animals, several mechanisms were proposed to describe an "imitative behavior", [1, 13, 4, 3]:

1. **Stimulus Enhancement** describes the general tendency to respond more vigorously toward those parts of the environment within which a conspecific is seen to interact. Seeing what are the important parts of the environment and which objects might be useful can speed up learning;
2. **Contextual Learning** describes the situation when an action is not learned, but the perception of a new object property can produce the desire to act upon it. If, for example, an animal sees someone throwing a coconut, it will learn the possibility of throwing it. In the context of our work, contextual imitation would amount to learning to employ an action, already known, in different circumstances.
3. **Response facilitation** is described in [1] as "a kind of social effect that selectively enhances responses: watching a conspecific performing an act, often resulting in a reward, increases the probability of an animal doing the same." Large flocks of birds fly in perfect synchronization. They are not imitating each other, but simply doing the same to protect themselves against predators.
4. **Emulation** can also lead to a behavioral match. Observing an action and the corresponding result might bring a desire to obtain the same goal. Learning that a coconut can be smashed to reach the inside will give the desire to eat the inside and thus producing the same behavior.

Although the mechanisms just described produce imitative behavior, they do not exactly correspond to imitation learning, in the sense that no "new actions" are learned from scratch or added to the existent repertoire. On the other hand, there is a second set of processes leading to imitative behavior where learning of new actions does actually occur. This is called *production learning* [13] and, as it is the most-powerful way of imitation, the "true-imitation" [3].

Byrne distinguishes two cases of production learning, namely **action-level** and **program-level** learning [13]:

- **Action-level learning** is defined as: "The indiscriminate copying of the actions of the teacher without mapping them onto more abstract motor representation." [3]. This is a perfect copy of the motions, if the kinematics of the systems are the same, even the joint level trajectories are the same.
- **Program-level learning** defined for the cases where not only the superficial motion is copied but when a broader description of the sequences, goals and the hierarchical structure of the behavior is inferred by the learner [14].

From the examples above we can see that many situations dubbed as imitation do not involve any actual learning, but only simultaneous/similar action. Response facilitation is just the equal answer that similar agents give when they are at the same state. Emulation and contextual learning can be explained as an improvement of the world model. The result of some action, or its relevant use in a given situation is added to the possible actions. In stimulus enhancement some task learning occurs, but the action is learned by trial-and-error, the demonstration only providing partial knowledge. In imitation, we expect the agent to learn how to complete the task or even the task itself.

2.1 Some implementation issues

Imitation cannot be reduced to supervised learning, where the agent is given the input and correct output. In imitation, the agent is given a set of observations of the environment and corresponding adequate actions. It must then *translate* this information *in terms of its own body*. This is the first difficulty in imitation: the observation is made from a different point-of-view. The different actions performed then must be *recognized* and *mapped to the agent's different capabilities*. Finally, the agent must *infer the important parts of the demonstration*. In imitation, all these problems must be carefully addressed, this being the reason why imitation is considered a complex cognitive task. We now discuss each of these three steps in detail.

Due to the problem of "seeing the world from another's viewpoint", the observed actions must be translated into the referential frame of the learner due to the different perceptual viewpoints, *i.e.*, the learner must perform a "mental rotation" to place the demonstrator's body (*allo-image*) in correspondence with the learner's own body (*ego-image*) [15, 9, 16].

Furthermore, when considering the problem of learning by imitation there is some *correspondence* assumed between the body of the demonstrator and that of the imitator. The *correspondence problem* is precisely defined as the mapping between the actions, states and effects of the demonstrator and those of the imitator. It is particularly relevant if the actions are performed by a specific body and should be replicated by a different body. Even when considering similar bodies, contextual knowledge or training may imply that the demonstrator and the imitator cannot use one same object in the same ways. And if this is not the case, there are always small differences in kinematics, size, dynamics or context that require the correspondence problem to be solved. This problem can be addressed using different methodologies. Examples include algebraic approaches [17], trajectory balance correction [18] and matching the effects of the actions [7].

Finally, it is necessary to *evaluate* the performance of the imitator. In other words, an agent needs a metric that, in a sense, allows it to determine if the imitation was successful or not. And, as expected, different metrics can will lead to different results. These *imitation metrics* evaluate how well the imitator was able to grasp underlying goal of the demonstrated task. These metrics can be selected using an algebraic formulation [8], by optimizing the learnt trajectories [19] or by considering the visual process involved [9].

Figure 1 combines the previous elements in an illustrative architecture that summarizes the relation between these elements of imitation learning [5]. In this paper we do not address the fundamental problems of view-point transformations or recognition. Instead, we assume that the learner receives the processed output of the blocks computing the VPT and performing the recognition, and focus in the problem of learning.

As will soon become apparent, we provide a unified framework to address imitation learning and reinforcement learning. We show that, in this setting, there is an imitation metric that arises naturally from the formulation of the problem of imitation. Furthermore, we describe several situations where such metric does not arise naturally from the problem formulation. We identify in each such situation

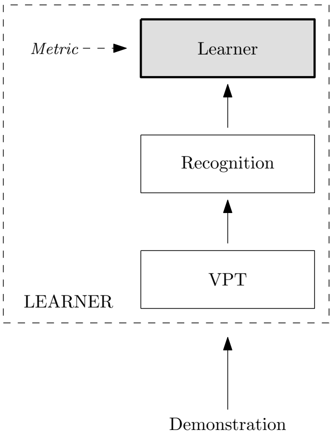


Figure 1. Architecture for the imitator.

a particular instance of *imitation-like behavior*, where the learning agent "appears" to imitate the demonstrator but where no actual imitation takes place.

3 REINFORCEMENT LEARNING

The general purpose of RL is to find a "good" mapping that assigns "perceptions" to "actions". Simply put, this mapping determines how a decision-maker reacts in each situation it encounters, and is commonly known as a *policy*. The use of evaluative feedback, by means of a *reinforcement signal*, allows the decision-maker to gradually grasp the *underlying purpose* of the task it must complete while optimizing the way of completing it.

In this section we describe *Markov decision processes*, the standard framework used to address RL problems. We also review some solution methods that we later employ in the context of imitation.

3.1 Markov decision processes

Let $\{X_t\}$ denote a controlled Markov chain, where the parameter t is the discrete time, and X_t takes values in a finite set $\mathcal{X}$, known as the *state-space*.

The distribution of each r.v. X_{t+1} is conditionally dependent on the past history $\mathcal{F}_t$ of the process according to the probabilities

$$\begin{aligned}\mathbb{P}\left[X_{t+1} = y \mid \mathcal{F}_t\right] = \mathbb{P}\left[X_{t+1} = y \mid X_t = x, A_t = a\right] = \\ = \mathsf{P}_a(x, y).\end{aligned}$$

We note that the transition kernel P depends at each time instant t on a parameter A_t, which takes values in a finite set $\mathcal{A}$. This parameter provides a decision-maker with a mechanism to "control" the trajectories of the chain by influencing the corresponding transition probabilities. We generally refer to the sequence $\{A_t\}$ as the *control process*; we refer to A_t as the *action at time instant* t and to $\mathcal{A}$ as the *action-set*.

Every time a transition from a state $x \in \mathcal{X}$ to a state $y \in \mathcal{X}$ occurs under a particular action $a \in \mathcal{A}$, the decision-maker is granted a numerical *reinforcement* $r(x, a, y)$. This reinforcement provides the evaluative feedback that the decision-maker must use to learn the desired task. The decision-maker must determine the control process $\{A_t\}$ maximizing the *expected total discounted reward*, as given by the functional

$$J(\{A_t\}, x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \mid X_0 = x\right],$$

where we denoted by R_t the reinforcement received at time t, given by $r(X_t, A_t, X_{t+1})$. Throughout the paper, we admit that the rewards are bounded, *i.e.*, , $|r(x, a, y)| \leq \mathcal{R}$ for some constant $\mathcal{R}$. Also, and to simplify the discussion, we admit r to be constant on the second and third parameters. The parameter $0 < \gamma < 1$ is a discount factor.

A *Markov decision process* (MDP) is a tuple $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$, where $\mathcal{X}$ is the state-space, $\mathcal{A}$ is the action-space, P represents the transition probabilities for the controlled chain and r is the reinforcement function.

3.2 Dynamic programming and stochastic approximation

We define a *policy* as being a state-dependent decision-rule, and denote it as a mapping $\delta_t : \mathcal{X} \times \mathcal{A} \longrightarrow [0, 1]$ assigning a probability $\delta_t(x, a)$ to each state action pair $(x, a) \in \mathcal{X} \times \mathcal{A}$. The value $\delta_t(x, a)$ represents the probability of $A_t = a$ when $X_t = x$. A policy δ independent of t is dubbed as *stationary*, and as *deterministic* if for each $x \in \mathcal{X}$ there is an $a \in \mathcal{A}$ such that $\delta_t(x, a) = 1$. In the latter case, we abusively denote by $\delta_t(x)$ the action determined by δ_t when $X_t = x$.

The *value function* associated with a policy δ_t is defined as a mapping $V^{\delta_t} : \mathcal{X} \longrightarrow \mathbb{R}$ defined for each state $x \in \mathcal{X}$ as

$$V^{\delta_t}(x) = J(\{A_t\}, x),$$

where the control process $\{A_t\}$ is generated from $\{X_t\}$ according to δ_t. Given an MDP $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$, there is at least one deterministic, stationary policy δ^* such that

$$V^{\delta^*}(x) \geq V^{\delta_t}(x),$$

for any policy δ_t and any state $x \in \mathcal{X}$. This policy can, in turn, be obtained from V^{δ^*} as

$$\delta^*(x) = \operatorname*{arg\,max}_{a \in \mathcal{A}} \Big[r(x) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}_a(x, y) V^{\delta^*}(y)\Big].$$

Any such policy is *optimal* and the corresponding value function V^{δ^*} is simply denoted by V^*. Clearly, V^* verifies the recursive relation

$$V^*(x) = \max_{a \in \mathcal{A}} \Big[r(x) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}_a(x, y) V^*(y)\Big],$$

known as the *Bellman optimality equation*. Notice that $V^*(x)$ is the expected total discounted reward along a trajectory of the Markov chain starting at state x obtained by following the optimal policy δ^*.

From V^* we define a function $Q^* : \mathcal{X} \times \mathcal{A} \longrightarrow \mathbb{R}$ as

$$Q^*(x, a) = r(x) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}_a(x, y) V^*(y).$$

The value $Q^*(x, a)$ is the expected total discounted reward along a trajectory of the chain verifying $X_0 = x$ and $A_0 = a$, obtained by following the optimal policy for $t \geq 1$.

Summarizing, we have the following relations

$$V^*(x) = \max_{a \in \mathcal{A}} Q^*(x, a); \tag{1a}$$

$$Q^*(x, a) = r(x) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}_a(x, y) \max_{b \in \mathcal{A}} Q^*(y, b); \tag{1b}$$

$$\delta^*(x) = \operatorname*{arg\,max}_{a \in \mathcal{A}} Q^*(x, a). \tag{1c}$$

Now given any functions $v : \mathcal{X} \longrightarrow \mathbb{R}$ and $q : \mathcal{X} \times \mathcal{A} \longrightarrow \mathbb{R}$, we consider the operators

$$(\mathbf{T}v)(x) = \max_{a \in \mathcal{A}} \Big[r(x) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}_a(x, y) v(y) \Big]$$

and

$$(\mathbf{H}q)(x, a) = r(x) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}_a(x, y) \max_{b \in \mathcal{A}} q(y, b).$$

It is straightforward to see that V^* and Q^* are fixed points of the operators $\mathbf{T}$ and $\mathbf{H}$. Each of these operators is a contraction in a corresponding norm and thus a simple fixed-point iteration can be used to determine V^* and Q^*.

The use of either $\mathbf{T}$ or $\mathbf{H}$ to determine V^* or Q^* by fixed-point iteration is a process known as *value iteration*. It is a dynamic programming approach that is often used to determine the function V^* and Q^* from which the optimal policy δ^* can be computed.

When this is not the case, *i.e.*, when P and r are unknown, many methods have been proposed that asymptotically converge to the desired functions [20, 2]. In this paper, we use use one of the most studied methods in the RL literature: the Q-learning algorithm [21]. This method uses sample trajectories of the Markov process, $\{x_t\}$, control process, $\{a_t\}$ and corresponding rewards $\{r_t\}$ to estimate the function Q^*. These estimates are updated according to the Q-learning update

$$\begin{aligned} Q_{t+1}(x_t, a_t) = (1 - \alpha_t(x_t, a_t)) Q_t(x_t, a_t) + \\ + \alpha_t(x_t, a_t) \big[r_t + \gamma \max_{b \in \mathcal{A}} Q_t(x_{t+1}, b) \big]. \end{aligned} \tag{2}$$

This algorithm will converge to Q^* w.p.1 as long as $\sum_t \alpha_t(x, a) = \infty$ and $\sum_t \alpha_t^2(x, a) < \infty$ for every $(x, a) \in \mathcal{X} \times \mathcal{A}$. This requires in particular that every state-action pair be infinitely often (there is sufficient exploration of the environment and the agent's actions).

4 LEARNING PARADIGMS USING EXPERT INFORMATION

In the previous sections we described two learning paradigms: learning by imitation and learning by reinforcement. In this section we move towards a combined learning framework, the *learning by observation and reinforcement* (LOR) framework. To this purpose, we consider a learning agent that must learn how to perform a sequential task using some prior knowledge and information provided by an expert.

The formalism considered herein borrows the fundamental ideas from the reinforcement learning framework described in the previous section, thus providing a unified framework to address both classes of learning processes. The fundamental assumptions usually considered in the reinforcement learning framework are:

- The task to be learnt can be described as a mapping from the set of states of the environment to the set of possible actions (a *policy*);
- The environment is *stationary*.

The first assumption simply states that in the same state of the environment the agent should always perform the same action. We remark that this assumption bears yet another important implication. If, as stated, the task to be accomplished can be fully described using a policy, then there is a reward function such that the *policy to be learnt is the optimal policy with respect to this reward function*, in the sense described in Section 3.

The second assumption above simply means that the policy used to fulfill the task should not change with time (the environment always responds to the agent's actions in the same way).

In what follows, we will consider two fundamental situations:

(i) The imitator knows the task to be learnt, but does not know how to perform this task;
(ii) The imitator does not know the task to be learnt.

From everything stated so far, it should be clear that, in terms of our formalism, the situation in (i) simply means that there is a previously defined reward function, known by the agent (since the reward function defines the task). Notice that if the agent is aware of this function, it can learn to perform the task by trial-and-error, given sufficient time. Clearly, the situation in (ii) means that *there is no reward function defined a priori*. This, of course, implies that the agent will not be able to learn any task without any further information.[2]

We analyze how different types of information provided by an expert can be integrated in learning the desired task. As will soon become apparent, models for the imitation-like behaviors described in Section 2 arise naturally in the LOR framework. We also show that, in the more complex scenario of an unknown task, it is possible to provide a natural interpretation for the used algorithm in terms of imitation metrics. The first case, where the agent does know the task, does not correspond to "real" imitation behavior as defined in Section 2. Only in the second situation, where the task is not defined beforehand, can we speak about true imitation. We further comment on this issue at the end of the section.

We consider each of the two situations (i) and (ii) in Subsections 4.1 and 4.2, respectively.

4.1 Known task

We consider that the interaction of the learning agent and the environment can be described as a controlled Markov chain, as in Section 3. This means that, at each time t, the state of the environment will move from a state $X_t = x$ to a state $X_{t+1} = y$ depending on the action A_t of the agent and according to the transition probabilities $\mathsf{P}_a(x, y)$. We suppose that an expert provides the learning agent with some information on how the task can be completed. We refer to such information generally as a *demonstration* and analyze how can this information be used in the learning process. We consider four distinct cases:

(i) The demonstration consists of a sequence of states,

$$\mathcal{H} = \{x_1, \ldots, x_N\},$$

obtained by following the optimal policy;

(ii) The demonstration consists of a sequence of state-actions pairs,

$$\mathcal{H} = \{(x_1, a_1), \ldots, (x_N, a_N)\},$$

"hinting" on which should be the optimal action a_i at each state x_i visited;

(iii) The demonstration consists of a sequence of transition triplets,

$$\mathcal{H} = \{(x_1, a_1, y_1), \ldots, (x_N, a_N, y_N)\},$$

providing the imitator with information on the behavior of the environment;

[2] We could argue that the situation in (ii) means that the agent *does not know* the reward function, but that the latter is defined. We do not adopt such position for the simple reason that, if a reward function *is* defined, the agent can still learn by trial-and-error and, therefore, there is no significative difference from (i).

(iv) The demonstration consists of a sequence of transition-reward tuples,

$$\mathcal{H} = \{(x_1, a_1, r_1, y_1), \ldots, (x_N, a_N, r_N, y_N)\},$$

providing the imitator with information on the behavior of the environment and on how the task should be completed.

First of all, we remark that, since we assume knowledge of r, (iii) and (iv) are redundant. Nevertheless, we will consider how to address the two situations distinctly, noting that in (iv) allows to address situations in which r is unknown.

We must further detail the idea behind each of the previous classes of demonstrations. The first situation, (i), addresses situations in which the learning agent *can not observe/recognize the actions of the demonstrator* but only their effect in the environment. This information will show the agent how the state of the environment should evolve when the optimal policy is implemented. A sequence as described in (ii) illustrates *how the task can be completed*. Each pair (x_i, a_i) is related through some deterministic policy δ that is "close" to optimal. Sequences as those described in (iii) and (iv) illustrate *the dynamics of the environment* in terms of transitions and transitions-rewards, respectively. Unlike the sequences described in (ii), it is not assumed that x_i and a_i in each tuple (x_i, a_i, y_i) or (x_i, a_i, r_i, y_i) are related by some policy.[3]

Another important aspect is that, at this stage, we are not concerned with the particular way by which the sequences $\mathcal{H}$ in (ii) through (iv) are obtained. Consider for example the situation in (ii). It may occur that the demonstrator illustrates how the task is completed by demonstrating the action to be chosen in an arbitrary set of states $\{x_1, \ldots, x_N\}$. Or, it may happen that the sequence of states $\{x_1, \ldots, x_N\}$ is actually a sample path of the process obtained with the control sequence $\{a_1, \ldots, a_{N-1}\}$.

We also remark that, in all 3 cases listed above, we assume that the imitator is able to perceive the information in the sequences $\mathcal{H}$ unambiguously. We could admit *partial observability*, meaning that the imitator was able to observe the states, actions and/or rewards in the sequences $\mathcal{H}$ only up to some degree of accuracy. This would imply that the imitator would have to *estimate* what the actual state, action and/or reward would have been. This, of course, would be the actual case in practical situations. Nevertheless, consideration of partial observability adds no useful insight to our formalization of the imitation problem and significantly complicates the presentation.

The four methods presented below all provide an initial estimate Q_0 for Q^* that integrates the information provided by the demonstration. We will see that this informed initialization brings a significant improvement in the learning performance of the agent.

Method 4.1.1: Sequence of states

Consider a sequence of states

$$\mathcal{H} = \{x_0, \ldots, x_N\},$$

obtained according to the optimal policy. As stated, this first scenario comprises situations where the learning agent is not able to observe/recognize the actions performed by the expert. Nevertheless, the sequence of states $\mathcal{H}$ provides the learning agent with an idea on how the environment evolves "under" the optimal policy.

Therefore, if the transition model is known, the agent can compute

$$Q_0(x, a) = r(x) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}_a(x, y) V^*(y),$$

where V^* is computed as $V^* = (\mathbf{I} - \gamma \mathsf{P}^*)^{-1} r$. The matrix $\mathbf{I}$ denotes the identity and the transition matrix P^* represents the transition model for the optimal policy, estimated from $\mathcal{H}$ as

$$\mathsf{P}^*(x, y) = \frac{N(x, y)}{\sum_{z \in \mathcal{X}} N(x, z)},$$

where $N(x, y)$ denotes the number of times that a transition from x to y occurred in $\mathcal{H}$. This method is similar to that proposed in [22].

Method 4.1.2: Sequence of state-action pairs

Consider a sequence of state-action pairs

$$\mathcal{H} = \{(x_1, a_1), \ldots, (x_N, a_N)\}.$$

Each demonstrated pair (x_i, a_i) provides significant information on the *optimal policy* at x_i. And even if the policy partially defined by $\delta(x_i) = a_i$ is not optimal, it is expectable that it is "close" to optimal. It is therefore reasonable that the imitator *uses δ as an initial policy to perform the task*. And, as it acquires further experience on the task, it should be able to improve from this initial policy, if there is room for such improvement. To incorporate this information in the initial estimate for Q^*, we set $Q_0(x_i, a_i) = 1$ for $i = 1, \ldots, N$ and 0 otherwise.

Method 4.1.3: Sequence of transition triplets

We now consider a sequence of transition triplets

$$\mathcal{H} = \{(x_1, a_1, y_1), \ldots, (x_N, a_N, y_N)\}.$$

As mentioned above, this sequence provides the imitator with *information on the behavior of the environment*. Clearly this is only useful if the transition probabilities are not known *a priori*. If this is the case, the information provided by the demonstrator can be used to improve the model of the environment by setting

$$\hat{\mathsf{P}}_a(x, y) = \frac{N(x, a, y)}{\sum_{z \in \mathcal{X}} N(x, a, z)},$$

where $N(x, a, y)$ denotes the number of times that the triplet (x, a, y) was observed in $\mathcal{H}$. This estimated transition model $\hat{\mathsf{P}}$ with the function r can be used to perform value iteration and obtain an initial estimate Q_0 for the learning algorithm.

Method 4.1.4: Sequence of transition-reward tuples

Finally, we consider a sequence of transition-reward tuples

$$\mathcal{H} = \{(x_1, a_1, r_1, y_1), \ldots, (x_N, a_N, r_N, y_N)\}.$$

This sequence provides the imitator with information on the behavior of the environment and on *the task*. This means that the tuples in $\mathcal{H}$ can be used to perform N iterations of Q-learning using (2). The resulting Q-function provides the initial estimate Q_0 for the learning algorithm.

[3] We make this distinction as each of the sequences described in (i) through (iv) provides the imitator with different information, to be used in different ways. This is not limiting in any way, as discussed below.

4.2 Unknown task

In this subsection, we use the exact same formulation considered in Subsection 4.1 above, but suppose that *no reward mechanism is defined*. This means that the imitator is no longer able to learn the task by trial-and-error if no demonstration is available.

However, if a demonstrator provides the imitator with some information on how the task can be completed, the imitator can *build* its own reward function and use it to learn how to perform the task. We also refer to such information generally as a *demonstration*.

Unlike in the previous situation, we only consider two scenarios: We consider four distinct cases.

(i) The demonstration consists of a sequence of states,

$$\mathcal{H} = \{x_1, \ldots, x_N\}.$$

(ii) The demonstration consists of a sequence of transition triplets,

$$\mathcal{H} = \{(x_1, a_1, y_1), \ldots, (x_N, a_N, y_N)\},$$

providing the imitator with information on the behavior of the environment.

Notice that, since there is no reward function defined, it is not possible to consider the situation where transition-reward tuples are observed. Also, and unlike Subsection 4.1, we now assume that the transition triplets in $\mathcal{H}$ considered in (ii) are obtained *using the policy to be learnt*. Therefore, (ii) includes both (ii) and (iii) from the previous subsection.

Method 4.2.1: Sequence of states

Consider a sequence of states

$$\mathcal{H} = \{x_0, \ldots, x_N\},$$

obtained according to the optimal policy. As in Subsection 4.1, this scenario comprises situations where the learning agent is not able to observe/recognize the actions performed by the expert.

We interpret the sequence of states in $\mathcal{H}$ as providing the learner with information *on the goal* of the task. In particular, we consider that $\mathcal{H}$ represents a *possible trajectory to a goal state*. Therefore, the learner will memorize the last state visited, x_N, as the goal state and build a simple reinforcement function defining the task "reach the goal state as fast as possible". An example of one such reward function is

$$r(x) = \begin{cases} +10 & \text{if } x = x_N; \\ -1 & \text{otherwise.} \end{cases}$$

The agent can now apply any preferred method to determine the optimal policy. For example, it can use value iteration if P is known, or Q-learning otherwise. The learner will thus learn a policy that will partially replicate the demonstration observed.

Method 4.2.2: Sequence of transition triplets

We now consider a sequence of transition triplets

$$\mathcal{H} = \{(x_1, a_1, y_1), \ldots, (x_N, a_N, y_N)\}$$

obtained using the "optimal policy". As in Subsection 4.1, this sequence can be used to improve the model of the environment. This model of the environment can, in turn, be used to determine the reward function that best translates the policy partially defined by $\delta(x_i) = a_i$, $i = 1, \ldots, N$. The approach considered here differs from that used in Method 4.2.1 in that the reward function is no longer built by considering only one final state. Instead, the learning agent will use the *whole demonstration* and apply inverse reinforcement learning to build the reward function [23]. We will show that this procedure is fundamentally different from the previous ones, and corresponds to "real imitation" in the sense of Section 2.

4.3 Classification of the learning paradigms

So far in this section we formalized several different methods by which an agent can use the information provided by an expert in learning how to accomplish a task. However, as discussed in Section 2, there are several learning paradigms that do exhibit imitative behavior but which cannot be truly classified as "imitation". And, as we show in the continuation, most of the methods described above actually fall in one of the following categories:

- Stimulus enhancement;
- Contextual learning;
- Response facilitation;
- Emulation.

We start with the Method 4.1.1. In this method, the learning agent seeks to replicate the *effect* of the actions of the demonstrator. This will actually lead to an initial replication of the demonstrator's policy, but the process by which this behavioral match is attained is *emulation*.

In Method 4.1.2, the imitator uses the demonstration to *bias its learning strategy*. Therefore, this method is actually a *stimulus enhancement* mechanism: the imitator observes some actions that can be useful for the task and uses this information to speed learning.

In Method 4.1.3, the imitator uses the demonstration to *improve its model of the world*. This means that the imitator gains further knowledge on what the consequences of some of its actions may be. We can classify this as a subtle form of *contextual learning*.

A similar thing occurs in Method 4.1.4. In this method, however, the imitator further observes *the rewards* obtained by the imitator. It realizes not only the consequences of some actions but also on *how these actions contribute to complete the task*. This use of the reward information allows us to realize that Method 3 combines contextual learning with *response facilitation*.

Notice that, in all these methods, the agent already knows the task to be learnt. This means that, with enough time, the agent could learn the task without any help from a demonstrator. Furthermore, independently of the policy used in the demonstration, the agent will eventually learn the correct policy, completely disregarding the demonstration if necessary. This means that the demonstration only provides a means for the agent to speed up its own learning process. Therefore, it is not surprising that all these situations do not correspond to "true-imitation" behaviors.

Moving to the the methods in Subsection 4.2, we start by noticing that, in Method 4.2.1 the agent seeks to replicate the final *effect* of the actions of the demonstrator. In fact, in this method, the agent focuses all its learning in *replicating the effect* observed in the demonstration (in terms of final state), displaying a flagrant example of *emulation*.

On the other hand, Method 4.2.2 seeks to *extrapolate the task behind the actions of the demonstrator*. From this information, the agent builds a reward function that will eventually lead to a replication of the demonstrator's policy. However, the actual method for computing this reward function (and, thus, realizing the task to be learnt) provides important insights into the problem of imitation, that we discuss next.

4.4 Inverse reinforcement learning and imitation metrics

As argued in Section 2, "true" imitation will occur if a broad description of the action sequences, goals and hierarchical structure of the desired behavior is inferred by the learner. As we have seen, in the RL formalism, the goals and structure of the desired behavior are "encoded" in the reward function. Therefore, learning the reward function and using it to determine the optimal policy would fit the above description of true imitation.

Notice that we consider Method 4.2.1 to be emulation because two completely different sequences ending in a common final state will lead the learning agent to infer the exact same reward function. This means that, as stated in the previous subsection, this method seeks to replicate the effect of the actions of the demonstrator rather than to extrapolate the task behind the actions of the demonstrator.

On the other hand, Method 4.2.2 does seek to extrapolate this information from the demonstration. To better realize how this method operates, we provide a brief description of its working [23].

Given the model of the environment (namely the transition probabilities in P), the inverse reinforcement learning method used (dubbed *Bayesian inverse reinforcement learning*—BIRL) searches the space of possible reward functions. To this purpose, the method considers a fine discretization of the referred space of reward functions. Then, given any initial reward function, the method evaluates the optimal Q-function, Q^*, for this reward function and evaluates the *likelihood of the demonstrated policy being optimal* given Q^*. This likelihood also takes into consideration a numerical parameter describing the *confidence on the optimality of the demonstrated policy*. The method will thus output the most likely reward given the demonstrated policy (obtained from $\mathcal{H}$) and the confidence parameter.

We emphasize several important aspects of this approach. First of all, this method considers the demonstration *as a whole*, instead of focusing on particular aspects. Therefore, the reward thus determined will more accurately the task "behind" the demonstration. On the other hand, the likelihood function used to compare different reward functions as well as the confidence parameter naturally provide an imitation metric for the problem. The inclusion of the confidence parameter is an important aspect that allows the agent to realize how strict it should follow the provided demonstration. A low confidence parameter will result in a learnt policy significatively more different from the demonstrated policy than a high confidence parameter.

Also notice that considering imitation metrics makes no sense in the other methods. In the methods in Subsection 4.1 the demonstration is only used to speed the learning. The agent is not trying to replicate the demonstration but to optimize its policy with respect to the pre-defined rewards. In Method 4.2.1, on the other hand, the agent is simply trying to reach the final state observed in the demonstration. Once again, is not trying to replicate the demonstration but to optimize the policy leading it to this goal state.

The reward function thus constructed will provide adequate evaluative feedback on the task and the imitator can use this evaluative feedback to optimize its own policy. We emphasize that, without the demonstration, the imitator *has no knowledge on the task*. The reward function built from the demonstration is, therefore, *new knowledge* that describes the task at hand and allows the imitator to learn how to perform it in an optimal fashion.

4.5 Discussion

With the methods above we conclude the presentation of the LOR framework. Within this framework, we model an agent's environment as a controlled Markov chain $\{X_t\}$. The demonstration provided by an expert is, in turn, described as a sequence $\mathcal{H}$ which can take various forms, depending on the information provided. The formalism considered herein borrows fundamental ideas from reinforcement learning and provides a unified framework to address both classes of learning processes.

We notice that the MDP model considered in this paper is the simplest model used in reinforcement learning. We are interested in establishing a unified framework to address both learning by imitation and reinforcement and thus focus on this simpler model for the sake of clarity. In Section 6 we briefly comment on how the fundamental framework considered herein can be extended to accommodate richer RL models (such as POMDPs).

As argued in Section 2, imitation cannot be reduced to supervised learning and, therefore, the framework presented here should not be seen as simple a combination of supervised learning and reinforcement learning.[4] Instead, it should be seen as a formalism to describe learning processes in which imitation and reinforcement learning can be properly modeled.

It is possible to find other works in the literature that combine learning by imitation and reinforcement. In [22], imitation arises *implicitly* in non-interactive multiagent scenarios. In it, a learning agent uses the trajectories observed from other agents to speed the learning of its individual task (which is generally independent of that of the others). In yet another example, [25], a learning method is proposed that learns a reinforcement function and dynamic model from the demonstration of an expert (human executer). This is then combined with a model-free, task-level direct learner to compensate for modeling errors.

Our work is fundamentally different from those considered above in that our aim is to understand how can the problem of imitation be modeled and how can imitative-like behaviors be distinguished with a formal perspective. Nevertheless, several methods described in our paper can be seen as simplified versions of the methods described in those papers.

Also, as argued in Section 2, we considered that in order for the learning mechanism to be properly classified as *imitation*, it should be able to *realize* the task from the demonstration. However, it should be flexible enough to feature two possible behaviors: to *replicate* the exact behavior of the demonstrator or, instead, to *perceive* the purpose of the task and, eventually, optimize beyond whatever it observed. As discussed in the previous subsection, the use of Method 4.2.2 verifies all these requisites. On the other hand, each of the remaining methods exhibits one of the above features, not all. This is the reason why we classified them as imitation-like.

Finally, we remark that the classical inverse reinforcement learning algorithms [26, 27] also determine a reward function given a policy. The difference from these methods to the one used here is that BIRL allows the policy to be only *partially specified* and *suboptimal*. This is an important advantage in the problems considered herein.

5 EXPERIMENTS

We conducted several simple experiments to evaluate the performance of proposed methods against that of simple trial-and-error. We evaluated each of the methods described in Section 4.

[4] Such approach is adopted, for example, in [24], where a supervisor is combined with an actor-critic learning architecture.

The task considered is a simple recycling game, where a robot must separate the objects in front of him according to its shape (Fig. 2). In front of the robot are two slots (Left and Right) where 3 types of objects can be placed: Large Ball, Small Ball and Box. The boxes should be dropped in the corresponding container and the small balls should be kicked out of the table. The large balls should be touched upon. Every time a large ball is touched, all objects are removed from the table.

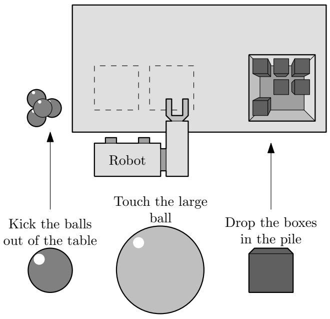


Figure 2. Simple recycling game.

The robot has, therefore, 6 possible actions: Touch Left (TL), Touch Right (TR), Kick Left (KL), Kick Right (KR), Grasp Left (GL) and Grasp Right (GR). We notice that, if the robot kicks a ball on the right while an object is lying on the left, the ball will remain in the same spot. The robot receives a reward of $+10$ every time the table is empty and -1 every other time.

The correct policy for this game is to touch the large ball, if there is any, or get rid of the object on the left and then of the object on the right (there are some situations where the order is not important). Every time the table is emptied, the game is restarted.

We tested the performance of the 4 Methods in Subsection 4.1 when the optimal policy is demonstrated and a suboptimal policy is demonstrated. We compared the performance of an agent using the information provided by the demonstration with that of an agent that has no previous information on the task. In all situations we allowed both agents to learn for 200 time steps using an ε-greedy policy with decaying ε.

Table 1 provides the percentage of time (out of the 200 time steps) that the agents are able to reach the goal state (empty table). For the sake of comparison, we also provide the performance of a "pure" reinforcement learner.

Table 1. Results obtained with Methods 4.1.1 through 4.1.4 using optimal and suboptimal demonstrations.

	Optimal	**Suboptimal**
Pure RL	34.6 %	32.4 %
Method 4.1.1	41.5 %	40.5 %
Method 4.1.2	41.5 %	37.5 %
Method 4.1.3	41.5 %	39.0 %
Method 4.1.4	42.0 %	41.0 %

From Table 1 it is evident that the performance of the learning algorithm is improved when considering a demonstration, since the agents were able to reach the goal state (and thus complete the task) more often. To have a clearer understanding of how this translates in terms of the learning process, we present in Figures 3 through 6 the total reward obtained during learning.

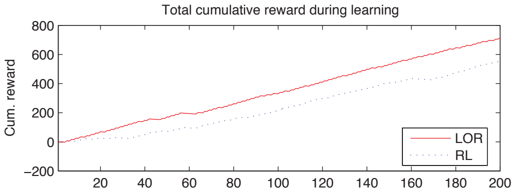


Figure 3. Total reward obtained with Method 4.1.1 over the time-frame of 200 steps when the demonstrator follows an optimal policy.

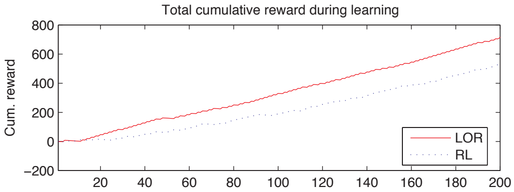


Figure 4. Total reward obtained with Method 4.1.2 over the time-frame of 200 steps when the demonstrator follows an optimal policy.

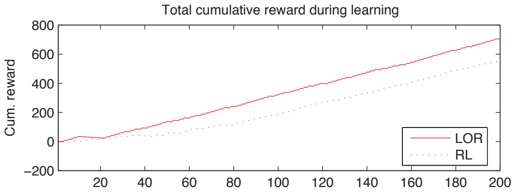


Figure 5. Total reward obtained with Method 4.1.3 over the time-frame of 200 steps when the demonstrator follows an optimal policy.

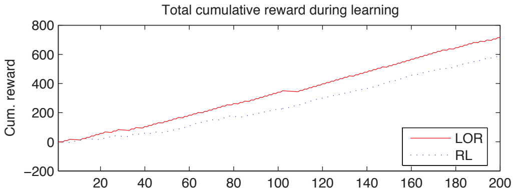


Figure 6. Total reward obtained with Method 4.1.4 over the time-frame of 200 steps when the demonstrator follows an optimal policy.

In the plots, the slope of the performance curve indicates how good is the learnt policy. It is clear that, in all methods, the provided information gives the learning agent a significative advantage: in the beginning of the learning process, the "greedy" action for the agents that were provided a demonstration is much more informed than that of the pure RL learner. This means that the demonstration provides the learner with a *knowledge boost* by improving the estimative of the optimal Q-function and thus speeding up the learning.

Notice that, in all these methods, the demonstration provides only informed initial estimates for Q^*, thus improving the initial performance of the agent. However, since this initial estimate is then properly adjusted by the learning algorithm, the sub-optimality of the demonstrated policy does not affect the performance of the learner.

In a second set of experiments we tested Method 4.2.1 from Subsection 4.2. To evaluate the performance of the method, we explicitly observed the learnt policy when the demonstrated policy is optimal and when it is not. The results are summarized in Table 2. We denoted by 0 the empty slot, by B the large ball, by c the cube and by b the small ball.

Notice that both learnt strategies are optimal. This is due to the fact that, in considering the same final state, the reward function obtained by Method 4.2.1 is the same independently of the actual policy used

Table 2. Learnt policies with Method 4.2.1 using optimal and suboptimal demonstrations.

	Optimal	**Suboptimal**
$(0,0)$	TL	TL
$(0,B)$	TR	TR
$(0,c)$	GR	GR
$(0,b)$	KR	KR
$(B,0)$	TL	TL
(B,B)	TR	TL
(B,c)	TL	TL
(B,b)	TL	TL
$(c,0)$	GL	GL
(c,B)	TR	TR
(c,c)	GR	GR
(c,b)	GL	GL
$(b,0)$	KL	KL
(b,B)	TR	TR
(b,c)	GR	KL
(b,b)	KL	KL

to demonstrate. And, in this particular case, it matches exactly the reward function considered in the previous examples, thus giving rise to the same policy.

Finally, we tested Method 4.2.2 from Subsection 4.2. As in the previous experiment, we evaluate the performance of the method by explicitly observing the learnt policy when the demonstrated policy is optimal and when it is not. The results are summarized in Table 3. In the third column we also present the results obtained with Method 4.2.2 using an optimal policy, but where the model is also estimated from the demonstration. The table elements in bold denote “suboptimal” actions.

Table 3. Learnt policies with Method 4.2.2 using optimal and suboptimal demonstrations.

	Optimal	**Suboptimal**	**No Model**
$(0,0)$	TL	TL	TL
$(0,B)$	TR	**TL**	**TL**
$(0,c)$	GR	**TR**	GR
$(0,b)$	KR	KR	KR
$(B,0)$	TL	**KL**	TL
(B,B)	TR	**GL**	TL
(B,c)	TL	**TR**	TL
(B,b)	TL	**TR**	**TR**
$(c,0)$	GL	**TL**	GL
(c,B)	TR	**GL**	**TL**
(c,c)	GR	GR	**TR**
(c,b)	GL	**KR**	**KL**
$(b,0)$	KL	KL	**TR**
(b,B)	TR	**KL**	**KL**
(b,c)	GR	**TL**	**TL**
(b,b)	KL	**TR**	**TR**

We emphasize the policy obtained with Method 4.2.2 when the demonstrated policy is suboptimal (and the agent has little confidence on the observed policy). Recall that this method determines a likely reward function for which demonstrated policy, we expect the performance of this method to be affected by the sub-optimality of the demonstrated policy. Notice that the policy learnt from a suboptimal demonstration is even worse than that learnt in the absence of a model with an optimal demonstration (third column of Table 3).

To conclude this section, we present the images obtained by experimenting Method 4.2.1 in a real robot. The robot is capable of recognizing the actions Grasp, Touch and Kick as well as the objects on the table (to details refer to [7]). Figure 7 presents the robot following the task it learned after having observed it.

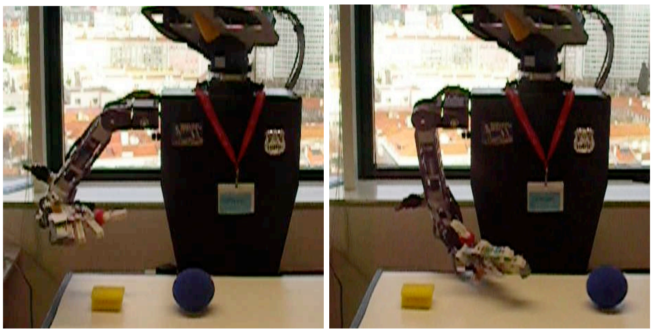


Figure 7. Robot following the learned task.

6 CONCLUSIONS

In this paper, we proposed an unified formalism to address imitation learning and RL problems. Using this formalism, we analyzed several imitation-like learning mechanisms, such as stimulus enhancement, response facilitation, contextual learning and emulation. These mechanisms can lead to imitative behavior without being imitation in the stricter sense of the concept. In this formalism, which we refer as the *learning by observation and reward* (LOR), these behaviors can be summarized as:

- *Stimulus enhancement*: biases exploration using the observed policy;
- *Contextual learning*: uses the observed transitions to improve the model of the world;
- *Response facilitation*: uses the observed transitions/rewards to improve the model of the world and accelerate learning;
- *Emulation*: uses the observed sequence of states, to either replicate the dynamics of the underlying Markov chain or final state.

One of the major contributions of the paper was to unify all of these mechanisms using a common formalism. We showed that this modelation is possible and the resulting behavior of the learner matches the descriptions of the corresponding behaviors in animals. We also discussed that, when learning a task from an expert, there are many sources of information and each of them can be exploited individually or in combination.

The results presented clearly established one of the known advantages of imitation learning: the imitation learner acquired the optimal policy for the problem faster than a learner following a standard trial-and-error learning strategy. We emphasize that, in the discussed cases of imitation-like behavior, the agent would still be able to learn the task on its own—the learner did not *infer* the solution from the demonstration. Instead, the demonstration provided *hints* on how to solve the task that the learner used to learn the task more efficiently.

It is interesting to note that, as these mechanisms do not rely completely on the details of the demonstration, they can also learn the optimal policy even when the demonstration is sub-optimal. The learner can thus look at someone performing a task and then understand the goal of the task and outperform the teacher.

We also emphasize the difference between imitation-like behaviors and “pure” imitation methods. In a pure imitation system, the found solution should not exist in the learner repertoire; or it should not be possible to know the task if it were not for the demonstration. In our formalism this translates into the fact that, without the demonstration, the agent does not know the task (there is no pre-defined reward mechanism). In imitation-like methods this reward function previously exists and the learner can always learn the task on its own.

The demonstrations used throughout the paper do not illustrate the full potential of the different methods, mainly due to the great simplicity of the task considered—the state and action spaces are small and the task is easily defined by a very simple reward function.

In our proposed LOR framework it is not easy to distinguish between action-level and program-level learning, since the important steps of the demonstration are abstract concepts that can be interpreted and implemented in different ways. We intend to address this problem with further detail by defining an hierarchical learner where we can define actions at several "resolutions". We also intend to study the effects of partial observability of state and action in learning by imitation.

Acknowledgements

This work was partially supported by Programa Operacional Sociedade do Conhecimento (POS_C) that includes FEDER funds, and by the EU-Project Robotcub. The first author acknowledges the PhD grant SFRH/BD/3074/2000.

References

[1] Richard Byrne. *The Thinking Ape Evolutionary Origins of Intelligence*. Oxford University Press, 1995.

[2] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[3] S. Schaal. Is imitation learning the route to humanoid robots. *Trends in Cognitive Sciences*, 3(6):233–242, 1999.

[4] S. Schaal, A. Ijspeert, and A. Billard. Computational approaches to motor learning by imitation. *Phil. Trans. of the Royal Society of London: Series B, Biological Sciences*, 358(1431):537–547, 2003.

[5] Manuel Lopes and José Santos-Victor. A developmental roadmap for task learning by imitation in robots. *IEEE Trans. Systems, Man, and Cybernetics - Part B: Cybernetics*, 37(2), 2007.

[6] Hideki Kozima, Cocoro Nakagawa, and Hiroyuki Yano. Emergence of imitation mediated by objects. In *2nd Int. Workshop on Epigenetic Robotics*, 2002.

[7] Luis Montesano, Manuel Lopes, Alexandre Bernardino, and José Santos-Victor. Learning affordances objects: From sensory motor maps to imitation. Technical report, Instituto de Sistemas e Robótica, Lisbon, Portugal, Feb 2007.

[8] Chrystioher L. Nehaniv and Kerstin Dautenhahn. Like me? - measures of correspondence and imitation. *Cybernetics and Systems*, 32:11–51, 2001.

[9] Manuel Lopes and José Santos-Victor. Visual transformations in gesture imitation: What you see is what you do. In *IEEE Int. Conf. Robotics and Automation*, 2003.

[10] R. Zöllner and R. Dillmann. Using multiple probabilistic hypothesis for programming one and two hand manipulation by demonstration. In *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2003.

[11] H. Bekkering, A. Wohlschläger, and M. Gattis. Imitation of gestures in children is goal-directed. *Quarterly J. Experimental Psychology*, 53A:153–164, 2000.

[12] György Gergely, Harold Bekkering, and Ildikó Király. Rational imitation in preverbal infants. *Nature*, 415:755, 2002.

[13] Richard W. Byrne. Imitation of novel complex actions: What does the evidence from animals mean? *Advances in the Study of Bahaviour*, 31:77–105, 2002.

[14] R. W. Byrne and A.E. Russon. Learning by imitation: a hierarchical approach. *Behav Brain Sci*, pages 667–84, 1998.

[15] J.S. Bruner. Nature and use of immaturity. *American Psychologist*, 27:687–708, 1972.

[16] Minoru Asada, Yuichiro Yoshikawa, and Koh Hosoda. Learning by observation without three-dimensional reconstruction. In *Intelligent Autonomous Systems (IAS-6)*, 2000.

[17] C. Nehaniv and K. Dautenhahn. Mapping between dissimilar bodies: Affordances and the algebraic foundations of imitation. In *European Workshop on Learning Robots*, 1998.

[18] S. Nakaoka, A. Nakazawa, K. Yokoi, H. Hirukawa, and K. Ikeuchi. Generating whole body motions for a biped humanoid robot from captured human dances,. In *ICRA*, Taipei, Taiwan, 2003.

[19] Aude Billard, Y. Epars, S. Calinon, G. Cheng, and S. Schaal. Discovering optimal imitation strategies. *Robotics and Autonomous Systems*, 47:2-3, 2004.

[20] Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[21] Christopher J. C. H. Watkins. *Learning from delayed rewards*. PhD thesis, King's College, University of Cambridge, May 1989.

[22] Bob Price and Craig Boutilier. Accelerating reinforcement learning through implicit imitation. *J. Artificial Intelligence Research*, 19:569–629, 2003.

[23] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. Proc. 20th Int. Joint Conf. Artificial Intelligence, 2007. (to appear).

[24] M.T. Rosenstein and A.G. Barto. Supervised actor-critic reinforcement learning. In *Learning and Approximate Dynamic Programming: Scaling Up to the Real World*. John Wiley & Sons, Inc., 2004.

[25] Christopher G. Atkeson and Stefan Schaal. Robot learning from demonstration. In *14th International Conference on Machine Learning*, pages 12–20. Morgan Kaufmann, 1997.

[26] Andrew Y. Ng and Stuart J. Russel. Algorithms for inverse reinforcement learning. In *Proc. 17th Int. Conf. Machine Learning*, pages 663–670, 2000.

[27] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proc. 21st Int. Conf. Machine Learning*, pages 1–8, 2004.

When Training Engenders Failure to Imitate in Grey Parrots (*Psittacus erithacus*)

Irene M. Pepperberg[1]

Abstract. The initial study on avian behaviour [1] was not designed to examine imitation, but nevertheless provided information concerning issues involving imitation. Four Grey parrots (*Psittacus erithacus*) were tested on their ability to obtain an item suspended from a string such that multiple, repeated, coordinated beak-foot actions were required for success (e.g., [2]). Those birds with little training to use referential English requests (e.g., "I want X") succeeded, whereas birds who could vocally request the suspended item failed to obtain the object themselves and instead engaged in repeated requesting [1]. Interestingly, even after subsequent, multiple observations of the actions of a successful parrot, the unsuccessful birds persevered in vocal requests or ignored the task, possibly retreating into learned helplessness. Such data emphasize three points: First, the entire behavioural repertoire and history must be examined in studies that try to determine whether animals act intelligently and/or can imitate; second, parrots can attempt to direct humans to assist them in achieving their goals, and such behaviour—although clearly complex—might lead them to fail certain tasks designed to test intelligence; third, even for a species known for imitative behaviour (physical as well as vocal [3]), imitation may not be expressed if it must overcome previous training.

1 INTRODUCTION

Defining and evaluating intelligence is a dauntless task with respect to humans (e.g., [4]) and is even more so with respect to nonhumans [5]: To examine nonhuman abilities, should an experimenter administer what are basically human tasks to nonhumans, making minimal concessions and adaptations to, for example, take into account their tendencies to peck a lit button rather than a computer keyboard, or instead restructure the tasks to accommodate any significantly different species-specific traits, such as poor vision and excellent olfaction? No simple solution exists, but one possible route through these difficulties is to examine not the ability to solve a specific problem but rather the *processes* whereby problems of ecological or ethological interest are solved. Consequently, researchers have become enamored of two types of studies—those involving insight and imitation. The first is favored because success suggests that the subject has formed a sophisticated representation of the problem and attained a solution via mental rather than physical trial-and-error, implying such an advanced understanding of—and memory for—actions and outcomes that physical experimentation is unnecessary. The second has become popular because success suggests that the subject can view, conceptualize, and then recreate from his/her own perspective, novel and improbable actions that lead to successful solution of a novel problem [6], also implying advanced cognitive processing skills. (The question also arises as to whether emulation—the attainment of the demonstrated goal via any means (e.g., [7])—is more or less advanced than imitation, but that is a separate issue). Of course, unless the experimenter knows the complete history of the subject, success or failure on a task might not be an accurate evaluation of capacities for insight or imitation, but rather relate to prior experience that may have either potentiated or blocked the targeted behaviour. And therein lies the question to be addressed in processing the results of both the prior [1] and present studies.

The initial experiment [1] was designed to examine whether Grey parrots (*Psittacus erithacus*) were capable of insightful behaviour; only later were the birds tested on their imitative competence. The task chosen, to obtain a special food treat suspended by a string, by reaching down, pulling up a loop of string onto the perch, stepping on the loop to secure it, and repeating the sequence several times (e.g., to demonstrate an understanding of intentional means-end behaviour; see review in [8]) has been previously used to assess "insight" in several avian species; simply reaching down for the food is not sufficient [2,9]. Not all birds succeed on this task [2,6,10; for more recent studies and reviews of older studies, see 11,12], suggesting that the necessary action pattern requires a higher-order cognitive ability that is prevalent neither among species nor within a given species.

Clearly, the extent to which the task is solved individually via insight might be affected by prior physical manipulative experience [11], but could prior training affect the ability to derive a solution via imitation of an expert? I had previously found [1] that for Grey parrots the capacity (or possibly willingness) to use insight could be tempered by a nonphysical type of training: specifically, that of my birds having learned to demand access to various objects vocally. Precisely because some of my birds can routinely request items from a human, without the need to work to obtain it on their own, two of the four birds tested (those with this vocal ability) failed the test of insight, persisting in their vocal requests. Possibly, a bird that responds with repeated requests, although ostensibly failing at the given task, could be considered to have demonstrated instead an alternative higher-order intelligence, in that it knows how to manipulate another individual to access its wants. Might, however, this ability to manipulate others interfere not only with the use of insight but also with the use of imitation? Two birds in the prior study observed but did not imitate after viewing a single trial by a successful demonstrator; the present study was performed to determine if repeated viewings of a demonstrator might be required to initiate any observational learning of a physical act.

I will review the initial study (reported in [1]), then describe subsequent trials to determine whether the parrots would engage in any behaviour related to observational learning of the string-pulling task. I compare my findings to

[1] Departments of Psychology, Harvard and Brandeis Universities, Cambridge and Waltham, MA, USA.

those using the same task for other avian species [11,12].

2 METHOD

2.1 Subjects

Four Grey parrots were the subjects of the insight part of the study [1] and three of the four were also involved in the imitation part of the study. Kyaaro, obtained from a breeder at 3 months old, was, when tested on insight in 1995, 4¾ years old and had had about four years of training on interspecies communication; much of his instruction had, however, involved unsuccessful video and audio exposure and his vocabulary was limited to a few object labels (i.e., parrots do not learn referential labels if training is via video or audiotapes [13,14,15]). He was removed from the laboratory in 2001, and did not participate in further experiments, such as the imitation study. At the time of most of the trials, in 2003, Alex was 27 years old and had been the subject of experiments on avian cognition and interspecies communication for 26 years [16]; his training involved human modeling [17] and his vocabulary included labels for over 50 different objects, seven colors, five shapes, numbers up to 6, three categories, and many functional phrases (e.g., "I want X", "Wanna go Y", "Come here", "Go pick up X", etc.). He had had one insight trial in 1995. In 2003, Griffin was 8 years old; he had been obtained from a breeder when 7½ weeks old and had been the subject of studies similar to those involving Alex; his unsuccessful video training experiments, unlike those of Kyaaro, had been limited to only a few labels (e.g., [15]). His vocabulary, although not as extensive as Alex's, therefore contained most of the same commands and functional phrases. He had also been the subject of a study on the simultaneous development of object and label combinations [18]. Arthur (aka Wart) was about 4½ years old in 2003; he had been obtained from his hand-feeder when he was about a year old, but most of his training had involved studies on animal-human computer interfaces [19], and his vocalizations were limited to just a few labels [14]; he could state "want some" when a trainer had something he desired, but could not specify an item for trainers to retrieve. Alex, Griffin, and Arthur participated in one imitation trial in 2003 and in additional imitation trials in 2006.

2.2 Apparatus

As reported in [1], birds were tested on parrot "T" stands. For Alex, Griffin, and Arthur, almonds (a favorite food) or pieces of blackboard chalk (a less favored item, but one with which they had interacted in the past) were suspended at the end of 60 cm long chains of plastic links hung from the end of a "T" stand; red or green oval links were ~ 2.5 cm long and 1.2 cm wide, blue triangular links were ~ 2.5 cm long and 2.5 cm at their widest. These chains would provide the birds with adequate purchase if they attempted to obtain the suspended items. For Kyaaro, a favorite bell was suspended at one end of the stand from a silken cord approximately 0.6 cm in thickness and about 60 cm in length; he could not chew through such cords and did not exhibit any fear of them but was afraid of the plastic chains we used for the other birds.

2.3 Procedure

Again as reported in [1], birds were initially examined individually on the test for insight. Each bird was placed on the "T" stand after a targeted item was suspended; trainers then pointed to the object. If a bird did not seem interested at first, it was told to "Pick up the nut/bell/chalk/treat". (All our birds respond to such commands if there is a single choice on, e.g., a tray; given multiple choices, they take their favorite item [16].) Birds were given several minutes in which to attempt the procedure; if they did not succeed, make an attempt, or demonstrate interest within 5 min, the trial was ended. Each bird except Alex was given three trials in its first session; Alex had only one trial in 1995. Two weeks after the first 2003 trials, Arthur was given a simultaneous choice between a nut hung from one chain (red, oval links) and chalk from another (blue, triangular links). Two months later (the delay was to avoid the possibility of training or massed-trial learning) Alex, Griffin, and Arthur were then given three more trials involving a single chain, with the less desirable item (a piece of chalk) suspended in the first two of the three trials, and a nut in the third. The intention was to see whether the type of reward affected their behaviour and if they could spontaneously solve the problem, not if they could learn to obtain their treat.

Note that none of the birds had received any training on this task prior to testing. Any toys hanging in their cages were suspended on short metal chains (at most, 7.5 cm long) such that each toy was at approximately beak level when birds were perched; thus they would not have been able to practice the maneuver. Only Arthur had had a toy suspended from a perch by a long chain (~30 cm) prior to testing, and it was his demonstration of the targeted behaviour as soon as the toy had been suspended that prompted formal testing.

After Alex and Griffin failed and Arthur succeeded on this test of insight (see **RESULTS**, full details are in [1]), the former two birds were allowed to watch Arthur once in 2003 and six times (twice each day for three non-consecutive days) in 2006. Each time, an almond was suspended from a plastic oval link chain at the end of Arthur's "T" stand as during the insight trials. Alex and Griffin were placed on their own "T" stands, less than 1 m from Arthur's stand (out of reach but in clear view), and then Arthur was placed on his stand and allowed to retrieve the nut. Alex's and Griffin's interest in the nut ensured that they observed Arthur's actions.

3 RESULTS

3.1 Insight trials

3.1.1 Kyaaro and Arthur

On their first exposures [1], both Arthur and Kyaaro immediately performed the targeted action of pulling, stepping, and repeating the behaviour so as to obtain the desired treat; they repeated their actions correctly each time without any hesitation for a total of three trials. For both birds, the actions were not necessarily performed smoothly (occasionally they had to make more than one attempt at grasping the chain or cord that sometimes began to swing as the trial progressed), but they acted consistently and with perseverance. (See video S1 in [1] for one of Arthur's trials.) Kyaaro had no more trials.

On the choice trials ([1], between nut and chalk), Arthur first performed the series of manipulations to obtain and eat the nut (i.e., chose the chain with the nut first), then repeated the manipulations with the chain holding the chalk. He dropped the chalk immediately after obtaining it.

On the final set of three trials [1], Arthur successfully performed the operations to obtain the chalk both times; he quickly discarded the chalk after extracting it from the chain. He also succeeded with the nut, which he dropped, seemingly by accident.

3.1.2 Alex and Griffin

On their first trials [1], neither Alex nor Griffin made any attempts at recovering the nuts. In Alex's only trial in 1995 and in his subsequent trials in 2003, he, like Griffin in 2003, looked at the nuts, looked at the trainer, and said "Want nut". To the trainer's command "Go pick up nut", they both replied "Want nut"; this verbal interplay was repeated several times during each trial. (See video S2 in [1] for part of one of Alex's trials.) In Alex's case, the volume and intensity of the request increased in one trial with the trainer's failure to comply.

In their final three trials [1], Alex and Griffin both completely ignored the chalk and, interestingly, then also ignored the nut; that is, they made no requests for either object nor did they engage in any action required to obtain either object.

3.2 Imitation trials

After Alex's and Griffin's first failure in 2003, they observed one of Arthur's successful trials, but their behaviour did not change; that is, they consistently requested the nut from the trainer and failed to make any attempts themselves [1]. The single-trial session was ended for both Alex and Griffin without their having succeeded.

In 2006, both birds were again given opportunities to observe Arthur's successful trials; they again failed to engage in any form of observational learning. On the first two trials (session one), they watched and requested the nuts vocally; on the next two trials (session two, held two days later), they watched but Alex did not make requests while Griffin continued to do so, and on the final two trials (session three, held about four weeks later), they again watched and both requested the nuts.

4 DISCUSSION

The results of these studies have implications for evaluating the effects of prior training on both insight and imitation. Detailed discussion of how training affects insightful behavior can be found in the original article [1], which I will review only briefly. I will concentrate instead on the effects of previous training on imitation and compare the results with data on other avian species.

In terms of insight, the noteworthy result of the prior experiment [1] was that the two parrots with limited vocabularies immediately acted out the correct physical tasks to obtain their treats, whereas the parrots that had received considerably more effective training in referential English speech attempted instead to manipulate their trainer. These birds' requesting behaviour appeared intentional: They were asking that trainers do something for them, in very specific, fairly stress-free circumstances and in a very direct manner [1]. They were not treating humans as a physical object to be used (e.g., as a stepping-stone to reach something desired; see [20]), but were engaging in deliberate communication as a problem-solving strategy, which is a fairly advanced stage of development, even for human infants (see [20,21]). These birds acted just as they do when they want other treats that are not within their reach (e.g., [16]), thereby cross-applying (transferring) behaviour patterns learned in one situation to another, which is also considered a hallmark of intelligent behaviour [22]. According to some researchers (e.g., [23]), the adaptive value of using a referential communication code to benefit oneself at the expense of others is viewed as an advanced, essentially human trait.

Why Alex and Griffin did not continue to request the suspended nuts in their final trials in 2003 nor Alex in his middle imitation trials in 2006 was not clear. They possibly had learned in previous trials that no trainer would assist them and that requesting the nut was useless (a form of learned helplessness [24]); other reasons for their lack of action (including string-pulling) are discussed in detail in [1].

Here, however, I wish to focus on the birds' lack of physical imitative ability. Why were Griffin and Alex unable—or at least unwilling—to reproduce an observed behaviour to acquire a desired treat? Several issues are of note.

First, my Grey parrots have shown, over the course of almost 30 years of study, a facility for accurately reproducing English speech. Moreover, these referential vocal abilities all derive from a social learning paradigm [16,17], thereby demonstrating the parrots' competence for observational learning. Although they use a different vocal apparatus than humans to produce speech, in many cases their articulatory acts would indeed seem to qualify as imitation [1,25,26,27]; such data suggest that some form of imitation is within the purview of the Grey parrot.

Second, a Grey parrot in different laboratory has been shown to reproduce human physical actions, such as waving a foot after seeing a human wave his hand [3]. Although the extent to which such behaviour patterns are novel and would fit Thorpe's definition of imitation [6] is unclear, the capacity of the bird in question to integrate observed physical actions into its behavioural repertoire suggests that this ability is also within the purview of the Grey parrot.

Third, in species such as goldfinches (*Carduelis carduelis*) and siskins (*Carduelis spinus*), not only do only a percentage of tested birds (23% of 52 the former, 62% of 29 of the latter) solve the string-pulling problem, but only another small percentage (25% of the former species, 10%

of the latter) who fail by themselves achieve any form of success after observing successful birds [11]. Too, those who achieve success via observation often did so by emulation—achieving the goal by a different method—rather than by imitating the actions of the demonstrators [11]. Most of those birds that consistently failed, even after being exposed to a demonstrator, did not fail because of lack of observational experience [11]. Assuming that such behaviour can be extrapolated to parrots—a likely supposition given the work of Huber and his colleagues [12,28,29], which demonstrated considerable individual differences and various levels of imitation and emulation in keas—birds (including parrots) likely exhibit individual differences in their ability or motivation to reproduce observed actions.

Given that Grey parrots have demonstrated competence in what appear to be related tasks of observational learning, I suggest that, whatever individual differences might exist between Alex, Griffin, and Arthur, that Alex's and Griffin's failure to reproduce Arthur's actions in the string-pulling task was a consequence of their previous training that emphasized the vocal mode and a paradigm in which humans would diligently respond to their vocal requests. Such training may have reduced their motivation to act (physically) on their own. Granted, neither Alex nor Griffin had had significant experience in the kind of physical manipulations (e.g., pulling at branches or twigs to obtain food) that might not only engender string-pulling but might also potentiate imitation of related physical actions [11], but neither had Kyaaro nor Arthur had such experience, and all birds had been given numerous objects that they could chew or tear apart, pick up or toss with foot or beak. Interestingly, Alex and Griffin, in contrast to Kyaaro and Arthur, were given tasks in which covers needed to be removed to expose hidden objects (e.g., [30] and references therein); Griffin also had demonstrated some proficiency in combining objects [18]. Note, however, that all actions were done with their beaks. Possibly, as was suggested in [1], for Alex and Griffin, successful vocal training may have caused communication (or at least beak-related) areas in the brain to develop to the detriment of those used to control complex, sequential physical actions involving both limbs and beaks (Heinrich, pers. comm.). The string-pulling task involves eye-foot-beak coordination and thus may have required brain areas in addition to those involved in solely beak-driven combinations such as stacking or removing cups and vocalizing. If true, this explanation does not detract from the complexity of the vocal behaviour, but rather provides a rationale for the Alex's and Griffin's vocal rather than physical actions.

Another issue might be the dominance hierarchy of the birds in the laboratory—would Alex and Griffin be willing to reproduce the behaviour of an individual in a position clearly subordinate to theirs? Arthur is the youngest, most recent addition to the lab, and by default the lowest ranking bird. One might imagine that having humans demonstrate the targeted string-pulling behaviour pattern, as they do with vocal patterns, might be preferable, but that option (hand-over-hand, or even an attempted mouth-hand demonstration) would not allow the birds to see how *they* might perform the task and could even be viewed by the birds as an acquiescence to their demands, not as a demonstration. (One such human demonstration, performed just before the writing of this manuscript, engendered the not-unexpected request for the retrieved nut.) Arguably, Alex's and Griffin's demands that the trainers do the task might be taken as evidence that they consider themselves dominant to the humans in the laboratory; clearly, humans do spend as much time acceding to their demands as querying and thus making demands of them. Consistent with such a view is the possibility that Arthur, subordinate to the other birds, might also be seen as subordinate to humans because he cannot ask trainers to carry out his demands and, thus, was unworthy of imitating.

I suspect that, in order to demonstrate that Alex and Griffin could engage in either insightful behaviour or a form of imitation that involves object manipulation, I would have to devise a task that would be intriguing and motivating enough to spontaneously override their prior training. For obvious reasons (continuing experiments on vocal learning and cognitive processing), extinguishing their previous training is not an option, and the parrots' overt distress upon the exit of trainers [1] precludes at present carrying out the study (Bugnyar, pers. comm.) involving videotaping Alex and Griffin in the absence of human observers.

5 CONCLUSION

In sum, two parrots that had limited use of vocal requests exhibited behaviour similar to the insightful food retrieval displays of, for example, Heinrich's ravens [2] and Funk's kakarikis [9]; the two parrots who could make specific vocal requests did so instead, and continued to do so even after observing the successful retrieval by another parrot. Such data emphasize three points. First, that the entire behavioural repertoire and history must be examined in studies that try to determine whether animals act insightfully or are capable of imitation; second, that parrots can attempt to direct humans to assist them in obtaining their goals; and third, that such behaviour—although clearly complex—might lead them to fail certain tasks.

ACKNOWLEDGMENTS

I thank the many donors to *The Alex Foundation* for their support of this research, particularly the Pearl Family Foundation, Janice Boyd, Kathryn and Walter McAdams, Janet Trumbule, John Paton, Nancy Clark and Bill Broach, Megumi Oka and the Makioka Foundation, Michael and Deborah Smith, Alex and Michael Shuman, Will Lipton, Greg LaMorte, Nancy Chambers, Carl and Leigh Ann Hartsfield, and Deborah and Robert Goodale.

REFERENCES

[1] I.M. Pepperberg, '"Insightful" string-pulling in Grey parrots (*Psittacus erithacus*) is affected by vocal competence', *Animal Cognition*, **6**, 263-266, (2004).

[2] B. Heinrich, 'An experimental investigation of insight in Common Ravens (*Corvus corax*)', *Auk*, **112**, 994-1003, (1995).

[3] B. Moore, 'Avian movement imitation and a new form of

mimicry: Tracing the evolution of a complex form of learning', *Behaviour*, **122**, 231-263, (1992).

[4] H. Gardner, *Multiple Intelligences: New Horizons.* Perseus Group, Basic Books, New York, 2006.

[5] I.M. Pepperberg, Evolution of avian intelligence. In *The Evolution of Intelligence*, R. Sternberg and J. Kaufman, eds., Erlbaum, Mahwah, NJ, 2001.

[6] W.H. Thorpe, *Learning and Instinct in Animals*, 2nd Ed. Harvard University Press, Cambridge, MA, 1963.

[7] M. Tomasello, *The Cultural Origins of Human Cognition*, Harvard University Press Cambridge, MA, 1999.

[8] P. Willatts, 'Development of means-end behavior in young infants: Pulling a support to retrieve a distant object', *Developmental Psychology*, **35**, 651-667, (1999).

[9] M.S. Funk, 'Problem solving skills in young yellow-crowned parakeets (*Cyanoramphus auriceps*)', *Animal Cognition*, **5**, 167-176, (2002).

[10] M.Vince, 'String pulling in birds. III. The successful response in greenfinches and canaries', *Behaviour*, **17**, 103-129, (1961).

[11] U. Seibt and W. Wickler, 'Individuality in problem solving: String pulling in two *Carduelis* species (Aves: Passerformes)', *Ethology*, **112**, 493-502, (2006).

[12] D. Werdenich and L. Huber, 'A case of quick problem solving in birds: String pulling in keas, *Nestor notabilis*', *Animal Behaviour*, **71**, 855-863, (2006).

[13] I.M. Pepperberg, 'Vocal learning in African Grey parrots: effects of social interaction', *Auk*, **111**, 300-313 (1994).

[14] I.M. Pepperberg and S.R. Wilkes, 'Lack of referential vocal learning from LCD video by Grey Parrots *(Psittacus erithacus)*', *Interaction Studies*, **5**, 75-97 (2004).

[15] I.M. Pepperberg, L.I. Gardiner, and L.J. Luttrell, 'Limited contextual vocal learning in the Grey parrot (*Psittacus erithacus*): the effect of co-viewers on videotaped instruction', *Journal of Comparative Psychology*, **113**, 158-172, (1999).

[16] I.M. Pepperberg, *The Alex Studies*, Harvard University Press, Cambridge, MA, 1999.

[17] I.M. Pepperberg, 'Functional vocalizations by an African Grey parrot (*Psittacus erithacus*)', *Zeitshrift für Tierpsychologie*, **55**, 139-160, (1981).

[18] I.M. Pepperberg and H. Shive, 'Simultaneous development of vocal and physical object combinations by a Grey Parrot (*Psittacus erithacus*): Bottle caps, lids, and labels', *Journal of Comparative Psychology*, **115**, 376-384, (2001).

[19] I.M. Pepperberg, *The Wired Kingdom*. Conference at the MIT Media Lab, April. 2000.

[20] J.C. Gómez, The emergence of intentional communication as a problem-solving strategy in the gorilla. In *"Language" and Intelligence in Monkeys and Apes: Comparative Developmental Perspectives*, S.T. Parker and K.R. Gibson eds., Cambridge University Press, New York, 1990.

[21] R. Case, *Intellectual Development: Birth to Adulthood.* Academic Press, Orlando, FL, 1984.

[22] P. Rozin, The evolution of intelligence and access to the cognitive unconscious. In *Progress in Psychobiology and Physiological Psychology*, (Vol. 6), J.M. Sprague and A.N. Epstein, eds., Academic Press, NY, 1976.

[23] D. Kemmerer, 'What about the increasing adaptive value of manipulative language use?', *Behavioral & Brain Sciences*, **19**, 546-548, (1996).

[24] M.E. Seligman and S.F. Maier, 'Failure to escape traumatic shock', *Journal of Experimental Psychology*, **74**, 1-9, (1967).

[25] D.K. Patterson and I.M. Pepperberg, 'A comparative study of human and Grey parrot phonation: Acoustic and articulatory correlates of stop consonants', *Journal of the Acoustical Society of America*, **103**, 2197-2213, (1998).

[26] D.K. Warren, D.K. Patterson, and I.M. Pepperberg, 'Mechanisms of American English vowel production in a Grey Parrot (*Psittacus erithacus*)', *Auk*, **113**, 41-58, (1996).

[27] I.M. Pepperberg, Training behavior by imitation: from parrots to people…to robots. In *Proceedings of the AISB '03 Second International Symposium on Imitation in Animals and Artifacts*, K. Dautenhahn and C. Nehaniv, eds., University of Wales, 2003.

[28] L. Huber, S. Rechberger, and M. Taborsky, 'Social learning affects object exploration and manipulation in keas, *Nestor notabilis*', *Animal Behaviour*, **62**, 945-954, (2001).

[29] G. Gajdon, N. Fijn, and L. Huber, 'Testing social learning in a wild mountain parrot, the kea (*Nestor notabilis*)', *Learning & Behavior*, **32**, 62-71. (2004).

[30] I.M. Pepperberg, M.R. Willner, and L.B. Gravitz, 'Development of Piagetian object permanence in a Grey parrot (*Psittacus erithacus*)', *Journal of Comparative Psychology*, **111**, 63-75, (1997).

Imitative learning in monkeys

Ludwig Huber[1], Bernhard Voelkl[1] and Friederike Range[1]

Abstract[1]. Imitative learning has received high levels of attention due to its supposed role in the development of culture, language and self-identification and the cognitive demands it poses on the individual. Although monkeys possess mirror neurons, show neonatal imitation, recognize when being imitated and copy an expert's use of a rule, their capacity of action imitation has been doubted by most imitation researchers so far. Here I will argue that imitation in the original definition of learning to do an act from seeing it done must be distinguished from other forms of "copying", in which the content of the copy is not the behavior of the model but the result of the demonstrated action, its goal or the intention of the demonstrator. Then I will describe several experiments with captive common marmosets (*Callithrix jacchus*) that show that these monkeys can use the same overall pattern of a technique to open a food box, or the same body part as the model, or – above all – can precisely copy the movement pattern of an action that a skilful model has demonstrated. On the basis of this cumulative evidence of imitation in non-human primates I will question the frequently expressed notion that imitation is a relatively recent invention in the hominoid lineage and will discuss its implications for the currently available theories of the underlying neuronal mechanism.

1 MONKEY SEE, MONKEY DO

According to Byrne [1], imitation research has focused on one of two distinct problems. The one favored by cognitive neuroscientists is the 'correspondence' problem, asking how is it possible for actions as seen to be matched with actions as imitated? The other, favored by ethologists and comparative psychologists, is the 'transfer of skill' problem, asking how is it possible for novel, complex behaviors to be acquired by observation? Despite various approaches to, and definitions of, imitation [2-5], most scholars agree that when an individual replicates an action that it has observed being performed by another individual it requires a matching system that allows conversion of observed actions by others into actions executed by oneself. In other words, visual input needs to be transformed into corresponding motor output. However, most currently available models of imitation require that the observers had possessed a motor representation of the demonstrated action already before they observe it being performed by the model. But how can new skills be acquired if the essence of imitation lies in the activation or facilitation of responses already in the repertoire of the observer? Imitative learning in the sense of the acquisition of new skills by observation must therefore be distinguished from response facilitation [6], priming, stimulus enhancement and other forms perception-motor coupling, let alone many forms of social influences [7-10].

A common conclusion about social learning among primates was that apes imitate in various forms [11], but that monkeys, despite a century's efforts, had not been shown to imitate [7, 12-14]. Although the sweet potato washing of Japanese macaques on Koshima Islet is perhaps the best known and most frequently cited example of the formation of traditions in nonhuman animals [15, 16], it has been questioned whether social learning, let alone imitation, is really involved [17]. Furthermore, Visalberghi and Fragaszy have made several attempts to find out whether Capuchin monkeys learn by observation of a skilful model how to use an object as a tool [13, 14]. However, all these attempts failed.

Recently, the picture of the monkeys' failure to imitate has been seriously doubted, because macaques show cognitive imitation by copying an expert's use of a rule [18], recognize when they are being imitated [19] and imitate adult facial movements as neonates [20]. Also, the discovery that rhesus monkeys have "mirror neurons"—neurons that fire both when monkeys watch another animal perform a goal-directed action and when they perform the same action [21-23] —suggests they possess the neural framework for perception and action that is associated with imitation. However, can monkeys also imitate by "learning to do an act from seeing it done" [24], restricted to the acquiring behaviors novel to the individual's repertoire (the 'transfer of skill' problem)? It has been suggested that in order for a response to be considered acquired through imitation it must be novel [25]. The behavior can be thought of as novel if the probability of the behavior is low at the start of the experiment and an increase in the behavior cannot be attributed to priming, motivational or attentional effects [10, 26].

2 IMITATION IN COMMON MARMOSETS

We have focused on one species of New World monkeys of the family *Callithrichidae*, the common marmoset (*Callithrix jacchus*). Callitrichid monkeys are small monkeys once thought to have retained many primitive primate characters and to be rather unsophisticated [27]. Therefore, marmosets and tamarins would not seem likely candidates for studies of complex cognition. However, this evaluation has changed [for a review, see 28], and it is currently accepted that they have developed a number of remarkably original adaptations for their unusual lifestyle [29]. More than this, Callitrichids are likely to locate food by using some sort of cognitive map [30], represent objects and their movements in an abstract manner [31], and benefit from social influences that aid in learning about new food by motivational and perceptual factors [32]. Marmosets and tamarins are remarkably sensitive and responsive to cues from other social companions, especially in the third and fourth month of life [33]. Their high level of tolerance

[1] University of Vienna, Vienna, Austria, email: ludwig.huber@univie.ac.at

during group foraging and also their sharing of food in both passive and active manners [34, 35] may be related to their cooperative breeding system [29]. All together, these social features may be responsible for their high degree of maintaining spatial and behavioral cohesion with their social partners in comparison to Capuchin monkeys (*Cebus* sp.). They are also more neophobic than capuchins, less likely to explore new places and, therefore, less likely to explore new foods or to solve new manipulative problems on their own [36].

These behavioral and social aspects of callithrichids inspired a number of experimental studies focusing on the ability of common marmosets to learn from conspecific demonstrators an food-processing technique [37-39]. All three studies used variants of the same experimental procedure (non-observer control): First, subjects (observers) were allowed to observe a physically separated conspecific (demonstrator) opening a novel apparatus – the "artificial fruit" [40] – in order to retrieve food from it, and thereafter these subjects themselves were given the opportunity to manipulate the apparatus on their own. The subjects' behavior was then compared with naïve animals that were confronted with the apparatus without prior observation of conspecifics (non-observers), and – in the Voelkl and Huber study [39] – also with observers which saw another demonstrator opening the apparatus in a different way (two-action procedure).

In the first study of this kind, Bugnyar and Huber [37] presented common marmosets a box with a pendulum door that could be either pushed or pulled to gain access to food inside the box. Observers were allowed to watch a conspecific demonstrator pull open the door. The observers showed less exploratory behavior than non-observers and, most importantly, two of them showed a strong tendency to use the demonstrated opening technique in the initial phase of the test. Only after some trials, in which they acquired own experience of opening the pendulum door, did they begin to perform the simpler solution of pushing, which was preferred by the non-observers. The authors argued that pulling the door in order to get access to food was not a simple act but a compound action-pattern. The authors distinguished four independent elements plus one dependent element in the pulling performance of the demonstrator: (1) using the left hand, (2) taking the door from the right gap, (3) pulling, (4) holding the door wide open with one hand, and (5) taking the food. Two observer marmosets copied all of these actions in the appropriate order, which is very unlikely to be due to chance, considering the combined probability for spontaneous occurrence of these actions.

In an attempt to provide data allowing a direct comparison between species, Caldwell and Whiten [38] used a marmoset-sized version of an artificial fruit that has been designed for studies of imitation in children and chimpanzees [41]. One demonstrator ('full' demonstrator) was trained to open the apparatus by removing a handle, while the other demonstrator ('partial' demonstrator) simply ate food from the lid of the apparatus. Unfortunately, none of the observers was successful in opening the apparatus – probably due to the technical sophistication of the opening mechanisms. However, the authors found clear response differences consistent with the different demonstration modes. Those animals that watched the 'partial' demonstrator performing predominantly mouthing behaviors used their mouth more frequently, while those that watched the 'full' demonstrator showing predominantly hand manipulation used their hands more frequently. The authors described these findings as body part copying, but they pointed out that the behavior of the observers might have been dependent on several other social learning effects as well. For instance, it may be the case that reaching or grasping behaviors are in some way contagious (i.e., triggered by the same response) in marmosets or that the fact that the movement of the apparatus was clearly different for both observer groups could account for the social learning seen.

Only a *two-action method*, which involves two demonstrators that differ in their body movements but create the same changes in the environment, controls for learning about the changes of state in the environment and therefore provides the most convincing evidence yet for imitative learning in animals [10, 42, 43]. Voelkl and Huber [39] applied this methodology, permitting two groups of marmosets to observe a demonstrator using one of two alternative techniques to remove the lids of baited film canisters and compared their initial test responses with one another and with a third group of marmosets that were never given the opportunity to observe a demonstrator. Furthermore, while one technique involved hand-opening common to marmosets, the other technique consisted of a behavioral 'peculiarity' (mouth-opening); that is, mouth opening was neither common in the animals under investigation nor necessary for lid removal. This requirement ensured that if the observers performed the technique, then they were most probably influenced by what they witnessed.

In fact, both groups of observers preferred to open the canisters using the same method as their demonstrator. Since hand and mouth demonstrators brought about identical changes to the canister (opening and exposing the food reward), the differential test behavior of the animals suggests that they indeed learned something about the demonstrator's behavior, rather than about certain properties of the canister. Furthermore, non-observers rarely opened the canister with the mouth, but they opened as many canisters as did members of both observer groups. An actual benefit to observer animals in terms of success rate could be found when the task was made more difficult by closing the lids of the canisters much more firmly. After this change, only the mouth-openers achieved opening the canisters and retrieving the desired mealworms. Thus, even 'slavish' copying (i.e., copying in the absence of insight) may therefore have beneficial effects for observers [26]. Furthermore, as emphasized by Caldwell and Whiten [38], social learning may provide particular practical benefits to individuals when it induces an individual to persist with unrewarded manipulations of an object, as individual learning (trial and error) is unlikely to be successful under such circumstances.

Although this study implies that learning by imitation is more widespread in the animal kingdom than recently assumed, both the studies that failed to demonstrate imitation in monkeys as well as our own experiences in previous studies suggests that imitation is a rarely used mechanism. This might be due to the special requirements of imitative learning. To allow detailed observations imitation requires proximity of the individuals. Thus imitation is more likely to be found in species with egalitarian social systems where the individual distances between all group members are small. Additionally the copied behavior itself must show certain characteristics to be appropriate for transmission by imitation. If the behavior in question is to simple it is more likely that it is learned by individual learning, while if the behavior is on the other hand to complex it is unlikely that it can be learned by observational learning at all. A further limiting factor in imitative learning – but widely neglected in discussions and experiments until today – is the attention of the observer. If the observer does not pay attention to the whole action or action sequence demonstrated, but looses interest before it is completed, it will not be able to learn the action sufficiently well or at all. The attention-holding process might vary according to the dominance, sex and relationship of the demonstrator, as well as the type of action demonstrated, and the vigilance, interest, and motivation of the observer. While we found generally quite short attention spans in marmosets, they are sufficiently long for the actions demonstrated in our experiments [44]. However, in the study with the film canisters, the observers have been found to be particularly attentive.

3 ACTION IMITATION

Recent theories of imitation have dissected the imitative act into two components, the body part used and the action performed [45]. With respect to *body part imitation*, the finding that marmosets would copy mouth versus hand use [39] and pigeons likewise would copy beak versus food use [46] was important in classing these as true imitation [9-11, 42, 47, 48]. Use of different body parts to deal with the same task relies on visuomotor mapping from seen parts of the model's body to equivalent parts of the self. But for an action to qualify as imitation in the restricted sense of "learning how to move" (*action imitation*), the observer must learn the specific response topography, i.e., the specific action by which the response is made [9]. Despite various approaches to, and definitions of, imitation [2-4], most scholars agree that when an individual replicates an action that it has observed being performed by another individual it requires a matching system that allows conversion of observed actions by others into actions executed by oneself. In other words, visual input needs to be transformed into corresponding motor output (the 'correspondence' problem).

The greatest challenge for an animal solving the correspondence problem is to perform imitation of 'perceptually opaque' actions, those model actions of which the observer's image is not similar to the sensory feedback received during performance of the same action [42]. This is particularly true if the action demonstrated by the model does not already exist in the observer's behavioral repertoire. So far, precise copying of novel actions is underspecified in theory and vague in evidence. The models currently available, including those that rest on mirror neurons, are not sufficiently competent to explain high matching fidelity in the imitation of *novel* actions, thereby solving both problems of imitation, the transfer of skill and the correspondence problem, at the same time. There is also no convincing evidence of movement copying in nonhuman animals with the trajectory of the movement or the dynamics of the model's action being replicated by the observer with high fidelity. The few cases in which animals have been reported as achieving some degree of matching are lacking rigorous quantitative analysis of the matching degree (e.g. only qualitative descriptions of the imitator's performance are provided [49, 50].

A paradigm that has come closest to the assessment of the precision of copying is the "do-as-I-do paradigm". I a replication of the classic study by Hayes and Hayes [51], Custance and colleagues [52] found only a modest degree of matching between tutor and subject. Coders blind to what each chimpanzee has actually watched identified some matching in relation to touching several parts of the body in sight, as well as out of sight, symmetric and asymmetric conjunctions of hands, digit movements, and hand or whole-body actions. However, the matching fidelity was low overall; only a small fraction of the novel actions were reproduced (13 or 20 from a total of 48 novel gestures), and the imitations were far from perfect. Similarly, in a study by Myowa-Yamakoshi [53], five female nursery-reared chimpanzees rarely reproduced a demonstrated action at the first attempt (less than 6% of the overall actions). In a further, more recent study, only 20% of the chimpanzee observers matched the demonstrator's actions, i.e., opening a tube with the hands [54].

4 THE PRECISE COPYING OF MOVEMENT TRAJECTORIES

To investigate imitation of movement patterns in marmoset monkeys we reanalyzed the actions shown by the mouth model and her six observers of the Voelkl and Huber [39] study and – for the sake of comparison – tested further 24 naïve animals (non-observers) that had never observed a model. In order to video-capture the movements from a fixed perspective, only one baited canister was attached to a wooden board that was placed in a six cm gap between a glass window in the front-side of the testing cage and an opaque partition wall. This setup ensured that animals could approach the artificial fruit from only two directions – in both cases approximately parallel to the window and the lens of the video camera. The completely shut canister required – due to the tightness of the lid – a powerful opening technique. Our marmosets have never achieved to open a completely shut canister by hand but only by mouth.

The model, five out of six observers, but only four out of 24 non-observers succeeded in opening the containers with their mouth (Voelkl & Huber, in prep.). These opening attempts provided six opening movements of the model, 14 of observers and 21 of non-observers. The head movement of the subjects was tracked by manually identifying the position of five morphological features in the face of the subject on a frame-to-frame basis (25 frames per s). We then defined five parameters to describe the movement, used discriminant function analysis of the orthogonalized data, and thus generated a function with clearly distinctive mean discriminant scores for movements of the model and the non-observers. As main result, we found the mean scores for the observers being closer to the mean of the model than to the non-observer. Thirteen out of 14 observer movements were classified as model movements. Thus, the movement patterns of the observers were clearly more similar to the movement pattern of the model than to those of the non-observers.

5 HOW DO BRAINS SOLVE THE IMITATIVE LEARNING PROBLEM?

More than a century of research on imitation has left us with a crucial functional problem: how are observers able to transform a visual presentation of an action into motor output? For most currently available theories of imitation the key to solution is automatic activation of *existing* motor representations. But here marmosets learned by observation a novel movement pattern, not available from the own behavior repertoire. Even if we assume that marmosets have already performed similar movements before, like biting into an object or levering it up with the head, how can we explain the exact matching of the observers, e.g. the convergence of the paths of the head, the same inclination of the head in the course of the opening action? A minimal requirement would be to adjust an action present in one's motor repertoire to a different observed action [55].

As evidenced by the significant difference in the movement shown by observers and non-observers, opening a film box is not an all-or-nothing behavior for marmosets. There are still many degrees of freedom for the exact performance, created by the movements of the head and the whole body when attempting to open the lid of the film canister. The common problem for imitation theories is to account for the close convergence of Bianca's (the model) and the observers' opening actions despite the actual variance of ways to achieve the common goal of opening the lid as evidenced by the non-observers. Which of the many theories currently available in the literature can offer a sufficient explanation of the creation of a novel action from using only visual information? Or more generally, what does this result tell us about what is actually involved in successful action imitation, i.e., learning how to move the body by observing the behavior of others? And what role do the mirror neurons play?

Mirror neurons might code the likely future actions of others so that observers are able to anticipate the intention of others [56] rather than to provide a form of motor learning. Macaques, for instance, might have used their mirror neurons to recognize being imitated [19]. However, the actions that they were shown by the human model have already existed in the observers' motor repertoire. Perhaps a first step in the direction of clarifying the potential role of mirror neurons for imitative learning might be the detection of a new type of visuomotor neurons, called *tool-responding mirror neurons*, in the lateral section of the macaque monkey's ventral premotor area F5 [57]. The neurons show experience-dependent responses and perhaps enable the observing monkey to extend action-understanding capacity to actions that do not strictly correspond to its motor representations. However, in contrast to our findings, these neurons were found to discharge only after a relatively long visual exposure to actions of a tool-using experimenter. It was therefore proposed that the changes in the body schema and/or in the motor representations of the observer are possible only for motor training [58], but that tool actions cannot be directly translated into own motor repertoire. The authors concluded with hypothesizing that a mirror mechanism evolved in monkeys for action understanding, but only emerged in human evolution as suitable neural substrate for imitation [57].

Recently, theoretical and empirical attempts have been made to explain imitative learning through reafferent feedback loops in the brain. As part of a conceptual framework for motor learning and sensorimotor control, the 'modular selection and identification for control model' (MOSAIC) is based on multiple pairs of 'predictor' and 'controller' models processing feedforward and feedback sensorimotor information, respectively [59-61]. Indeed, the MOSAIC model has been shown to learn a simple acrobat task (swinging up a jointed stick to the vertical) through action observation and imitation [62]. The results of functional magnetic resonance experiments suggested the superior temporal sulcus (STS) as the region at which the observed actions, and the reafferent motor-related copies of actions made by the imitator, interact [63]. Furthermore, in the macaque there seems to be a circuitry composed of the STS, providing a higher-order visual description of the observed action, the rostral sector of the inferior parietal lobule (PF) and the ventral premotor cortex (area F5) that codes the action of others and maps it onto the motor repertoire of the observer [64]. Thus, imitative learning is supported by interaction of the core circuitry of imitation with the dorsolateral prefrontal cortex and perhaps motor preparation areas — namely, the mesial frontal, dorsal premotor and superior parietal areas. In humans, this direct route of visuo-motor conversion on a sensory-motor level of imitation is used especially for transforming a novel or meaningless action into a motor output, while a semantic mechanism, working on the basis of stored memories, allows reproductions of known actions on an intentional level of processing [65-67]. It remains to be shown whether non-human animals can also use multiple routes of action imitation.

In conclusion, the present findings suggest that monkeys are not only able to reproduce known actions shown by others or to recognize when others reproduce actions they themselves have executed before, but are also able to learn new actions by observation. Such abilities are functional by providing a type of learning that avoids remaining with insufficient or ineffective variants or time-consuming trial-and-error learning.

ACKNOWLEDGEMENTS

This work has received funding from the Austrian Science Fund (P12011-BIO) and the European Community (NEST 012929).

REFERENCES

1. Byrne, R.W., *Imitation as behaviour parsing.* Philos Trans R Soc Lond B Biol Sci, 2003. **358**(1431): p. 529-36.
2. Heyes, C.M. and B.G.J. Galef, eds. *Social Learning in Animals: The Roots of Culture*. 1996, Academic Press: San Diego.
3. Meltzoff, A.N. and W. Prinz, eds. *The imitative mind: development, evolution, and brain bases*. 2002, Cambridge University Press: Cambridge, UK. 353.
4. Hurley, S. and N. Chater, eds. *Perspectives on Imitation: From Neuroscience to Social Science - Volume 1: Mechanisms of Imitation and Imitation in Animals*. 2005, MIT Press: Cambridge, MA.
5. Zentall, T.R. and B.G. Galef, Jr., eds. *Social learning: Psychological and biological perspectives*. 1988, Erlbaum: Hillsdale, New Jersey.
6. Byrne, R.W., *The evolution of intelligence*, in *Behaviour and Evolution*, P.J.B. Slater and T.R. Halliday, Editors. 1994, Cambridge University Press: Cambridge. p. 223-264.
7. Byrne, R.W. and A.E. Russon, *Learning by imitation: a hierarchical approach.* Behavioral and Brain Sciences, 1998. **21**(5): p. 667-84; discussion 684-721.
8. Heyes, C.M., *Social learning in animals: categories and mechanisms.* Biological Reviews, 1994. **69**: p. 207-231.
9. Zentall, T., *Imitation by animals: how do they do it?* Current Directions in Psychological Science, 2003. **12**(3): p. 91-95.
10. Zentall, T.R., *Imitation in animals: evidence, function, and mechanisms.* Cybernetics and Systems: An International Journal, 2001. **32**: p. 53-96.
11. Whiten, A., et al., *How do apes ape?* Learning and Behavior, 2004. **32**(1): p. 36-52.
12. Tomasello, M. and J. Call, *Primate cognition*. 1997, Oxford: Oxford University Press.
13. Visalberghi, E. and D.M. Fragaszy, *Do monkeys ape?*, in *"Language" and intelligence in monkeys and apes. Comparative developmental perspectives*, S.T. Parker and K.R. Gibson, Editors. 1990, Cambridge University Press: Cambridge. p. 247-273.
14. Visalberghi, E. and D.M. Fragaszy, *''Do monkeys ape?'' Ten years after*, in *Imitation in animals and artefacts*, K. Dautenhahn and C. Nehaniv, Editors. 2002, MIT Press: Cambridge. p. 471–499.
15. Kawai, M., *Newly-acquired pre-cultural behaviour of the natural troop of Japanese monkeys in Koshima Islet.* Primates, 1965. **6**: p. 1-30.
16. Hirata, S., K. Watanabe, and M. Kawai, *"Sweet-potato washing" revisited*, in *Primate origins of human cognition and behavior*, T. Matsuzawa, Editor. 2001, Springer: Tokyo. p. 487-508.
17. Galef, B.G.J., *The question of animal culture.* Human Nature, 1992. **3**: p. 157-178.
18. Subiaul, F., et al., *Cognitive imitation in rhesus macaques.* Science, 2004. **305**(5682): p. 407-410.
19. Paukner, A., et al., *Macaques (Macaca nemestrina) recognize when they are being imitated.* Biology Letters, 2005. **1**: p. 219-222.
20. Ferrari, P.F., et al., *Neonatal imitation in rhesus macaques.* PLoS Biol, 2006. **4**(9): p. e302.
21. Gallese, V., et al., *Action recognition in the premotor cortex.* Brain, 1996. **119 (Pt 2)**: p. 593-609.
22. Rizzolatti, G., et al., *Premotor cortex and the recognition of motor actions.* Brain Res Cogn Brain Res, 1996. **3**(2): p. 131-41.
23. Rizzolatti, G. and L. Craighero, *The mirror-neuron system.* Annu Rev Neurosci, 2004. **27**: p. 169-92.
24. Thorndike, E.L., *Animal intelligence: An experimental study of the associative processes in animals.* Psychol. Rev. Monogr. Suppl., 1898. **2**: p. 79-.
25. Thorpe, W.H., *Learning and instinct in animals.* 1956, London: Methuen.
26. Huber, L., *Movement imitation as faithful copying in the absence of insight.* Behavioral and Brain Sciences, 1998. **22** (5): p. 694.
27. Hershkovitz, P., *Living New World Monkeys (Platyrrhini), with an Introduction to Primates. Vol. 1.* 1977, Chicago: Chicago University Press.
28. Huber, L. and B. Voelkl, *Social and physical cognition in marmosets and tamarins*, in *The Smallest Anthropoids: The Callimico/Marmoset Radiation*, S.M. Ford, L.C. Davis, and L. Porter, Editors. in press, Springer: New York.
29. Snowdon, C.T., *Social processes in communication and cognition in callitrichid monkeys: a review.* Animal Cognition, 2001. **4**: p. 247-257.
30. Garber, P. and P. Hannon, *Modeling monkeys: a comparison of computer-generated and naturally occurring foraging patterns in two species of Neotropical primates.* International Journal of Primatology, 1993. **14**: p. 827-852.
31. Mendes, N. and L. Huber, *Object Permanence in Common Marmosets (Callithrix jacchus).* Journal of Comparative Psychology, 2004. **118**(1): p. 103-112.
32. Voelkl, B., C. Schrauf, and L. Huber, *Social contact influences the response of infant marmosets towards novel food.* Animal Behaviour, 2006. **72**(2): p. 365-372.
33. Schiel, N. and L. Huber, *Social influences on the development of foraging behavior in free-living common marmosets (Callithrix jacchus).* American Journal of Primatology, 2006. **68**: p. 1-11.
34. Feistner, A.T.C. and E.C. Price, *Food offering in new world primates: two species added.* Folia Primatologica, 1991. **57**: p. 165-168.
35. Ferrari, S.F., *Food transfer in a wild marmoset group.* Folia Primatologica, 1987. **48**: p. 203-206.
36. Fragaszy, D.M. and E. Visalberghi, *Socially biased learning in monkeys.* Learning and Behavior, 2004. **32**: p. 24-35.
37. Bugnyar, T. and L. Huber, *Push or pull: an experimental study on imitation in marmosets.* Animal Behaviour, 1997. **54**(4): p. 817-831.
38. Caldwell, C.A. and A. Whiten, *Testing for social learning and imitation in common marmosets, Callithrix jacchus, using an artificial fruit.* Anim Cogn, 2004. **7**(2): p. 77-85.
39. Voelkl, B. and L. Huber, *True imitation in marmosets.* Animal Behaviour, 2000. **60**(2): p. 195-202.
40. Whiten, A., et al., *Imitative learning of artificial fruit processing in children (Homo sapiens) and chimpanzees (Pan troglodytes).* Journal of Comparative Psychology, 1996. **110**: p. 3-14.
41. Whiten, A., et al., *Imitative learning of artificial fruit processing in children (Homo sapiens) and chimpanzees (Pan troglodytes).* J Comp Psychol, 1996. **110**(1): p. 3-14.
42. Heyes, C.M. and E.D. Ray, *What is the significance of imitation in animals?*, in *Advances in the Study of Behavior*, P.J.B. Slater, et al., Editors. 2000, Academic Press: New York. p. 215-245.

43. Whiten, A. and R. Ham, *On the nature and evolution of imitation in the animal kingdom: Reappraisal of a century of research*, in *Advances in the study of behavior*, P.J.B. Slater, et al., Editors. 1992, Academic Press: New York. p. 239-283.
44. Range, F. and L. Huber, *Attention span of common marmosets - Implications for social learning experiments.* Animal Behaviour, in press.
45. Chaminade, T., A.N. Meltzoff, and J. Decety, *An fMRI study of imitation: action representation and body schema.* Neuropsychologia, 2005. **43**(1): p. 115-27.
46. Zentall, T.R., J.E. Sutton, and L.M. Sherburne, *True imitative learning in pigeons.* Psychological Science, 1996. **7**(6): p. 343-346.
47. Miklósi, A., *The ethological analysis of imitation.* Biological Reviews, 1999. **74**: p. 347-374.
48. Heyes, C.M., *Causes and consequences of imitation.* Trends in Cognitive Science, 2001. **5**: p. 253–261.
49. Moore, B.R., *Avian movement imitation and a new form of mimicry: tracing the evolution of a complex form of learning.* Behaviour, 1992. **122**(3-4): p. 231-263.
50. Tayler, C.K. and G.S. Saayman, *Imitative behaviour by Indian Ocean bottlenose4 dolphins (Torsiops aduncus) in captivity.* Behaviour, 1973. **44**: p. 286-298.
51. Hayes, K.J. and C. Hayes, *Imitation in a home-raised chimpanzee.* Journal of Comparative and Physiological Psychology, 1952. **45**: p. 450-459.
52. Custance, D.-M., A. Whiten, and K.A. Bard, *Can young chimpanzees (Pan troglodytes) imitate arbitrary actions? Hayes & Hayes (1952) revisited.* Behaviour, 1995. **132**(11-12): p. 837-859.
53. Myowa-Yamakoshi, M., *Evolutionary foundation and development of imitation*, in *Primate origins of human cognition and behavior*, T. Matsuzawa, Editor. 2001, Springer: Tokyo.
54. Call, J., M. Carpenter, and M. Tomasello, *Copying results and copying actions in the process of social learning: chimpanzees (Pan troglodytes) and human children (Homo sapiens).* Anim Cogn, 2005. **8**(3): p. 151-63.
55. Rizzolatti, G., *The mirror neuron system and imitation*, in *Perspectives on imitation. From neuroscience to social science*, S. Hurley and N. Chater, Editors. 2005, MIT Press: Cambridge, MA. p. 55-76.
56. Fogassi, L., et al., *Parietal lobe: from action organization to intention understanding.* Science, 2005. **308**(5722): p. 662-7.
57. Ferrari, P.F., S. Rozzi, and L. Fogassi, *Mirror neurons responding to observation of actions made with tools in monkey ventral premotor cortex.* Journal of Cognitive Neuroscience, 2005. **17**: p. 212-226.
58. Iriki, A., M. Tanaka, and Y. Iwamura, *Coding of modified body schema during tool use by macaque postcentral neurones.* Neuroreport, 1996. **7**(14): p. 2325-30.
59. Haruno, M., D.M. Wolpert, and M. Kawato, *Mosaic model for sensorimotor learning and control.* Neural Computation, 2001. **13**: p. 2201–2220.
60. Wolpert, D.M., K. Doya, and M. Kawato, *A unifying computational framework for motor control and social interaction.* Philos Trans R Soc Lond B Biol Sci, 2003. **358**: p. 593-602.
61. Wolpert, D.M. and M. Kawato, *Multiple paired forward and inverse models for motor control.* Neural Networks, 1998. **11**: p. 1317–1329.
62. Doya, K., et al., *Recognition and imitation of movement patterns by a multiple predictor–controller architecture.* Technical Rep. IEICE, 2000. **TL2000-11**: p. 33–40.
63. Iacoboni, M., et al., *Reafferent copies of imitated actions in the right superior temporal cortex.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13995-9.
64. Iacoboni, M., *Neural mechanisms of imitation.* Curr Opin Neurobiol, 2005. **15**(6): p. 632-7.
65. Rumiati, R.I. and H. Bekkering, *To imitate or not to imitate? How the brain can do it, that is the question!* Brain and Cognition, 2003. **53**: p. 479-482.
66. Rumiati, R.I. and A. Tessari, *Imitation of novel and wellknown actions: The role of short-term memory.* Experimental Brain Research, 2002. **142**: p. 425–433.
67. Rumiati, R.I., et al., *Common and differential neural mechanisms supporting imitation of meaningful and meaningless actions.* J Cogn Neurosci, 2005. **17**(9): p. 1420-31.

Visuo-Cognitive Perspective Taking for Action Recognition

Matthew Johnson and **Yiannis Demiris** [1]

Abstract. Many excellent architectures exist that allow imitation of actions involving observable goals. In this paper, we develop a Simulation Theory-based architecture that uses continuous visual perspective taking to maintain a persistent model of the demonstrator's knowledge of object locations in dynamic environments; this allows an observer robot to attribute potential actions in the presence of goal occlusions, and predict the unfolding of actions through prediction of visual feedback to the demonstrator. The architecture is tested in robotic experiments, and results show that the approach also allows an observer robot to solve Theory-of-Mind tasks from the 'False Belief' paradigm.

1 Introduction

When we see another person performing an action, we are usually able to understand the purpose and intention underlying the action, and can reproduce the action for ourselves. The HAMMER architecture [5, 16] can be used to equip a robot with this common human ability. The HAMMER architecture achieves the mapping between observed and self-generated action by directly involving the observer robot's motor system in the action recognition process; during observation of the demonstrator's actions, all the observer's inverse models (akin to motor programs) are executed in parallel in simulation using forward models. The simulated actions generated by the inverse models are compared to the observed action, and the one that matches best is selected as being the observed action. The internal action simulation, combined with the comparison to the observed action, achieves the mapping between observed action and self-generated action that is required for imitation [4].

By using the motor system to achieve action recognition, the HAMMER architecture is taking a Simulation Theory approach to solving the imitation problem. In the 'Theory of Mind' paradigm, the Simulation Theory is used to attribute mental states to other people by using one's own cognitive decision-making mechanism as a manipulable model of other's minds, taken off-line and placed into the context of their situation [13, 9, 8]. For this to work, the state of the 'target' agent is used instead of one's own state, but transformed into an egocentric format that our first-person decision-making and behaviour-generation mechanisms will accept.

Similarly, in order to provide meaningful data for comparison, the simulated actions used by the HAMMER architecture during recognition must be generated as though from the point of view of the demonstrator. Since the HAMMER architecture uses a Simulation Theory approach, the observer's inverse models require first-person data in order to generate actions, and so spatial and visual perspective taking are used to achieve the egocentric 'shift' from the observer to the demonstrator. The data required for the inverse models to operate is therefore derived from consideration of the demonstrator's physiospatial and sensory circumstances, and not the observer's, using perspective taking [11].

However, it is not only instantaneous sensory information that informs goal selection and action planning. It is through keeping in memory details of objects that are seen that *cognitive maps* are built up, which are critical to action generation. In this paper, we present a Simulation Theory approach to perspective taking that allows an observer robot to use its visual perceptual mechanisms in simulation to determine what the demonstrator is seeing; by performing this process continually, the observer's first-person cognitive map generation routines can be used to build up and maintain a representation of the demonstrator's own cognitive map. Taking into account the demonstrator's knowledge of the world in this manner allows more accurate state and goal information to be fed to the HAMMER architecture.

2 Background

In common and also academic use, the term 'perspective taking' has many meanings in many different situations. There are such definitions as:

- "People's ability to experience and describe the presentation of an object or display from different vantage points" [1]
- "Imagining oneself in another's shoes" [7]
- "Understand[ing] how others perceive space and the relative positions of objects around them- [...] the ability to see things from another person's point of view" [15]
- "Consider[ing] the needs and wants of the opponent" [6]

In this paper we focus on equipping robots with perceptual and cognitive perspective-taking abilities, through a Simulation Theory approach, in order to improve the quality of the state information fed to the HAMMER architecture. In this architecture, a cognitive map is defined as being a *representation in memory of the location of observed objects*. This memory is updated continually from observation of the environment, and is available to the action generation system for action planning. The cognitive map is used also as a manipulable spatial model of the environment to facilitate perspective taking; to enable visual perceptual perspective taking, the objects in the cognitive map are linked with visuo-spatial representations that are used to re-create the visual image seen by the demonstrator.

[1] BioART, ISN Group, Department of Electrical and Electronic Engineering, Imperial College London. Email: {matthew.johnson, y.demiris}@imperial.ac.uk

2.1 HAMMER

The HAMMER (Hierarchical Attentive Multiple Models for Execution and Recognition) architecture is a Simulation-Theoretical architecture for action recognition and imitation, based on the hierarchical coupling of internal models to produce simulation loops. HAMMER achieves first-person action generation using coupled *inverse* and *forward* models, and uses the same arrangement to achieve imitation, but fed from a perceptual perspective-taking process involving internal inverse and forward *vision models*. The perceptual perspective taking process has been shown to improve the performance of action recognition in situations where the observer must take into account visual occlusions and visual cues provided to the target [11].

2.2 Internal Inverse and Forward Models

One of the core components of the HAMMER architecture is the *inverse model* for motor control. Inverse models represent functionally specialised units for generating actions to achieve certain goals. The generic inverse model takes as input the *current state* of a system, a *goal state* that is the system's desired state, and produces as output the action required to move the system from its current state to the goal state [12, 18]. In the control theory literature, the inverse model is known as a *controller* and its outputs are control signals; when applied to robotics, the current state is the state of the robot and its environment, and the outputs are motor commands. In that context, inverse models are known as *behaviours*.

Inverse models have several *internal states*, that are used in action execution and recognition [10]. One of these states is the *applicability* of the inverse model. When presented with a goal, the inverse model will calculate its level of applicability through simulation with its coupled forward model. The applicability is a measure of how useful the inverse model is for achieving the goal. A low applicability level means that the inverse model cannot achieve the goal from its current state, for example, the "Place object on shelf" inverse model when the shelf is too high to reach. The applicability level is explained in more detail in Section 3.3.

Forward models of causal dynamics are used in predictive control systems. The classic forward model takes as input a system state and the dynamics currently acting on the system, and produces as output the *predicted next state of the system*. In the HAMMER architecture, multiple forward models are *coupled* to inverse models to create a simulation process. This approach is similar to that used in other internal model-based systems for motor control [21, 20]. When coupled to an inverse model, a forward model receives the action output from the inverse model; the forward model then generates a prediction of the state that would result, if the action was to be performed.

2.3 Inverse and Forward Vision Models

In [11], the capacity for *visual perceptual perspective taking* was introduced to the HAMMER architecture. In keeping with the simulation theoretical approach, this was achieved through a biologically inspired *simulation of visual perception*. In the same way as action recognition and imitation is achieved in the HAMMER architecture through coupled inverse and forward models as used in control, visual perception and perspective taking is performed here through coupled inverse and forward models of the *visual* process. The *inverse vision model* is defined as taking two inputs, the first being a camera image, and the second being the visual parameters with which to process that image. The output from the model is the state output from processing. A *forward vision model* is defined as having two inputs and one output. The forward vision model takes as input visual object properties retrieved from the cognitive map (e.g. colour, shape, etc), and their desired state (e.g. positions and orientations taken from the cognitive map), and produces as output the visual image that results from reconstructing these inputs. Inverse and forward vision models are described in detail in [11].

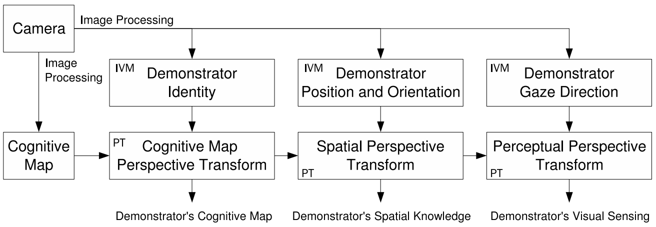

Camera
Image Processing
Image Processing
IVM Demonstrator Identity
IVM Demonstrator Position and Orientation
IVM Demonstrator Gaze Direction
Cognitive Map
PT Cognitive Map Perspective Transform
Spatial Perspective Transform PT
Perceptual Perspective Transform PT
Demonstrator's Cognitive Map
Demonstrator's Spatial Knowledge
Demonstrator's Visual Sensing

Figure 1. The perspective-taking process. Image information from the camera, and the robot's own knowledge held in the cognitive map, are fed into a cascade of perspective transform 'filters'. The outputs at each stage are used as the 'pretend states' fed into the HAMMER architecture. 'PT' indicates a perspective transform stage, and 'IVM' indicates an inverse vision model for performing image processing.

The coupling of inverse and forward vision models results in simulation of perception, and gives HAMMER the ability to consider what the demonstrator *sees*, as well as its position. This enables the observer robot to take into account visual occlusions effecting the demonstrator, and through *continual* usage, the observer can keep track of what objects the demonstrator has seen in the past and potentially stored in its cognitive map. Because the demonstrator sees different things to the observer due to its differing viewpoint, it becomes necessary for the observer robot to maintain a representation of the demonstrator's cognitive map in order to predict and recognise actions. In keeping with the simulation theory approach, this may be achieved by recruiting the observer's own cognitive map creation and updating processes, but fed with information derived from visual perceptual perspective taking instead of first-person visual information. Figure 2 shows the perception simulation process.

3 Implementation

The perspective taking architecture shown in Figures 1 and 2 was implemented in C++ for experiments involving an observing observer robot and a demonstrator robot. The target robots were ActivMedia Peoplebots, equipped with grippers and firewire cameras. A version of HAMMER was implemented and linked to the perspective taking architecture.

3.1 Inverse and Forward Vision Models

Inverse vision models were implemented using the ARToolkit Plus, an extension of the ARToolkit [2]. When presented with an image containing two-dimensional square markers (fiducials) of known size, the ARToolkit can calculate the position and orientation of the markers in world co-ordinates. A set of three objects was therefore produced with fiducials attached, and in order to extract the demonstrator robot's position and orientation at any point in time, a cubic AR 'hat' was made, with a fiducial on each vertical face. This ensured that no matter which direction the demonstrator robot was facing, the observer robot would be able to determine its location and orientation.

To construct visual scenes from the transformed cognitive map, the forward vision models used the OpenGL graphics library (www.opengl.org). To ensure that the same inverse vision models as used for first-person visual processing would work with the reconstructed image for the demonstrator robot, the fiducials used by the ARToolkit were added in as OpenGL textures and linked to the object entries in the cognitive map.

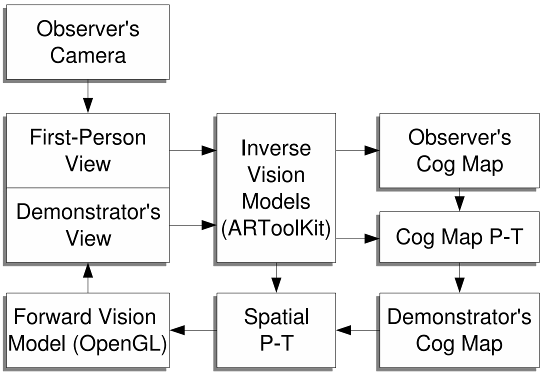


Figure 2. The perception simulation loop. The observer's first-person view of the scene is used to build up the observer's cognitive map of the scene. The cognitive map is filtered through the cognitive map perspective transform that 'filters' the observer's cognitive map to make it like the demonstrator's. This is then used as a basis for the perceptual perspective transform, that begins with a spatial geometric transform to 're-centre' on the demonstrator, and then fed through the forward vision model to re-create what the demonstrator is seeing. The observer can then use its inverse vision models on the image to update its representation of the demonstrator's cognitive map, in the same way as it would update its own.

3.2 Cognitive Map Definition

The cognitive map was defined as being a list of objects, held in memory. It was assumed that the robots already knew what each object was and could identify them through the inverse vision models (i.e. the inverse vision models were programmed to recognise the objects, through use of the fiducials, and extract relevant information). When visual information for the objects was available from the inverse vision models, the cognitive map entries for those objects were updated with world position and orientation values. Linked with this information was a three-dimensional model of each object, and visual information (e.g. colour and texture) that would be used by the observer's forward vision model to re-create the image of the scene from the point of view of the demonstrator. As can be seen from Figure 1, the perspective-taking process is then comprised of the following steps:

1. The demonstrator is identified and the correct cognitive map perspective transform, comprising the differences from the observer's own cognitive map, is applied;
2. A spatial perspective transform is applied to the resulting cognitive map, to re-centre it upon the demonstrator;
3. The forward vision model takes in the re-centred spatial data, and accesses the visual information linked to the objects in the cognitive map, to re-construct the image that the demonstrator is seeing.

This image is then processed by the observer's inverse vision models, to update the demonstrator's cognitive map transform, and to provide state information to HAMMER.

3.3 Inverse and Forward Models

Inverse models for the HAMMER architecture were implemented as PID controllers, generating robot velocities and delta-angle headings in order to minimise the distance between the robot grippers and a goal object. The state information required for the inverse models was taken from either the observer's own cognitive map, or its representation of the demonstrator's cognitive map. When used for action simulation, the applicability level A_t of the inverse model was calculated for the n^{th} simulation iteration according to:

$$A_{t,n} = \begin{cases} 0 & \text{at } n = 0 \\ A_{t,n-1} + \frac{1}{S_d} \times \frac{1}{n} & \text{if } \frac{dS_d}{dt} < 0 \\ A_{t,n-1} - \frac{1}{S_d} \times \frac{1}{n} & \text{if } \frac{dS_d}{dt} \geq 0 \end{cases} \tag{1}$$

The applicability accumulation is discounted over time and is increased (rewarded) if the inverse model is making progress towards achieving its goal, and decreased (punished) if it is not. The state distance between current state S_t and the goal state vector λ is defined as:

$$S_d = \sum_{i=1}^{M} |\lambda_i - S_{t,i}| \tag{2}$$

When S_d was less than a completion threshold ϵ_1, the inverse model became complete and did not generate motor commands even when instructed to execute. In the following experiments, ϵ_1 was chosen to be 0.04.

Forward models used Euler integration to form 1-timestep predictions based on the current state, and the robot velocity and heading motor commands generated by the inverse models. As in [11], the forward models were equipped with collision models of the robot and objects in the environment, allowing the forward model to predict position and velocity states in situations when the robot ran into tables or other objects.

3.4 Perspective Taking Visual Representations — 'Ghosts'

Through coupling the perception simulation loop to the simulation of action enabled by the HAMMER architecture, it is possible to *predict the visual feedback* arising from the action. By processing this visual feedback with the inverse vision models, and updating the cognitive map representation, the action simulation can continue further into the future, and the *outcome* of actions predicted. In section 4, experiments are described in which this approach is used to create multiple *Perspective Taking Visual Representations* (PTVRs), parallel instances of the observer's own perception and action mechanisms, each driven by a different inverse model. Using perspective taking, the observer can place these 'ghosts' in the place of the demonstrator, and use them to predict what the visual feedback will be from possible actions the demonstrator may perform. This allows the observer to predict the changes to the demonstrator's knowledge of the world during the course of a possible action, and how this may effect the course of the action. Figure 3 shows the arrangement.

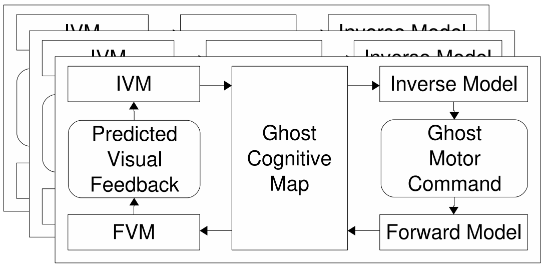


Figure 3. Perception and action simulation coupled for the generation of PTVRs ('ghosts'). Multiple ghosts can be instantiated and used in parallel, each one driven by a different inverse model. By coupling the HAMMER action simulation loop to the perception simulation loop, it is possible to predict the visual feedback to the ghost and therefore the updates to the ghost's cognitive map. A ghost can represent either the observer or the demonstrator performing a certain action.

4 Experiments

The implemented perspective-taking architecture was deployed onto the robots arranged in the scenario shown in Figure 4.

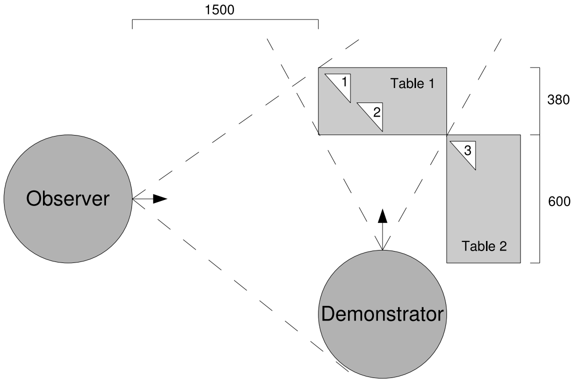


Figure 4. Plan view of the experimental setup. The observer robot faces two tables. Table 1 has objects 1 and 2, and table 2 has object 3. The demonstrator robot is positioned so that it, and the tables and the objects, can be clearly seen by the observer. The demonstrator is placed facing and close to table 1, so that initially it is able only to see objects 1 and 2. The dashed lines indicate the fields-of-view of the robots. The plan is not to scale, but measurements have been provided to indicate size and relative position. Measurements are in millimetres. The objects were 150mm across their long edge. The ARToolKit fiducials had 120mm edge, and were mounted as per Figure 5.

Three ARToolKit fiducials were attached to triangular objects, to enable both the observer and the demonstrator to identify and locate those objects. The objects were placed so that both the observer and the demonstrator could see the objects, however, initially, the observer could see all three objects whereas the demonstrator could see only the first two. The objects and the demonstrator robot were given simple 3D models to enable their reconstruction into the image produced by the forward vision model. Similarly, the implemented HAMMER architecture was provided with three 'nudge object' inverse models, one for each of the objects used during the experiments. These inverse models, when activated, produced motor commands for moving the robots from their current position to the specific object, and stopped when the robot gripper touched the object. During the experiments it was assumed that the demonstrator's camera was kept stationary with respect to the robot's frame, pointing directly ahead of the robot.

4.1 Experimental Scenarios

The first experiment was designed to test the architecture's ability to update its cognitive map representations, in the presence of demonstrator movements, and object movements both seen and unseen by the demonstrator. There were three parts:

1. The observer takes the perspective of the demonstrator robot, and initialises its representation of the demonstrator's cognitive map;
2. Demonstrator rotates 45 degrees to its right. The observer, through continual perspective taking, updates its cognitive map representations;
3. Object 1 is moved, unseen to the demonstrator, but seen by the observer.

Experiment Two took this further, by having the observer maintain its cognitive map representations over five episodes of object and demonstrator movements, in which objects were occluded from both the demonstrator and the observer. The observer also had to use its cognitive map representations to attempt to determine what actions the demonstrator *believed* it could perform, through using perspective taking and action simulation to calculate inverse model applicabilities. The sequence was the following:

1. Demonstrator can see objects 1 and 2. Object 1 is not graspable, and the demonstrator cannot see object 3 (Figure 4);
2. Object 1 is moved close to the demonstrator, occluding object 2 from the observer;
3. Demonstrator rotates 45 degrees to its right. Object 3 becomes observable (as per Figure 7 B);
4. Object 1 is moved back to original position. The observer can see this, but the demonstrator cannot;
5. Demonstrator rotates back to original position.

In Experiment Three the observer robot had to maintain its cognitive map representations over four episodes of *simultaneous* demonstrator and object movements. The observer also had to predict the visual feedback during each potential demonstrator action using its PTVRs, in order to predict the impact of false beliefs on the performance of the actions. The sequence was the following:

1. Demonstrator can see objects 1 and 2. Object 1 is not graspable, and the demonstrator cannot see object 3 (Figure 4);
2. Demonstrator rotates 45 degrees to its right, then object 1 is moved to a graspable position (unseen by demonstrator);
3. Objects 1 and 2 are moved away (unseen by demonstrator);
4. Demonstrator rotates back to original position. Object 3 moved away (unseen by demonstrator).

5 Results

5.1 Experiment 1

Figure 6 shows the observer's view of the scene during part 1 of the experiment. The demonstrator robot and the objects are visible. Figure 6 A shows the thresholded camera image fed to the observer's inverse vision models, and Figure 6 B shows the resulting reconstruction of the visual scene, using data from the observer's cognitive map and its forward vision model. The ARToolKit has successfully extracted the position and orientation of the objects and the demonstrator, and Table 1 shows the contents of the observer's cognitive map resulting from the processing. The X, Y, Z position and angle of objects are extracted and updated in the observer's cognitive map representations while the objects are visible.

Table 1. Cognitive map entries for centroid positions and orientations of objects when viewed by the observer robot (first-person perspective). The values shown are relative to the observer's camera position and orientation. The results correspond to the scene shown in Figure 6.

Object	**X** (m)	**Y** (m)	**Z** (m)	**Angle** (Degrees)
Demonstrator	-0.56	0.20	1.80	357.51
Object 1	0.32	-0.28	1.61	44.03
Object 2	0.20	-0.27	1.75	50.39
Object 3	-0.15	-0.24	2.21	12.13
Table 1	0.37	0.00	1.75	90.00
Table 2	-0.15	0.00	2.30	0.00

Figure 7 shows the result of the perceptual perspective taking. Figure 7 A is what the observer determines the demonstrator to be seeing during the first part of the experiment; Figure 5 shows the demonstrator's actual camera image of this scene—the simulation of perception has clearly resulted in accurate perspective taking. Objects 1 and 2 are observed, but object 3 is outside the field-of-view on the table to the right. Figure 7 B shows the scene during part 2, after the demonstrator has rotated 45 degrees to the right. The observer robot realises through perspective taking that object 3 is now visible to the demonstrator, and objects 1 and 2 are not.

Table 2 shows the results from part 3 — moving object 1 while it can be seen by the observer, but not by the demonstrator. Through perceptual perspective taking the observer knows that the demonstrator cannot see the object being moved — and so, it updates its *own* cognitive map with the change in position, but not the demonstrator's. This leads to the discrepancy between the object 1 position values for the observer and the demonstrator, as shown in the table. The demonstrator *believes* that the object is in the place where it last saw it, whereas the observer *knows* it to be somewhere else.

Table 2. Cognitive map entries for observer and demonstrator after movement of object 1 inside the observer's field of view but outside of the demonstrator's. Through perceptual perspective taking, the observer knows that the demonstrator cannot see object 1 while it is being moved, and so the demonstrator's cognitive map is not updated with the changes in position and orientation as the object is moved.

Object	**X** (m)	**Y** (m)	**Z** (m)	**Angle** (Degrees)
Observer's Cognitive Map				
Object 1	0.12	0.06	0.55	20.96
Object 2	0.24	-0.26	1.82	55.07
Object 3	0.05	-0.23	2.17	40.86
Demonstrator's Cognitive Map				
Object 1	0.26	-0.03	0.83	64.40
Object 2	0.24	-0.26	1.82	55.07
Object 3	0.05	-0.23	2.17	40.86

5.2 Experiment 2

Building on the success of Experiment 1, the perspective taking was linked to the HAMMER architecture for action simulation experiments. Figure 8 shows the results of the applicability calculations for Experiment 2. Each episode is separated by a period of zero applicability, before the action simulations begin and then reset five seconds later. Figure 8 A shows the applicability levels when the observer is drawing on its representation of the demonstrator's cognitive map to generate state information for the inverse models, and in Figure 8 B the observer is using its own cognitive map. The final applicability levels achieved by each inverse model are shown in Table 3.

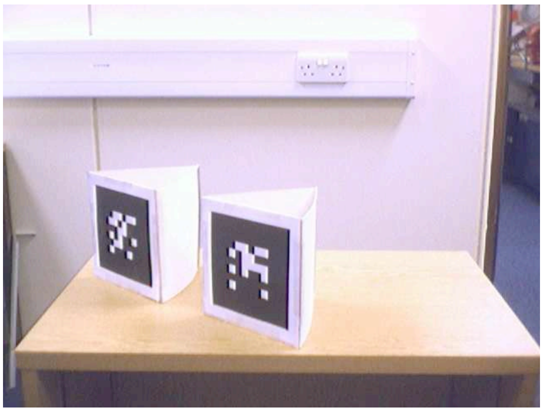


Figure 5. The demonstrator's view of the table and objects 1 and 2. Figure 7 (A) shows an internal simulation of this viewpoint by the observer.

The top graph effectively shows the observer's attempt to determine, through simulation, what actions the demonstrator *believes* it can perform; the lower graph is the observer calculating what actions the demonstrator can actually perform, given the state of the world as the observer knows it to be. In the first three episodes, the demonstrator's cognitive map and the observer's own are in agreement as to what inverse models are applicable: the 'nudge object 3' inverse model is not simulated for the demonstrator in the first two episodes as the observer determines that the demonstrator is unaware of object 3's existence (through the perceptual perspective taking). While the demonstrator is looking at object 3, object 1 is moved to an un-nudgeable position; the demonstrator does not see this, but the observer does, the result being that the observer calculates that the demonstrator still believes that touching object 1 is possible, even though it itself knows that the action cannot be accomplished. Upon the demonstrator rotating back to observe objects 1 and 2 in episode 5, the false belief is resolved and the applicability levels are once again in agreement.

5.3 Experiment 3

Experiment 3 took the perspective taking-action simulation of Experiment 2 further, by having the observer use its PTVRs to predict the visual feedback resulting from potential demonstrator actions, and through this, predict the updates to the demonstrator's cognitive map and how this would effect the outcome of each action. Figure 9 shows the results. Figure 9 A shows the applicability levels of the three inverse models over the four episodes, as determined by the observer when observing the demonstrator and basing its action simulations on its representation of the demonstrator's cognitive map. Figure 9 B, C and D show the cognitive map updates predicted by each of the three 'ghosts', as used by the observer during prediction of visual feedback.

In this experiment, the demonstrator robot may not see an object being moved *at the time*, but if it believes an action with that object

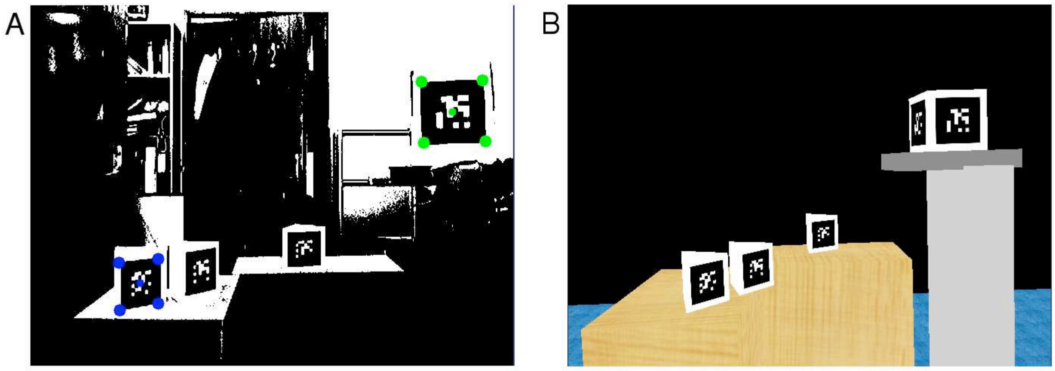


Figure 6. The observer's view of the scene. (A) shows the thresholded camera image sent to the inverse vision models. Objects 1 and 2 are on the table facing the demonstrator, and object 3 is on the table facing the observer. (B) shows the observer's cognitive map, rendered by OpenGL. The three fiducial markers can clearly be seen on the tables, and the demonstrator robot (and its 'hat') can be seen to the right.

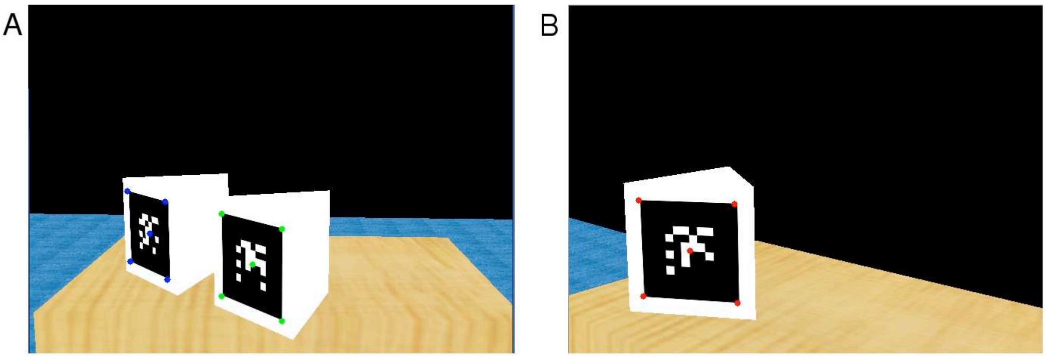


Figure 7. The demonstrator's view of the scene, re-created by the observer in simulation. (A) shows the what the demonstrator sees at the beginning of the experiment, objects 1 and 2. In (B), the demonstrator robot has rotated 45 degrees to the right, and the observer determines that it is able to see object 3. The demonstrator's actual view of (A) is shown in Figure 5.

is still possible and begins to execute it, then after it has rotated and seen the new object configuration, its cognitive map will be updated, the applicability of the action re-calculated, and then it will stop execution since it realises the action is now impossible. Episodes 3 and 4 show this; while the demonstrator is looking at object 3, objects 1 and 2 are moved away from the edge of the table. The demonstrator still believes that the objects are touchable, and so the observer sends out 'ghosts' to simulate how the action may unfold. The spikes in Figure 9 B and C show the predicted updates to the demonstrator's cognitive map when it sees that the objects have moved; as can be seen from figure Figure 9 A, negative applicabilities are calculated and the observer predicts the demonstrator will stop executing those actions. In episode four, the demonstrator rotates to observe the new configuration of objects 1 and 2, and unseen, object 3 is moved away. Figure 9 D shows the resulting cognitive map update for that episode. Again, the result is that the inverse model is no longer applicable and the action is halted mid-execution.

6 Discussion

In developmental psychology, several experimental tasks have been devised in order to investigate the development of cognitive perspective-taking abilities in the paradigm of *false belief*. One of the first tasks in this field was devised by the developmental psychologists Heinz Wimmer and Josef Perner, in response to Daniel Dennett's critique of the experimental protocols used by David Premack and Guy Woodruff in their seminal article that originated the term 'Theory of Mind' [19, 14]. This is the *action prediction* task (also known as the "unexpected transfer" task).

The action prediction task tests an observer determining what a target agent will do when holding a false belief about the world. The test subject, usually a child, observes a puppet-show involving the main character, "Maxi", and his mother. In the show, Maxi watches his mother place a chocolate bar inside a box. Maxi then leaves the room and his mother transfers the chocolate from that box into a different one. Maxi then returns, and the subject is asked where he will look for the chocolate. Further questions include what Maxi would tell to someone he wants to deceive as to the location of the chocolate, and someone he would want to tell the truth to. The result of this task is that four-year-old children give predictions based on correctly attributing the false belief, whereas younger children do not.

Through the use of the cognitive map perspective taking described in this paper, the observer would be able to solve this task. By being able to represent the cognitive map of the demonstrator robot separate to its own, the observer robot is intrinsically able to represent the concept that the demonstrator may possess a false belief about the

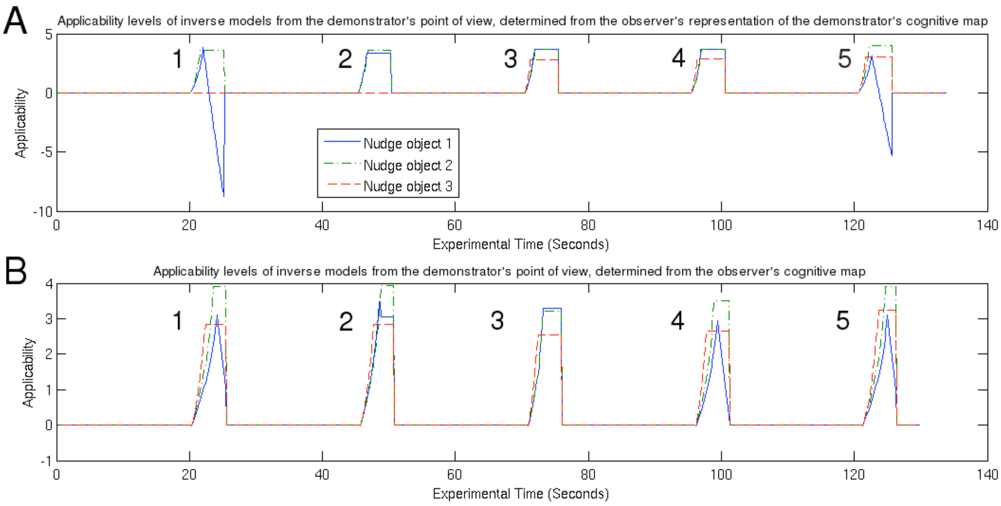


Figure 8. Applicability levels of the observer's inverse models to the demonstrator, over five repeated episodes. Each episode lasted five seconds, after which the applicability levels were reset to zero. Table 3 shows the final applicability levels for each inverse model at the end of the each episode.

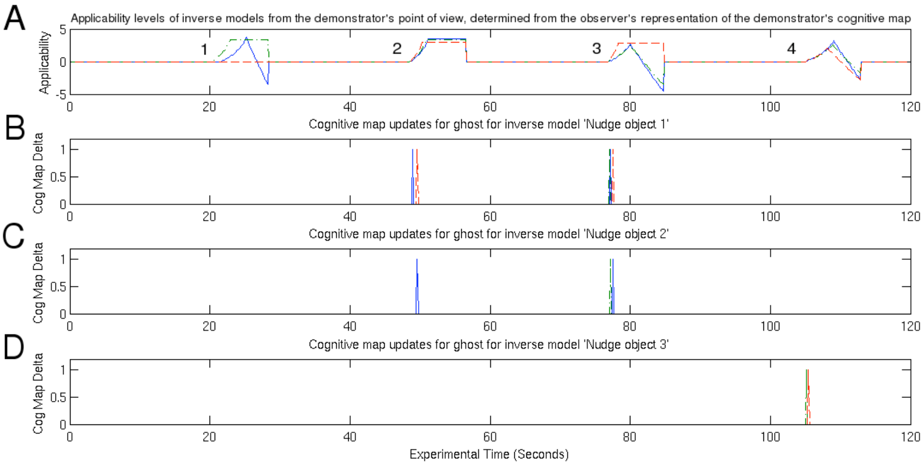


Figure 9. **A.** Applicability levels of the observer's inverse models to the demonstrator, over four repeated episodes. **B, C, D.** Cognitive map updates for each of the three 'ghosts', executing the inverse models 'nudge object' 1, 2, and 3 respectively. A spike indicates that the 'ghost' has seen something that necessitates a change to the cognitive map, and an update is made accordingly. The legend for these graphs is the same as for Figure 8.

location of objects in the world, due to objects moving outside the field-of-view, or object movement being obscured due to occlusions within the field of view. When asked to make predictions as to what the demonstrator may do in such situations, the observer robot is then able to take into account the false belief in the demonstrator's goal setting and action planning. This is illustrated through the results to part 3 of experiment 1, detailed in section 5.1. Through perceptual perspective taking the observer knows that the demonstrator cannot see object 1 being moved — and so, it updates its *own* cognitive map with the change in position, but not the demonstrator's.

Knowledge of this kind, as to the presence of false beliefs in observed agents, can be used by an observer to determine what actions a target agent considers to be available to it, as opposed to what actions it can in fact perform. This information is useful when priming a Simulation-Theory based architecture, such as the HAMMER architecture, with the action simulations it requires for action recognition. The demonstrator will derive its own action goals from what it believes to be the state of the world and move accordingly, and without a representation of the demonstrator's cognitive map, the observer will feed its perspective transform with its own world-state beliefs and potentially end up hypothesizing different goals for its action generation systems — this results in the comparison between internally generated action and observed action being meaningless. In other words, using perceptual perspective taking alone means that we see the world as *we* believe it is from the demonstrator's point of view, whereas what we need to do, in order to infer intention, is see the world as the *demonstrator* believes it is from the demonstrator's point of view. To do the former is to risk not recognising the demonstrator's movements and their action context at all, or to mis-recognise the action as being something else, or to be unable

Table 3. Final applicability levels for each inverse model shown in Figure 8. 'D' indicates that the observer is using its representation of the demonstrator's cognitive map when determining what inverse models are applicable; 'O' indicates that the observer is using its own cognitive map to determine the applicability level of the inverse models. The numbers highlighted in bold type, for 'Nudge object 1' in episode 4, indicates the situation where the observer determines that the demonstrator may possess a *false belief* as to the actions it can make. The absence applicability levels for 'Nudge object 3' in episodes 1 and 2 is due to the demonstrator robot being unaware, at that stage, of the existence of object 3.

	Episode 1		**Episode 2**		**Episode 3**		**Episode 4**		**Episode 5**	
I-Model	D	O	D	O	D	O	D	O	D	O
Nudge object 1	-8.80	1.07	3.35	3.05	3.65	3.28	**3.65**	**0.26**	-5.28	0.98
Nudge object 2	3.59	3.89	3.59	3.93	3.69	3.22	3.69	3.50	3.97	3.90
Nudge object 3	—	2.85	—	2.86	2.83	2.55	2.87	2.65	3.01	3.25

to interpret the demonstrator's goal, and therefore be unable to imitate or learn. The results for experiments 2 and 3 show how through coupling the perspective taking architecture developed in this paper to the action simulation capabilities of HAMMER, the observer can successfully *predict* and *attribute* actions to the demonstrator, while taking into account prior knowledge and experience, and potential false beliefs.

In previous research, the HAMMER architecture was used to model and make predictions regarding the visuomotor 'mirror' neurons found in area F5 of macaque monkey premotor cortex [3]. These neurons are active both when observing an object-directed action, and when performing the same action, leading to suggestions that they underly the imitation capability. Recently, it was found that a subset of these neurons fire even when the object goal of the action is hidden from view, so long as the observer has prior knowledge of the object's presence [17]. With the addition of the cognitive map mechanism described in this paper, HAMMER gains this capability, by keeping a long-term memory of the locations of objects. This can be seen in the results for episode 2 of Experiment 2, where object 1 occludes object 2 from the observer's sight, but the action simulation is still performed. Furthermore, the results of section 5 offer a further prediction—that when a demonstrator performs an action based on a known *false belief* as to the presence of an object, the observer's mirror neurons will fire. Although there is currently no evidence either way, this would lend support to the hypothesis that the mirror neurons encode *intention* and underly action understanding, in addition to action recognition.

7 Conclusions

In this paper we have presented a perspective-taking architecture that uses simulation of visual perception to build up and maintain representations of the cognitive map of a demonstrator. This mechanism, used to improve the state information provided to the HAMMER imitation architecture, was deployed onto robots for perspective-taking and action-prediction experiments, in which an observer successfully attributed potential actions and action predictions to a demonstrator possessing false beliefs regarding the environment. In future work, the mechanism will extended and investigated in experiments involving the observer inferring false beliefs from the actions of a demonstrator.

REFERENCES

[1] E. K. Ackermann, 'Perspective-taking and object construction: Two keys to learning', in *Constructionism in Practice: Designing, Thinking and Learning in a Digital World*, eds., Y. Kafai and M. Resnick, chapter 2, 25–37, Mahwah, New Jersey: Lawrence Erlbaum Associates, first edn., (1996).

[2] M. Billinghurst, H. Kato, and I. Poupyrev, 'The magicbook: A transitional AR interface', *Computers and Graphics*, 745–753, (November 2001).

[3] Y. Demiris, 'Imitation, mirror neurons, and the learning of movement sequences', in *Proceedings of the International Conference on Neural Information Processing (ICONIP-2002)*, pp. 111–115. IEEE Press, (2002).

[4] Y. Demiris and M. Johnson, 'Distributed, predictive perception of actions: A biologically inspired robotics architecture for imitation and learning', *Connection Science*, **15**(4), 231–243, (December 2003).

[5] Y. Demiris and B. Khadhouri, 'Hierarchical attentive multiple models for execution and recognition', *Robotics and Autonomous Systems*, **54**, 361–369, (2006).

[6] A. Drolet, R. Larrick, and M. W. Morris, 'Thinking of others: How perspective taking changes negotiators' aspirations and fairness perceptions as a function of negotiator relationships', *Basic and Applied Social Psychology*, **20**(1), 23–31, (1998).

[7] A. D. Galinsky, G. Ku, and C. S. Wang, 'Perspective-taking and self-other overlap: Fostering social bonds and facilitating social coordination', *Group Processes and Intergroup Relations*, **8**(2), 109–124, (2005).

[8] V. Gallese, 'The manifold nature of interpersonal relations: the quest for a common mechanism', *Phil. Trans. of the Royal Society of London B*, **358**, 517–528, (2003).

[9] R. M. Gordon, 'Simulation vs theory-theory', in *The MIT Encyclopædia of the Cognitive Sciences*, eds., R. A. Wilson and F. Keil, 765–766, MIT Press, (1999).

[10] M. Johnson and Y. Demiris, 'Hierarchies of coupled inverse and forward models for abstraction in robot action planning, recognition and imitation', in *Proceedings of the AISB 2005 "Third International Symposium on Imitation in Animals and Artifacts"*, (April 2005).

[11] M. Johnson and Y. Demiris, 'Perceptual perspective taking and action recognition', *International Journal of Advanced Robotic Systems*, **2**(4), 301–308, (December 2005).

[12] K. S. Narendra and J. Balakrishnan, 'Adaptive control using multiple models', *IEEE Transactions on Automatic Control*, **42**(2), 171–187, (February 1997).

[13] S. Nichols and S. P. Stich, *Mindreading*, Oxford University Press, 2003.

[14] D. Premack and G. Woodruff, 'Does the chimpanzee have a theory of mind?', *Behavioral and Brain Sciences*, **4**, 515–526, (1978).

[15] A. C. Schultz and J. G. Trafton, 'Towards collaboration with robots in shared space: spatial perspective and frames of reference', *Interactions*, 22–23, (March-April 2005).

[16] G. Simmons and Y. Demiris, 'Object grasping using the minimum variance model', *Biological Cybernetics*, **94**(5), 393–407, (May 2006).

[17] M. A. Umilta, E. Kohler, V. Gallese, L. Fogassi, L. Fadiga, C. Keysers, and G. Rizzolatti, 'I know what you are doing: A neurophysiological study', *Neuron*, **31**, 155–165, (July 2001).

[18] Y. Wada and M. Kawato, 'A neural network model for arm trajectory formation using forward and inverse dynamics models', *Neural Networks*, **6**, 919–932, (1993).

[19] H. Wimmer and J. Perner, 'Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception', *Cognition*, **13**, 103–128, (1983).

[20] D. M. Wolpert, K. Doya, and M. Kawato, 'A unifying computational framework for motor control and social interaction', *Phil. Trans. of the Royal Society of London B*, **358**, 593–602, (2003).

[21] D. M. Wolpert and M. Kawato, 'Multiple paired forward and inverse models for motor control', *Neural Networks*, **11**, 1317–1329, (1998).

Learning by Observation: Comparison of Three Methods of Embedding Mentor's Knowledge in Reinforcement Learning Algorithms

Natalia Akchurina[1]

Abstract. Using knowledge of already successfully functioning agents can help to avoid expensive exploration that is so vital to some domains of reinforcement learning. Three methods to embed mentor's knowledge are proposed: initialization of Q function, reward shaping and implementing mentor's decisions in a separate action-value function. The speed of convergence of these methods in combination with Q-learning algorithm with different amount of information on mentor's decisions and their robustness to the quality of mentor are compared on four domains from the benchmarks for testing and comparing reinforcement learning algorithms "Reinforcement Learning Benchmarks and Bake-offs".

1 INTRODUCTION

Reinforcement learning has attracted rapidly increasing interest in the machine learning and artificial intelligence communities in the last years. Its promise is very tempting — a way of programming agents by reward and punishment without specifying how the task is to be achieved. The framework permits the simplifications necessary in order to make progress in domains that are clearly beyond our current capabilities due to many unsolved key problems in learning and representation.

Reinforcement learning is the problem faced by an agent that must learn to choose optimal sequences of actions through trial-and-error interactions with its environment. This very general problem covers tasks such as learning to control a mobile robot (robot control problems, such as navigation, pole-balancing, or juggling are the canonical reinforcement-learning problems), learning to optimize operations, learning to play board games (TD-Gammon program [16], based on reinforcement learning, has become a world-class backgammon player) or set prices on virtual markets [2]. The task of the agent is to learn from indirect, delayed reward a policy that maps states of the world to the actions the agent should take to maximize its cumulative reward. Trial-and-error search and delayed reward — are the two most important distinguishing features of reinforcement learning. One of the most important breakthroughs in reinforcement learning was the development of Q-learning algorithm. Q-learning can acquire optimal policies from delayed rewards, even when the agent has no prior knowledge of the effects of its actions on the environment.

While learning the agent must choose whether to explore unknown states and actions (to gather new information), or hold to states and actions that it has already learned will yield high rewards (to maximize its cumulative reward).

For the agent learns through trial and error the quality of its decisions is rather poor at the beginning. Some mistakes in exploration can even lead to vital consequences. For example, while learning to drive a bicycle by reinforcement [12] the robot fell so much and got deformed so badly that could not be used anymore for any purpose let alone to balance on the bicycle. On virtual markets a real profit can be wasted while learning how to set the prices [2, 17].

How can this expensive exploration be avoided?

In many environments already exist agents that function successfully and whose knowledge can be of use to avoid committing vital mistakes and making expensive exploration.

But how can we tell the difference between "experts" and those agents whose policies are not so effective? In general case we don't know in advance how good the agent's decisions are.

To implement agent's experience under this condition we need new methods that must possess the following properties:

- They must be robust to the quality of the expert (they must learn the optimal policy even if the expert's decisions turned out to be not so good)
- Usage of bad decisions should not radically worsen the speed of the convergence of the methods
- Usage of "real expert's" decisions should improve the convergence of the methods even if the data were scarce

Evident solution seems to embed somehow experience in existing good reinforcement learning algorithm.

In this paper we are proposing three methods to implement expert's knowledge in Q-learning algorithm (though they can be used with any action-value function based reinforcement learning algorithms): initialization of Q function, reward shaping and implementing mentor's decisions in a separate action-value function. We suppose that expert's knowledge is available in the form of state→action pairs (what action is better to take in this particular state) but this state-action pairs don't constitute the whole policy but come as a result of observing the expert agent for some time (to obtain such state-action pairs by observation is quite possible for many environments [14, 17, 5, 6]). This is a much weaker assumption than the supposition that the agent can get hold of state→action→next state→reward sequences [9, 7, 8, 15, 18, 19]. It also allows us not only to avoid problems with communication but to learn from a broader range of agents including those who just don't care to share their experience with us as well as competitive agents who in no way concern them-

[1] International Graduate School of Dynamic Intelligent Systems, University of Paderborn, Germany, email: anatalia@mail.uni-paderborn.de

selves with our successful learning[2].

The problem of learning a control policy to maximize cumulative reward is in general one of learning to control sequential processes. Subsection 2.1 is devoted to the introduction of the general formulation of this problem based on Markov decision process. In subsection 2.2 we present Q-learning algorithm. Three methods to embed expert's knowledge in action-value reinforcement learning algorithms are given in section 3. Section 4 is devoted to the depiction of the domains on which the three methods were tested and section 5 — to the results of the experiments. The discussion how well the proposed methods satisfy the above formulated requirements is presented in section 6. In a few words, reward shaping turned out to be quite inapplicable for the above problem. Q-initialization has proven to be robust to mentor's level of expertise, meanwhile the last method requires a very small amount of data on mentor's decisions.

2 PRELIMINARIES

2.1 Definitions

Here we introduce a general formulation of the problem of learning sequential control strategies, based on Markov decision process.

In a Markov decision process (MDP) [11] the agent can perceive a set S of distinct states of its environment and has a set A of actions that it can perform. At each discrete time step t, the agent senses the current state s_t, chooses a current action a_t, and performs it. The environment responds by giving the agent a reward $r_t = r(s_t, a_t)$ and by producing the succeeding state s_{t+1} with the transition probability $P_{s_t a_t}(s_{t+1})$. Here the function r and the transition probabilities $P_{sa}(\cdot)$ are part of the environment and are not necessarily known to the agent. In an MDP, r and $P_{sa}(\cdot)$ depend only on the current state and action, and not on earlier states or actions. We consider only the case in which S and A are finite.

Formally, a *finite nondeterministic MDP* is a tuple $M = (S, A, \delta, \gamma, r)$, where

- S is a finite set of m *states*
- $A = \{a_1, a_2, \dots, a_n\}$ is a set of n *actions*
- $\delta = \{P_{sa}(\cdot) | s \in S, a \in A\}$, where $P_{sa}(\cdot)$ are the *state transition probabilities* upon taking action a in state s
- $\gamma \in [0, 1)$ is the *discount factor*
- $r : S \times A \to \mathbb{R}$ is the *reward function*, bounded in absolute value by r_{max}

A *policy* is a function $\pi : S \to A$, that maps the current observed state s_t to action a_t, $\pi(s_t) = a_t$. The expected value $V^\pi(s_t)$ of the *discounted cumulative reward* achieved by following an arbitrary policy π from an arbitrary initial state s_t is defined as follows:

$$V^\pi(s_t) \equiv E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots] \equiv E[\sum_{i=0}^{\infty} \gamma^i r_{t+i}]$$

where the sequence of rewards r_{t+i} is generated by beginning at state s_t and by repeatedly using the policy π to select actions as described above (i.e., $a_t = \pi(s_t)$, $a_{t+1} = \pi(s_{t+1})$, etc.). Here $0 \le \gamma < 1$ is the *discount factor* that determines the relative value of delayed versus immediate rewards.

Agent's learning task is to learn a policy π that maximizes $V^\pi(s)$ for all states s. We will call such a policy an *optimal policy* and denote it by $\pi^\star$.

$$\pi^\star \equiv arg \max_\pi V^\pi(s), (\forall s)$$

$V^{\pi^\star}(s)$ gives the maximum expected value of the discounted cumulative reward that the agent can obtain starting from state s; that is, the expected value of the discounted cumulative reward obtained by following the optimal policy beginning at state s.

The optimal action in state s is the action a that maximizes the expected value of the sum of the immediate reward $r(s, a)$ plus the value $V^{\pi^\star}$ of the immediate successor state, discounted by γ.

$$\pi^\star(s) \equiv arg \max_a E_{s' \sim P_{sa}(\cdot)}[r(s, a) + \gamma V^{\pi^\star}(s')]$$

(where the notation $s' \sim P_{sa}(\cdot)$ means that s' is drawn according to the distribution $P_{sa}(\cdot)$)

Let us define the evaluation function $Q(s, a)$ so that its value is the maximum expected value of the discounted cumulative reward that can be achieved starting from state s and applying action a as the first action. In other words, the value of Q is the expected value of the reward received immediately upon executing action a from state s, plus the value (discounted by γ) of following the optimal policy thereafter.

$$Q(s, a) \equiv E_{s' \sim P_{sa}(\cdot)}[r(s, a) + \gamma V^{\pi^\star}(s')]$$

Therefore, optimal policy in terms of $Q(s, a)$ will be

$$\pi^\star(s) = arg \max_a Q(s, a)$$

2.2 Q-learning

One of the most important breakthroughs in reinforcement learning was the development of Q-learning algorithm. Q-learning can acquire optimal control policies from delayed rewards, even when the reward function and / or state transition probabilities are not known. All that is required for correct convergence is that the system can be modeled as a nondeterministic Markov decision process, the reward function r is bounded, and actions are chosen so that every state-action pair is visited infinitely often.

```
Q-learning algorithm for episodic tasks (with terminal states):
Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
• Initialize $s$
• Repeat (for each step of episode):
    – Choose $a$ from $s$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
    – Take action $a$, observe $r$, $s'$
    – $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$
    – $s \leftarrow s'$;
• until $s$ is terminal
```

Parameter α in Q-learning algorithm must satisfy $0 < \alpha \le 1$.

Choosing a from s using ϵ-greedy policy derived from Q means that most of the time we choose an action a that has maximal estimated action value $Q(s, a)$, but with probability ϵ we instead select an action at random.

[2] For the sake of consistency, we will use the terms "mentor" and "expert" to describe any agent from which we can learn by observation even if the mentor (expert) is hostile and we only suspect it to possess enough level of expertise to function effectively in the environment

3 SUPERVISED REINFORCEMENT LEARNING METHODS

Here we are proposing three methods to embed expert's decisions in Q-learning algorithm.

We suppose that state-action pairs (s, a) of expert's choice are available, that is not such bold an assumption for many environments [14, 17, 5, 6].

3.1 Q-initialization

Q-initialization was before mentioned in [14] to use prior knowledge to accelerate learning.

To embed the expert's knowledge in Q-learning algorithm we can use it thus:

Let's consider state s.

Let $\{a_1, a_2, \ldots, a_n\}$ be the set of the possible actions in this state s.

If the expert agent has chosen action a_k, $1 \leq k \leq n$, at this state it means in the terms of Q-function that:

$$
\begin{gathered}
Q(s, a_k) > Q(s, a_1) \\
Q(s, a_k) > Q(s, a_2) \\
\vdots \\
Q(s, a_k) > Q(s, a_{k-1}) \\
Q(s, a_k) > Q(s, a_{k+1}) \\
\vdots \\
Q(s, a_k) > Q(s, a_n)
\end{gathered}
$$

By initializing $Q(s, a_k)$ to some nonzero value and $Q(s, a_i)$ to zero, $i = 1, \ldots, n$, $i \neq k$, we can take into account the expert's decision.

3.2 Reward shaping

Reward shaping is a technique of changing rewards to accelerate learning and turned out to be very useful and even vital for the following tasks:

- A foraging task of a group of mobile robots [10]
- Learning to drive a bicycle [12]

The idea of this promising approach [10, 12], which is borrowed from behavioral psychology, is to give the learning agent a series of relatively easy problems building up to the harder problem of ultimate interest. The term was first proposed in [13], who studied the effect on animals, especially pigeons and rats. To train an animal to produce a certain behavior, the trainer must find out what subtasks form the desired behavior, and how these should be reinforced. By reward shaping horses can be brought to recognize numbers and pigs to eat at a table [4].

To use reward shaping for our task we need two reward functions instead of one:

- τ — artificial reward function
- r — usual reward function from the environment

Let's consider state s.

Let $\{a_1, a_2, \ldots, a_n\}$ be the set of the possible actions in the state s.

If the expert agent has chosen action a_k, $1 \leq k \leq n$, at this state, then we initialize $\tau(s, a_k)$ with some small positive value, and $\tau(s, a_i) = 0$, $i = 1, \ldots, n$, $i \neq k$.

The idea is to bias the action choice to the expert's one.

3.3 Two action-value functions

The idea to use two action-value functions was first proposed in [18]. But there the second action-value function was used to implement immediate feedback on agent's performance provided by external critic. Here we use two action-value function to implement the idea always to trust the expert's choice and to explore states in background mode and to use its results only in case of the lack of information about the expert's decision.

Here we derive policy from the sum of two action-value functions $Q + B$:

- B — action-value function reflecting the expert's choice
- Q — usual action-value function

Let's consider state s.

Let $\{a_1, a_2, \ldots, a_n\}$ be the set of the possible actions in the state s.

If the expert agent has chosen action a_k, $1 \leq k \leq n$, at this state, then we initialize $B(s, a_k)$ so that $B(s, a_k) > r_{max}$, and $B(s, a_i) = 0$, $i = 1, \ldots, n$, $i \neq k$.

4 BENCHMARKS

We have tested the above proposed methods on four domains from the benchmarks for testing and comparing reinforcement learning algorithms "Reinforcement Learning Benchmarks and Bake-offs" [1] that were the result of two workshops "NIPS Workshop on Reinforcement Learning: Benchmarks and Bake-offs" 2004 and 2005.

4.1 Gridworld with mines

The figure 1 shows a standard gridworld, with start and goal states, but with one difference: there are mines located at various positions on the grid. These mines are stationary and cause the agent to be destroyed if touched.

- *State space*: The state is represented by an integer value of the state label, comprised of the row index (x) multiplied by the number of columns plus the column index (y), resulting in 108 unique states
- *Starting states*: Agent is started in a new random starting state at the beginning of each episode
- *Terminal states*: Mines or goal state
- *Actions*: The actions are the standard four: *up*, *down*, *right* and *left*
- *Rewards*: $+10$ for reaching the goal, $-\ 10$ for hitting a mine and -1 otherwise

4.2 Distributed sensor network (DSN)

The distributed sensor network (DSN) problem is a sequential decision making variant of the distributed constraint optimization problem described in [3].

The network consists of two parallel chains of an arbitrary, but equal, number of sensors. The area between the sensors is divided into cells. Each cell is surrounded by exactly four sensors and can be occupied by a target. With equal probability a target moves to the cell to its left, to the cell to its right or remains on its current position. Actions that move a target to an already occupied cell are not executed. It is the goal of the sensors to capture all targets. See figure 2 for a configuration with eight sensors and two targets.

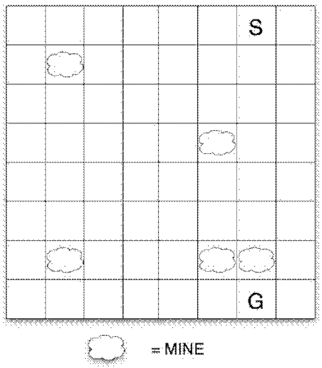


Figure 1. Gridworld with mines

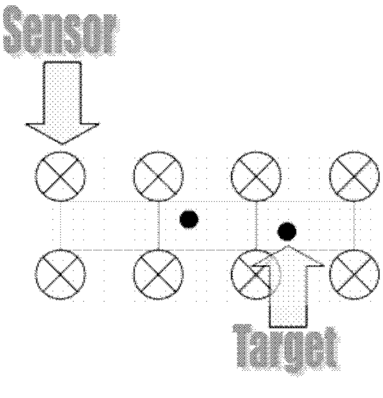


Figure 2. Sensor network configuration with eight sensors and two targets

Each sensor is able to perform three actions: track a target in the cell to its immediate left, to its immediate right, or don't track at all. Every track action has a small cost (reward of -1). When in one time step at least three of the four surrounding sensors track a target, it is 'hit'. Each target starts with a default energy level of three. Each time a target is hit its energy level is decreased by one. When it reaches zero the target is captured and removed. The three sensors involved in the capture are each provided with a reward of $+10$. An episode finishes when all targets are captured. Each joint action comprises single actions and the received reward is the sum of the individual agent rewards.

- *State space* (37): 9 states for each of the 3 configurations with 2 agents, 9 for those with one agent and 1 for those without any agents
- *Starting states* (3): $[3, 3, 0]$, $[3, 0, 3]$, $[0, 3, 3]$ (where the starting state $[3, 3, 0]$ corresponds to the initial situation when the two targets are in the first two cells and possess the maximum energy level 3)
- *Terminal states* (1): when both targets have zero hit points $[0, 0, 0]$
- *Actions* (6561): the actions of each sensor (0—don't track, 1—track left cell, 2—track right cell) are packed into a single integer a; the *i*th sensor's action is: $\left(\frac{a}{3^{7-i}}\right) mod3$
- *Rewards* ($[-8, 54]$): -1 for each sensor focus, $+30$ for eliminating a target

This problem has a relatively small state space compared to the action space. Furthermore, the problem involves multiple targets forcing the sensors to coordinate their actions.

4.3 Cat and mouse

There is a cat, a mouse, a piece of cheese as well as some obstacles in the cat and mouse world. The mouse tries to avoid getting caught by the cat, at the same time trying to get to the cheese. The mouse is the one learning, the cat is already programmed to go for the mouse. The obstacle layout is generated randomly. Both the cat and mouse have 8 degrees of movement: up, down, left and right, as well as the four diagonals. The mouse gets positive reward ($+50$) for getting the cheese. The mouse gets the cheese when it is in the same square as the cheese. The mouse gets negative reward (-100) for getting caught, by simply moving to the same square as the cat. The episode ends when the cat catches the mouse. Single cat and piece of cheese are in the maze at one time.

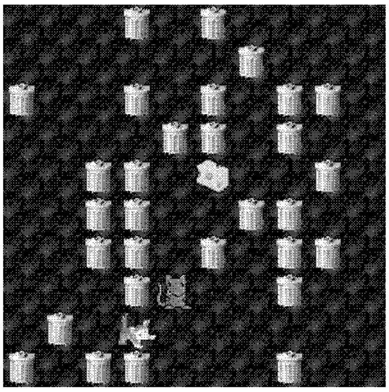


Figure 3. Cat and mouse game [1]

- *State space*: mouse position, cat position, cheese position
- *Starting states*: Mouse and cat start in a new random starting state at the beginning of each episode
- *Terminal states*: The episode ends when the cat catches the mouse
- *Actions*: Mouse has 8 degrees of movement: up, down, left and right, as well as the four diagonals
- *Rewards*: The mouse gets positive reward ($+50$) for getting the cheese. The mouse gets negative reward (-100) for getting caught

4.4 Blackjack

The problem is based on the Blackjack problem described in Sutton-Barto's book [14]. Standard rules of blackjack hold, but modified to allow players to (mistakenly) hit on 21. Episodes are prevented from starting in terminal states.

- *State space*: 1-dimensional integer array of 3 elements:
  - element[0] – current value of player's hand (4-21)
  - element[1] – value of dealer's face-up card (2-11)
  - element[2] – player does not have usable ace (0/1)
- Region for *starting states*: player has any 2 cards (uniformly distributed), dealer has any 1 card (uniformly distributed)
- *Terminal states*: Terminates when player "sticks" or is "bust"
- *Actions*: discrete integers 0-HIT, 1-STICK
- *Rewards*: -1 for a loss, 0 for a draw and 1 for a win

5 EXPERIMENTS

The above considered domains were used for testing the speed of convergence of the proposed methods in combination with Q-learning algorithm with different amount of information about mentor's decisions and their robustness to the quality of the mentor.

As we want to avoid expensive exploration we have also no chance to tune parameters and so we used the same parameters for all the domains: $\gamma = 0.9$, $\alpha = 0.1$ and $\epsilon = 0.1$.

We considered the rate of reduction in exploration on the basis of cumulative reward (the more cumulative reward is — the better reduction has been achieved) on episodes necessary for the methods to converge (10000 episodes were necessary for cat and mouse problem, and 1000 for the others). As mentor we chose the agent trained with the use of Q-learning algorithm for all the domains except blackjack (Q-learning algorithm failed to converge to optimal blackjack strategy [14] with the above parameters in 1000 episodes, so we used optimal blackjack player [14] as the mentor). We also initialized the corresponding functions so that they reflected random action choice in every state to research the robustness of our methods in case of a "poor mentor". For in general we don't possess information even about the order of the rewards in the environment, we used 0.1 as the initial value for functions in Q-initialization and reward shaping methods.

For gridworld with mines, DSN and cat and mouse problems the following comparison on the cumulative reward was held for every method proposed in subsections 3.1, 3.2 and 3.3:

- Q-learning — Q-learning algorithm presented in subsection 2.2
- 5 mentor episodes — current method with data on the mentor's state $\dashrightarrow$ action choice during 5 episodes
- 10 mentor episodes — current method with data on the mentor's state $\dashrightarrow$ action choice during 10 episodes
- 100 mentor episodes — current method with data on the mentor's state $\dashrightarrow$ action choice during 100 episodes
- 1000 mentor episodes — current method with data on the mentor's state $\dashrightarrow$ action choice during 1000 episodes
- Randomly initialized — the corresponding function was initialized so that it reflected random action choice in every state (to research the robustness of the current methods in case of a "poor mentor")

For blackjack the following comparison on the cumulative reward was held for every proposed method:

- Optimal policy — the optimal policy known for blackjack game [14] (player with this strategy was used as the mentor)
- Q-learning — Q-learning algorithm presented in subsection 2.2
- 5 mentor episodes — current method with data on the optimal player's state $\dashrightarrow$ action choice during 5 games
- 10 mentor episodes — current method with data on the optimal player's state $\dashrightarrow$ action choice during 10 games
- 100 mentor episodes — current method with data on the optimal player's state $\dashrightarrow$ action choice during 100 games
- 1000 mentor episodes — current method with data on the optimal player's state $\dashrightarrow$ action choice during 1000 games
- Randomly initialized — the corresponding function was initialized so that it reflected random action choice in every state (to research the robustness of the current methods in case of a "poor mentor")

As we consider the rate of reduction in exploration on the basis of cumulative reward on episodes necessary for the methods to converge the slope on such graphics corresponds to the quality of the learnt policy and the initial low values of cumulative reward reflect the errors in exploration (the initial bad quality of agent's decisions).

5.1 Gridworld with mines

- As we can see on figure 4 the agent with original Q-learning algorithm (without exploiting any expert knowledge) learnt good policy approximately after 200 episodes for gridworld with mines benchmark. Q-initialization method with any (even random) initialization acquired good policy also approximately after 200 episodes (the slopes of the graphics became the same as for Q-learning method). Even the usage of 5 mentor's episodes allowed to reduce the errors in exploration in the beginning but the agent stopped committing mistakes only when 100 mentor's episodes became available. And the more expert's decisions it used the better reduction in exploration it got (the larger the cumulative reward on figure 4). Agent with randomly initialized Q function at first committed many mistakes but nevertheless after 200 episodes it learnt a good policy.
- The results of testing reward shaping method on gridworld with mines domain is presented on the figure 5. As we can see only the usage of 1000 mentor episodes is comparable with initial Q-learning (these two graphics coincide on the figure). Embedding mentor's decisions in Q-learning algorithm with the use of reward shaping worsened the convergence of the original method. And the problem is not just with the exploration phase (the time till the agent learnt a good policy — corresponds to the point on the graphic where the cumulative reward starts to grow) but the agent also needed more episodes to learn it (the growth of corresponding cumulative reward curves comes later). Though as we can see on the figure 5 reward shaping with any initialization allowed to learn a good policy (at the end the slopes of the graphics are the same as for Q-learning method).

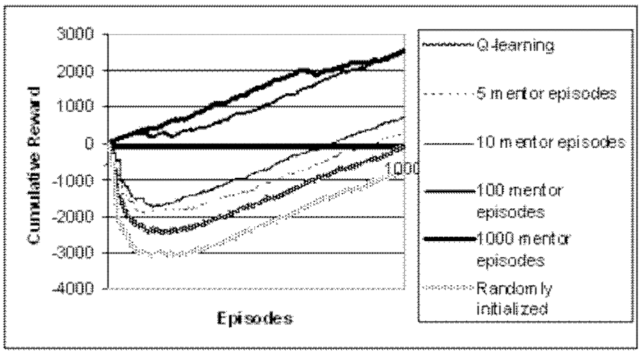


Figure 4. Gridworld with mines: cumulative reward of Q-initialization

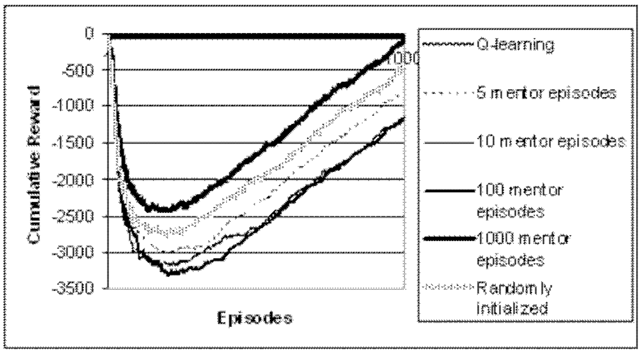


Figure 5. Gridworld with mines: cumulative reward of reward shaping

- As it can be seen on figure 6 usage of even a small number of expert's decisions (even 5 episodes) with two action-value function method prevented the agent from making mistakes. And the agent began to accumulate rewards from the very beginning. But as it can also be seen from figure 6 two action-value function method didn't show the robustness to the quality of mentor's decisions (the corresponding graphic of cumulative reward in case of randomly initialized value function goes inevitably down). So actually the agent that embedded decisions of poor agent with the help of Q-learning with two action-value functions couldn't learn a good policy at all. From figure 6 we can draw a conclusion that the more expert decisions we used the better reduction in exploration we got (the corresponding graphics of cumulative reward are higher).

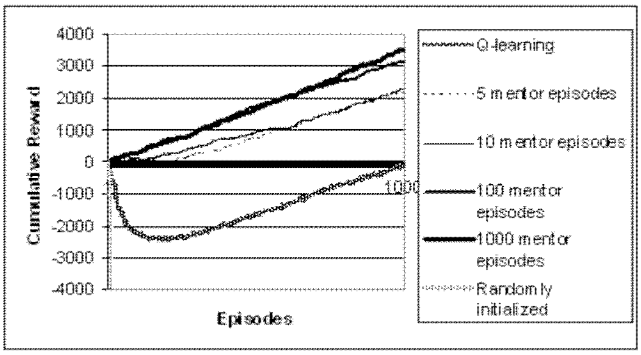


Figure 6. Gridworld with mines: cumulative reward of two functions

5.2 Distributed sensor network (DSN)

- As it can be seen from the figure 7 the more expert decisions we use for the Q-initialization method the better reduction in exploration we get for DSN domain.

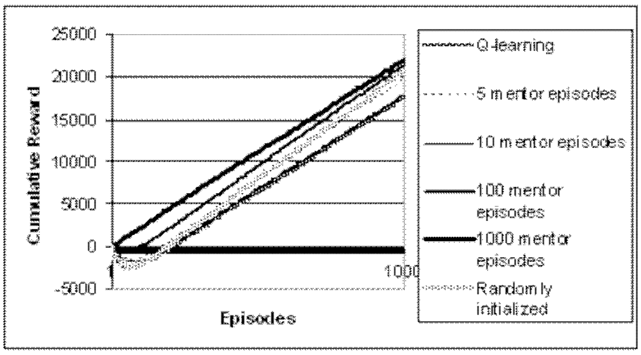


Figure 7. DSN: cumulative reward of Q-initialization

- The usage of any quantity of expert's decisions (see figure 8) didn't influence the convergence of reward shaping method for DSN benchmark.
- As it can be seen from figure 9 the usage of even a small number of expert's decisions with two action-value function method prevented the agent from making mistakes. But in case the mentor

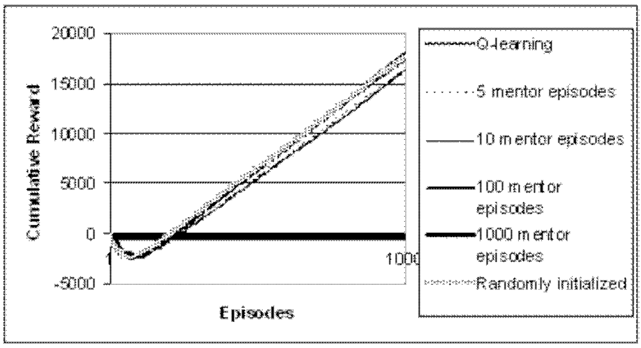


Figure 8. DSN: cumulative reward of reward shaping

itself doesn't know how to function successfully in the environment the consequences are devastating. As it ensues from the corresponding graphic two function method didn't allow the agent in this case to learn a good policy.

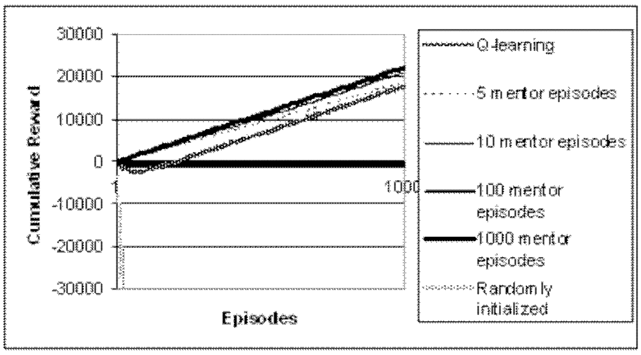


Figure 9. DSN: cumulative reward of two functions

5.3 Cat and mouse

- The figure 10 represents the results of testing Q-initialization method on cat and mouse benchmark. As we can see approximately after 5000 episodes Q-initialization as well as original Q-learning method allowed to learn a good policy (the slopes of the corresponding graphics after 5000 episodes are approximately the same). But usage of good mentor's decisions allowed to collect larger cumulative reward and in general the more mentor's episodes the agent observed the larger cumulative reward it got. In case of randomly initialized Q-learning function the cumulative reward is smaller then without any initialization at all (original Q-learning method).
- As it can be seen from figure 11 embedding mentor's decisions with the help of reward shaping method only worsened original Q-learning.
- We can draw a conclusion from figure 12 that the more data on mentor's decisions we used for two action-value function method the earlier the agent started to function effectively in cat and mouse benchmark. In case of a poor mentor the agent using this method had no chance to learn a good policy.

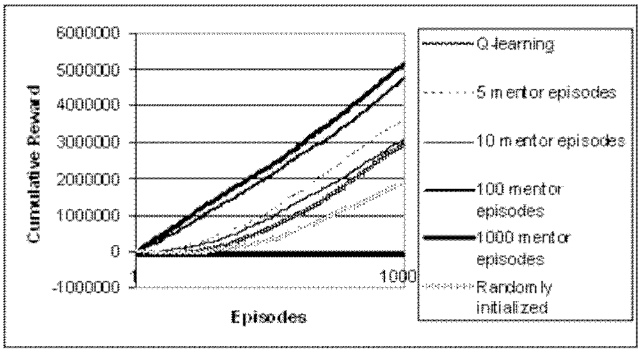


Figure 10. Cat and mouse: cumulative reward of Q-initialization

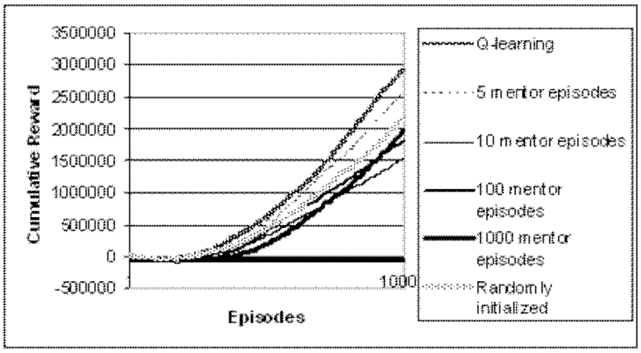


Figure 11. Cat and mouse: cumulative reward of reward shaping

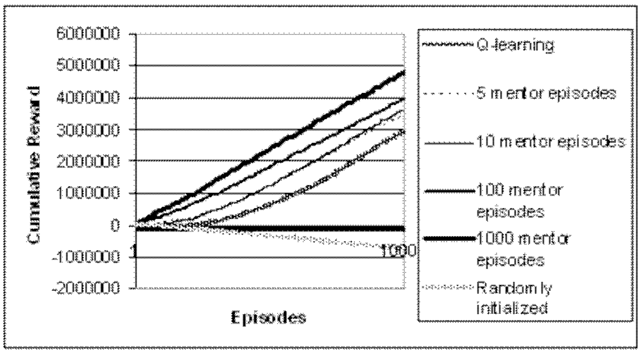


Figure 12. Cat and mouse: cumulative reward of two functions

5.4 Blackjack

When the agent can use decisions the optimal player took during 1000 games it learns optimal strategy either by Q-initialization or with two function method (see figures 13 and 15). On observing less number of games it can function only as well as Q-learning player.

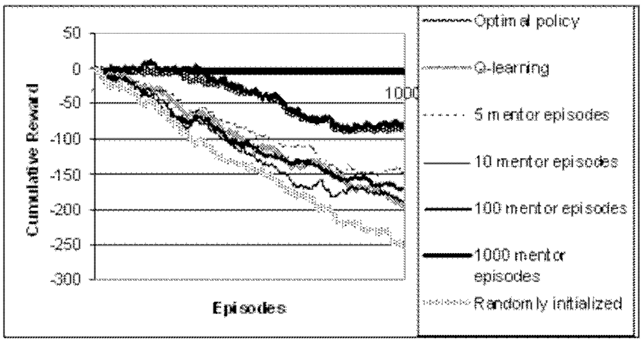


Figure 13. Blackjack: cumulative reward of Q-initialization

As in the above considered domains reward shaping method didn't yield any positive results for blackjack (see figure 14). Original Q-learning turned out to be better than reward shaping with any initialization.

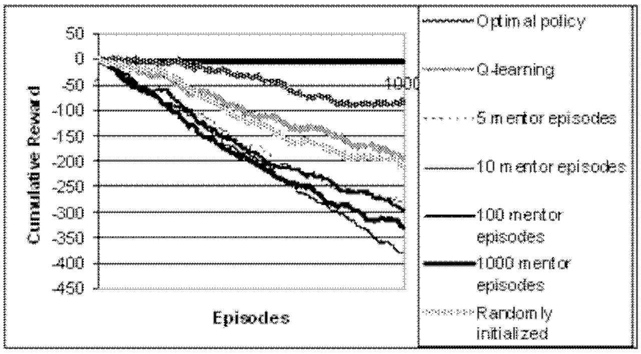


Figure 14. Blackjack: cumulative reward of reward shaping

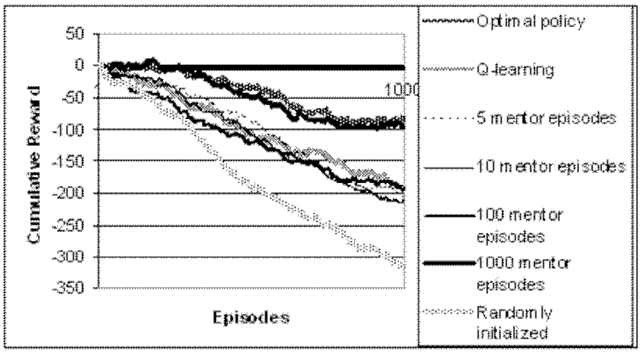


Figure 15. Blackjack: cumulative reward of two functions

6 DISCUSSION

Q-initialization and two action-value function methods proved to be useful for exploration reduction problem. The effect in reduction in case of Q-initialization got visible starting from 100 mentor episodes, while two function method started to show good results already with 5 available episodes. Meanwhile only Q-initialization showed good robustness to the quality of the mentor (that quite corresponds with theoretical results for Q-initialization method is the only one considered that is bound to converge to optimal policy for initial Q-learning algorithm converges to optimal strategy in case every state-action pair is visited infinitely often for any initial values of Q function). Reward shaping turned out to be too capricious (we tried it with different initialization values as well) but nothing yielded any good stable results. This though is quite in accordance with the first unsuccessful attempts of tuning extra reward function in reward shaping later successful applications [10, 12]. But our task differs from the one that was solved in [10, 12]. We are not interested in the solution however good it is if it requires tuning (tuning means a lot of exploration). What is clear is that reward shaping cant be seriously considered as general method for embedding expert's knowledge in reinforcement learning algorithms.

7 CONCLUSION

In this paper three methods: initialization of Q function, reward shaping and implementing mentor's decisions in a separate action-value function were proposed for an actual problem of reducing expensive exploration by means of using knowledge of already successfully functioning agent. Testing of these three methods in combination with Q-learning algorithm on four domains from the benchmarks for testing and comparing reinforcement learning algorithms "Reinforcement Learning Benchmarks and Bake-offs" has shown that reward shaping even in the best case requires too much tuning (that itself does need a lot of exploration) and quite inapplicable for the above problem. Q-initialization has proven to be robust to mentor's level of expertise, meanwhile the last method has shown better results in exploration reduction and requires less data on mentor's decisions.

ACKNOWLEDGEMENTS

We would like to thank the referees for their comments which helped improve this paper.

REFERENCES

[1] http://rlai.cs.ualberta.ca/rlbb/top.html.

[2] Natalia Akchurina and Hans Kleine Büning, 'Virtual markets: Q-learning sellers with simple state representation', *Proceedings of the Workshop on Autonomous Intelligent Systems: Agents and Data Mining (AIS-ADM 07) (to appear). Lecture Notes in Computer Science*, (2007).

[3] Syed Ali, Sven Koenig, and Milind Tambe, 'Preprocessing techniques for accelerating the dcop algorithm adopt', in *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pp. 1041–1048, New York, NY, USA, (2005). ACM Press.

[4] Richard C. Atkinson, Edward E. Smith, Daryl J. Bem, and Susan Nolen-Koeksema, *Hilgard's Introduction to Psychology*, Harcourt Brace College Publishers, 12'th edition, 1996.

[5] Amy R. Greenwald and Jeffrey O. Kephart, 'Shopbots and pricebots', in *Agent Mediated Electronic Commerce (IJCAI Workshop)*, pp. 1–23, (1999).

[6] Jeffrey O. Kephart, James E. Hanson, and Jakka Sairamesh, 'Price-war dynamics in a free-market economy of software agents', in *ALIFE: Proceedings of the sixth international conference on Artificial life*, pp. 53–62, Cambridge, MA, USA, (1998). MIT Press.

[7] Long-Ji Lin, 'Programming robots using reinforcement learning and teaching.', in *AAAI*, pp. 781–786, (1991).

[8] Long-Ji Lin, 'Self-improving reactive agents based on reinforcement learning, planning and teaching', *Machine Learning*, **8**(3-4), 293–321, (1992).

[9] Long-Ji Lin, *Reinforcement Learning for Robots using Neural Networks*, Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, 1993. CMU-CS-93-103.

[10] Maja J. Mataric, 'Reward functions for accelerated learning', in *International Conference on Machine Learning*, pp. 181–189, (1994).

[11] Tom Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[12] Jette Randlov and Preben Alstrom, 'Learning to drive a bicycle using reinforcement learning and shaping', in *Machine Learning: Proceedings of the Fifteenth International Conference (ICML'98)*, p. 117. MIT Press, (1998).

[13] Burrhus F. Skinner, *The Behavior of Organisms: An Experimental Analysis*, Prentice Hall, Englewood Cliffs, New Jersey, 1938.

[14] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*, The MIT Press, 1998.

[15] Ming Tan, 'Multi-agent reinforcement learning: Independent vs. cooperative learning', in *Readings in Agents*, eds., Michael N. Huhns and Munindar P. Singh, 487–494, Morgan Kaufmann, San Francisco, CA, USA, (1997).

[16] Gerald Tesauro, 'Temporal difference learning and td-gammon', *Commun. ACM*, **38**(3), 58–68, (1995).

[17] Gerald Tesauro, 'Pricing in agent economies using neural networks and multi-agent Q-learning', *Lecture Notes in Computer Science*, **1828**, 288–400, (2001).

[18] Steven D. Whitehead, 'A complexity analysis of cooperative mechanisms in reinforcement learning.', in *AAAI*, pp. 607–613, (1991).

[19] Steven D. Whitehead, *Reinforcement Learning for the Adaptive Control of Perception and Action*, Ph.D. dissertation, University of Rochester, 1992.

Shared Intentional Plans for Imitation and Cooperation: Integrating Clues from Child Development and Neurophysiology into Robotics

Peter Ford Dominey

Abstract. One of the long-term goals in the domain of human-robot interaction is that robots will approach these interactions equipped with some of the same fundamental cognitive capabilities that humans use. This will include the ability to perceive and understand human action in terms of an ultimate goal, and more generally to represent shared intentional plans in which the goal directed actions of the robot and the human are interlaced into a shared representation of how to achieve a common goal in a cooperative manner. The current research takes specific experimental protocols from studies of cognitive development to define behavior milestones for a perceptual-motor robotic system. Based on a set of previously established principals for defining the "innate" functions available to such a system, a cognitive architecture is developed that allows the robot to perform cooperative tasks at the level comparable to that of an 18 month old human child. Structural and functional properties of the primate neurophysiological mechanisms for action processing are used to provide further constraints on how the architecture is implemented. At the interface of cognitive development and robotics, the results on cooperation and imitation provide (1) a concrete demonstration of how cognitive neuroscience and developmental studies can contribute to human-robot interaction fidelity, and (2) a demonstration of how robots can be used to experiment with theories on the implementation of cognition in the developing human.

1. INTRODUCTION

One of the current open challenges in cognitive computational neuroscience is to understand the neural basis of the human ability to observe and imitate action. The results from such an endeavor can then be implemented and tested in robotic systems. Recent results from human and non-human primate behavior, neuroanatomy and neurophysiology provide a rich set of observations that allow us to constrain the problem of how imitation is achieved. The current research identifies and exploits constraints in these three domains in order to develop a system for goal directed action perception and imitation.

An impressive body of research exists on human imitation (62K responses to "human imitation" in Google Scholar), which has been empirically studied for over 100 years [15]. One of the recurrent findings across these studies is that in the context of goal directed action, it is the goal itself that tends to take precedence in defining what is to be imitated, rather than the means [1, 6, 25, 28, 29]. Of course in some situations it is the details (e.g. kinematics) of the movement itself that are to be imitated (see discussion in [6, 7]), but the current research focuses on goal based imitation. This body of research helped to formulate questions concerning what could be the neurophysiological substrates for goal based imitation. In 1992 di Pellegrino in the Rizzolatti lab [8] published the first results on "mirror" neurons, whose action potentials reflected both the production of specific goal-directed action, and the perception of the same action being carried by the experimenter. Since then, the premotor and parietal mirror system has been studied in detail in monkey (by single unit recording) and in man (by PET and fMRI) [see 25 for review].

In the context of understanding imitation, the discovery of the mirror system had an immense theoretical impact, as it provided justification for a common code for action production and perception. In recent years a significant research activity has used simulation and robotic platforms to attempt to link imitation behavior to the underlying neurophysiology at different levels of detail (see [24 and 27] for recent reviews from different perspectives, edited volumes [22, 23], and a dedicated special issue of Neural Networks [2]). Such research must directly address the question of how to determine what to imitate. Carpenter and Call [6] distinguish three aspects of the demonstration to copy: the physical action, the resulting change in physical state, and the inferred goal – the internal representation of the desired state. Here we concentrate on imitation of the goal, with the advantage of eliminating the difficulties of mapping detailed movement trajectories across the actor and imitator [7].

Part of the novelty of the current research is that it will explore imitation in the context of cooperative activity in which two agents act in a form of turn-taking sequence, with the actions of each one folding into an interleaved and coordinated intentional action plan. We use the term "shared intentional plan" to insist on the idea that multiple agents have a shared intention that will be realized through their use of a corresponding plan – the shared intentional plan. With respect to constraints derived from behavioral studies, we choose to examine child development studies, because such studies provide well-specified protocols that test behavior that is both relatively simple, and pertinent. The expectation is that a system that can account for this behavior should extend readily to more complex behavior, as demonstrated below.

Looking to the developmental data, Warneken, Chen and Tomasello [31] engaged 18-24 month children and young chimpanzees in goal-oriented tasks and social games which required cooperation. They were interested both in how the cooperation would proceed under optimal conditions, but also how the children and chimps would respond when the adult had a problem in performing the task. The principal finding was that children enthusiastically participate both in goal directed

P. F. Dominey is with the CNRS, 67 Bd Pinel 69675 Bron Cedex, France (phone: 33-437-911266; fax: 33-437-9112110; e-mail: dominey@ isc.cnrs.fr).

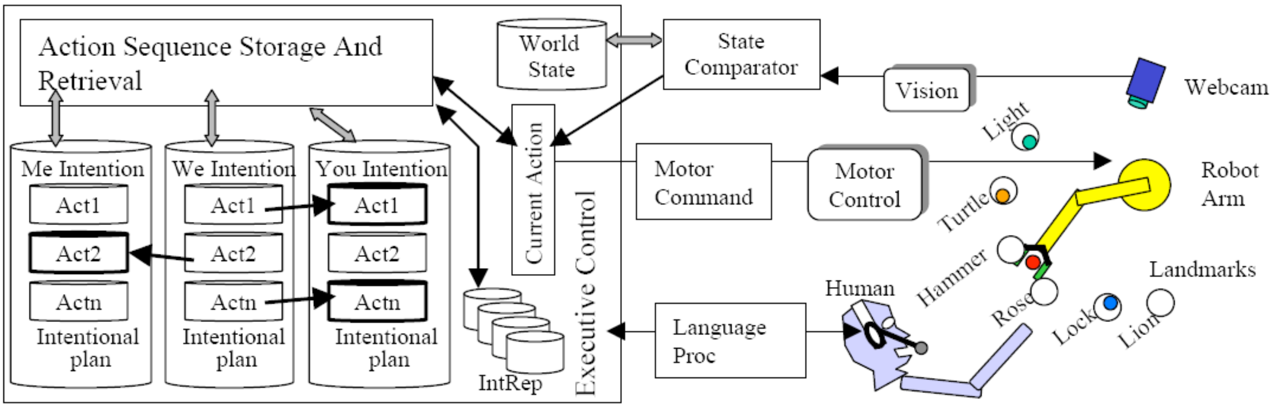


Fig 1. Cooperation System. In a shared work-space, human and robot manipulate objects (green, yellow, read and blue circles corresponding to dog, horse, pig and duck), placing them next to the fixed landmarks (light, turtle, hammer, etc.). *Action*: Spoken commands interpreted as individual words or grammatical constructions, and the command and possible arguments are extracted using grammatical constructions in Language Proc. The resulting Action(Agent, Object, Recipient) representation is the Current Action. This is converted into robot command primitives (Motor Command) and joint angles (Motor Control) for the robot. *Perception*: Vision provides object location input, allowing action to be perceived as changes in World State (State Comparator). Resulting Current Action used for action description, imitation, and cooperative action sequences. *Imitation*: The user performed action is perceived and encoded in Current Action, which is then used to control the robot under the supervision of Executive Control. *Cooperative Games*. During observations, individual actions are perceived, and attributed to the agent or the other player (Me or You). The action sequence is stored in the We Intention structure, that can then be used to separately represent self vs. other actions..

cooperative tasks and social games, and spontaneously attempt to reengage and help the adult when he falters. In contrast, chimps are uninterested in non-goal directed social games, and appear wholly fixed on attaining food goals, independent of cooperation. Warneken et al. thus observed what appears to be a very early human capacity for (1) actively engaging in cooperative activities for the sake of cooperation, and (2) for helping or reengaging the perturbed adult [30, 31].

In one of the social games, the experiment began with a demonstration where one participant sent a wooden block sliding down an inclined tube and the other participant caught the block in a tin cup that made a rattling sound. This can be considered more generally as a task in which one participant moves an object so that the second participant can then in turn manipulate the object. This represents a minimal case of a coordinated action sequence. After the demonstration, in Trials 1 and 2 the experimenter sent the block down one of the tubes three times, and then switched to the other, and the child was required to choose the same tube as the partner. In Trials 3 and 4 during the game, the experimenter interrupted the behavior for 15 seconds and then resumed.

Behaviorally, children successfully participated in the game in Trials 1 and 2. In the interruption Trials 3 and 4 they displayed two particularly interesting types of response that were (a) to attempt to perform the role of the experimenter themselves, and/or (b) to reengage the experimenter with a communicative act. This indicates that the children had a clear awareness both of their role and that of the adult in the shared coordinated activity. This research thus identifies a set of behavioral objectives for robot behavior in the perception and execution of cooperative intentional action. Such behavior could, however, be achieved in a number of possible architectures.

In order to begin to constrain the space of possible solutions we can look to recent results in human and primate neurophysiology and neuroanatomy. It has now become clearly established that neurons in the parietal cortex and the premotor cortex encode the goal of simple actions both for the execution of these actions as well as for the perception of these same goal-directed actions when performed by a second agent [8, 25]. This research thus corroborates the emphasis from behavioral studies on the importance of the goal (rather than the details of the means) in action perception [1, 6, 25, 28, 29]. It has been suggested that these “mirror” neurons play a crucial role in imitation, as they provide a common representation for the perception and subsequent execution of a given action. Interestingly, however, it has been clearly demonstrated that the imitation ability of non-human primates is severely impoverished when compared to that of humans [25, 29-31]. This indicates that the human ability to imitate novel actions and action sequences in real time (i.e. after only one or two demonstrations) relies on additional neural mechanisms.

In this context, a recent study of human imitation learning [5] implicates Brodmann’s area (BA) 46 as responsible for orchestrating and selecting the appropriate actions in novel imitation tasks. We have recently proposed that BA 46 participates in a dorsal stream mechanism for the manipulation of variables in abstract sequences and language [14]. Thus, variable “slots” that can be instantiated by arbitrary motor primitives during the observation of new behavior sequences, are controlled in BA 46, and their sequential structure is under the control of corticostriatal systems which have been clearly implicated in sensorimotor sequencing (see [14]). This allows us to propose that this evolutionarily more recent cortical area BA 46 may play a crucial role in allowing humans to perform compositional operations (i.e. sequence learning) on more primitive action representations in the ventral premotor and parietal motor cortices. In other words, ventral premotor and parietal cortices instantiate shared perceptual and motor representations of atomic actions, and BA46 provides the capability to compose arbitrary sequences of these atomic actions, while relying on well known corticostriatal neurophysiology for sequence storage and retrieval. The functional result is the human ability to observe and represent novel behavioral action sequences. We further claim that this system can represent behavioral sequences from the “bird’s eye view” or third person perspective, as required for the cooperative tasks of Warneken et al. [31]. That is, it can allow one observer to perceive and form an integrated representation of the coordinated actions of two other agents engaged in a cooperative activity. The observer can then use this representation to step in and play the role of either of the two agents.

2. IMPLEMENTATION

In a comment on Tomasello et al [29] on understanding and sharing intention, Dominey [10] analyses how a set of initial capabilities can be used to provide the basis for shared intentions. This includes capabilities to

1. perceive the physical states of objects,
2. perceive (and perform) actions that change these states,
3. distinguish between self and other,
4. perceive emotional/evaluation responses in others, and
5. learn sequences of predicate-argument representations.

The goal is to demonstrate how these 5 properties can be implemented within the constraints of the neurophysiology data reviewed above in order to provide the basis for performing these cooperative tasks. In the current experiments the human and robot cooperate by moving physical objects to different positions in a shared work-space as illustrated in Figures 1 and 2. The 4 moveable objects are pieces of a wooden puzzle, representing a dog, a pig, a duck and a cow. These pieces can be moved by the robot and the user in the context of cooperative activity. Each has fixed to it a vertically protruding metal screw, which provides an easy grasping target both for the robot and for humans. In addition there are 6 images that are fixed to the table and serve as landmarks for placing the moveable objects, and correspond to a light, a turtle, a hammer, a rose, a lock and a lion, as partially illustrated in Figures 1 & 2. In the interactions, human and robot are required to place objects in zones next to the different landmarks, so that the robot can more easily determine where objects are, and where to grasp them. Figure 1 provides an overview of the architecture, and Figure 2, which corresponds to Experiment 6 provides an overview of how the system operates.

2.1 Representation

The structure of the internal representations is a central factor determining how the system will function, and how it will generalize to new conditions. Based on the neurophysiology reviewed above, we use a common representation of action for both perception and production. Actions are identified by the agent, the object, and the target location to move that object to. As illustrated in Figure 1, by taking the short loop from vision, via Current Action Representation, to Motor Command, the system is thus configured for a form of goal-centered action imitation. This will be expanded upon below.

A central feature of the system is the World Model that represents the physical state of the world, and can be accessed and updated by vision, motor control, and language, similar to the Grounded Situation Model of [21]. The World Model encodes the physical locations of objects that is updated by vision and proprioception (i.e. robot action updates World Model with new object location). Changes in the World Model in terms of an object being moved allows the system to detect actions in terms these object movements. Actions are represented in terms of the agent, the object and the goal of the action, in the form MOVE(object, goal location, agent). These representations can be used for commanding action, for describing recognized action, and thus for action imitation and narration, as seen below.

In order to allow for more elaborate cooperative activity, the system must be able to store and retrieve actions in a sequential structure. This form of real time sequence learning for imitation is not observed in non-human primates. Interestingly, in this context, an fMRI study [5] that addressed the human ability to observe and program arbitrary actions indicated that a cortical area (BA46) which is of relatively recent phylogenetic origin is involved in such processes. Rizzolatti and Craighero [25] have thus suggested that the BA 46 in man will orchestrate allow the real-time capability to store and retrieve recognized actions, and we can further propose that this orchestration will recruit canonical brain circuitry for sequence processing including the cortico-striatal system (see [14] for discussion of such sequence processing).

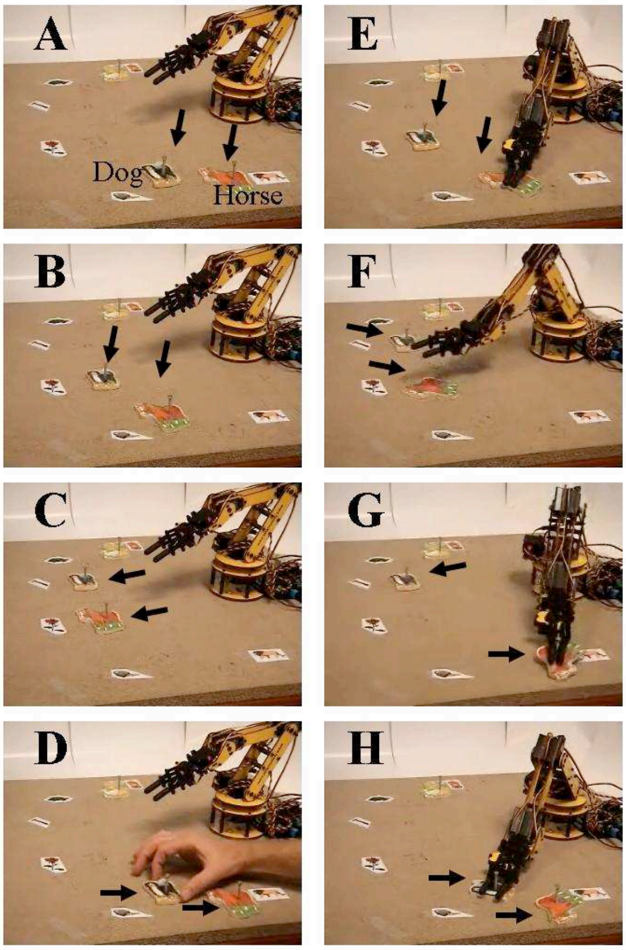


Figure 2. Cooperative task of Exp 5-6. Robot arm, with 6 landmarks (Light, turtle, hammer, rose, lock and lion from top to bottom). Moveable objects include Dog and Horse. In A-D, human demonstrates a "horse chase the dog" game, and successively moves the Dog then Horse, indicating that in the game, the user then the robot are agents, respectively. After demonstration, human and robot "play the game". In each of E – F user moves Dog, and robot follows with Horse. In G robot moves horse, then in H robot detects that the user is having trouble and so "helps" the user with the final move of the dog. See Exp 5 & 6.

In the current study we address behavioral conditions in which focus on the observation and immediate re-use of an intentional (goal directed) action plan. However, in the more general case, one should consider that multiple intentional action plans can be observed and stored in a repertory (IntRep or Intentional Plan Repertory in Figure 1). When the system is subsequently observing the behavior of others, it can compare the ongoing behavior to these stored sequences. Detection of a match with the beginning of a stored sequence can be used to retrieve the entire sequence. This can then be used to allow the system to "jump into" the scenario, to anticipate the other agent's actions, and/or to help that agent if there is a problem.

2.2 Visual perception

Visual perception is a challenging technical problem. To simplify, standard lighting conditions and a small set (n = 10) of visual object to recognize are employed (4 moveable objects and 6 location landmarks). A VGA webcam is positioned at 1.25 meters above the robot workspace. Vision processing is provided by the Spikenet Vision System (http://www.spikenet-technology.com/). Three recognition models for each object at different orientations (see Fig. 3) were built with an offline model builder. During real-time vision processing, the models are recognized, and their (x, y) location in camera coordinates are provided. Our vision post-processing eliminates spurious detections and returns the reliable (x, y) coordinates of each moveable object in a file. The nearest landmark is then calculated.

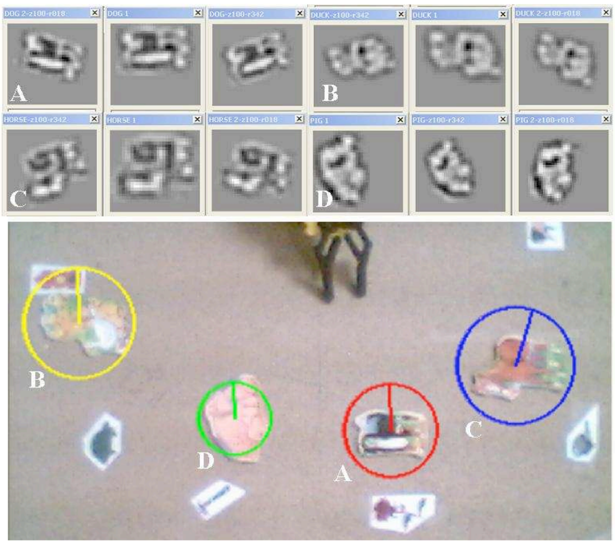


Figure 3. Vision processing. Above: A. – D. Three templates each for the Dog, Duck, Horse and Pig objects at three different orientations. Below, encompassing circles indicate template recognition for the four different objects near different fixed landmarks, as seen from the camera over the robot workspace

2.3 Motor Control & Visual-Motor Coordination

While visual-motor coordination is not the focus of the current work, it was necessary to provide some primitive functions to allow goal directed action. All of the robot actions, whether generated in a context of imitation, spoken command or cooperative interaction will be of the form *move(x to y)* where *x* is a member of a set of visually perceivable objects, and *y* is a member of the set of fixed locations on the work plan.

Robot motor control for transport and object manipulation with a two finger gripper is provided by the 6DOF Lynx6 arm (www.lynxmotion.com). The 6 motors of the arm are coordinated by a parallel controller connected to a PC computer that provides transmission of robot commands over the RS232 serial port.

Human users (and the robot) are constrained when they move an object, to place it in one of the zones designated next to each of the six landmarks (see Fig 3). This way, when the nearest landmark for an object has been determined, this is sufficient for the robot to grasp that object at the prespecified zone.

In a calibration phase, a target point is marked next to each of the 6 fixed landmark locations, such that they are all on an arc that is equidistant to the center of rotation of the robot arm base. For each, the rotation angle of Joint 0 (the rotating shoulder base) necessary to align the arm with that point is then determined, along with a common set of joint angles for Joints 1 – 5 that position the gripper to seize any of the objects. Angles for Joint 6 that controls the closing and opening of the gripper to grasp and release an object were then identified. Finally a neutral position to which the arm could be returned in between movements was defined. The system was thus equipped with a set of primitives that could be combined to position the robot at any of the 6 grasping locations, grasp the corresponding object, move to a new position, and place the object there.

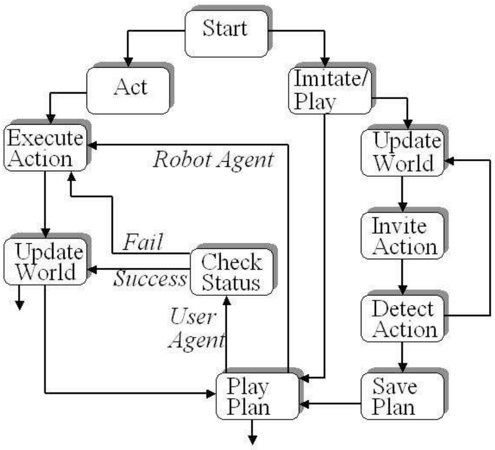


Figure 4. Spoken Language Based Cooperation flow of control. Interaction begins with proposal to act, or imitate/play a game. Act – user says an action that is verified and executed by robot. World Model updated based on action. Downward arrow indicates return to Start. Imitate/Play – user demonstrates actions to robot and says who the agent should be when the game is to be played (e.g. "You/I do this"). Each time, system checks the state of the world, invites the next action and detects the action based on visual object movement. When the demo is finished, the plan (of a single item in the case of imitation) is stored and executed (Play Plan). If the user is the agent (encoded as part of the game sequence), system checks execution status and helps user if failure. If robot is agent, system executes action, and then moves on to next item.

2.4 Cooperation Control Architecture

The spoken language control architecture illustrated in Fig 4 is implemented with the CSLU Rapid Application Development toolkit (http://cslu.cse.ogi.edu/toolkit/). This system provides a state-based dialog management system that allows interaction with the robot (via the serial port controller) and with the vision processing system (via file i/o). It also provides the spoken language interface that allows the user to determine what mode of operation he and the robot will work in, and to manage the interaction via spoken words and sentences.

Figure 4 illustrates the flow of control of the interaction management. In the Start state the system first visually observes where all of the objects are currently located. From the start state, the system allows the user to specify if he wants to ask the robot to perform actions (Act), to imitate the user, or to play (Imitate/Play). In the Act state, the user can specify actions of the form "Put the dog next to the rose" and a grammatical construction template [9, 11-14] is used to extract the action that the robot then performs.

2.5 Imitation and Learning Shared Intentional Plans

In the Imitate state, the robot first verifies the current state (Update World) and then invites the user to demonstrate an action (Invite Action). The user shows the robot one action. The robot re-observes the world and detects the action based on changes detected (Detect Action). In particular, it will observe that an object has been moved to a new location. This corresponds to the action of moving the object to that location. This action is then saved and transmitted (via Play the Plan with Robot as Agent) to execution (Execute action). A *predicate(argument)* representation of the form Move(object, landmark) is used both for action observation and execution, thus radically simplifying the correspondence problem [see 27]. Imitation is thus a minimal case of Playing in which the "game" is a single action executed by the robot.

In the more general case, the robot should learn to play a game that involves a succession of moves executed by the user and robot in a specific turn-taking sequence. For a given game, the user can demonstrate multiple successive actions, and indicate the agent - by saying "You/I do this" - for each action. The resulting intentional plan specifies what is to be done by whom. When the user specifies that the plan is finished, the system moves to the Save Plan. In this state, the system stores the shared intentional plan, consisting of a sequence of actions and a specification of the agent for each of these action. Control then moves on to the Play Plan state. For each action, the system recalls whether it is to be executed by the robot or the user. Robot execution takes the standard Execute Action pathway. User execution performs a check (based on user response) concerning whether the action was correctly performed or not. If the user action is not performed, then the robot communicates with the user, and performs the action itself. Thus, "helping" was implemented by combining an evaluation of the user action, with the existing capability to perform a stored action representation.

Once the shared intentional plan has been stored or "learned" it can then be re-used in the future. This, when entering the Imitate/Play state, the user is given the option of playing the most recently learned game, or learning a new one.

3. EXPERIMENTAL RESULTS

For each of the 6 following experiments, equivalent variants were repeated at least ten times to demonstrate the generalized capability and robustness of the system. In less than 5 percent of the trials, errors of two types were observed to occur. Speech errors resulted from a failure in the voice recognition, and were recovered from by the command validation check (Robot: "Did you say …?"). Visual image recognition errors occurred when the objects were rotated beyond 20° from their upright position. These errors were identified when the user detected that an object that should be seen was not reported as visible by the system, and were corrected by the user re-placing the object and asking the system to "look again". At the beginning of each trial the system first queries the vision system, and updates the World Model with the position of all visible objects. It then informs the user of the locations of the different objects, for example "The dog is next to the lock, the horse is next to the lion." It then asks the user "Do you want me to act, imitate, play or look again?", and the user responds with one of the action-related options, or with "look again" if the scene is not described correctly.

3.1 Experiment 1: Validation of Sensorimotor Control

In this experiment, the user says that he wants the "Act" state (Fig 4), and then uses spoken commands such as "Put the horse next to the hammer". Recall that the horse is among the moveable objects, and hammer is among the fixed landmarks. The robot requests confirmation and then extracts the predicate-argument representation - *Move(X to Y)* - of the sentence based on grammatical construction templates. In the Execute Action state, the action *Move(X to Y)* is decomposed into two movement primitives [27] of *Get(X)*, and *Place-At(Y)*. *Get(X)* queries the World Model in order to localize X with respect to the different landmarks, and then performs a grasp at the corresponding landmark target location. Likewise, *Place-At(Y)* simply performs a transport to target location Y and releases the object. Decomposing the *get* and *place* functions allows the composition of all possible combinations in the *Move(X to Y)* space. Ten trials were performed moving the four objects to and from different landmark locations. Experiment 1 thus demonstrates (1) the ability to transform a spoken sentence into a Move(X to Y) command, (2) the ability to perform visual localization of the target object, and (3) the sensory-motor ability to grasp the object and put it at the specified location. In ten experimental runs, the system performed correctly.

3.2 Experiment 2: Imitation

In this experiment the user chooses the "imitate" state. As stated above, imitation is centered on the achieved ends – in terms of observed changes in state – rather than the means towards these ends. Before the user performs the demonstration of the action to be imitated, the robot queries the vision system, and updates the World Model (Update World in Fig 4) and then invites the user to demonstrate an action. The robot pauses, and then again queries the vision system and continues to query until it detects a difference between the currently perceived world state and the previously stored World Model (in State Comparator of Fig 1, and Detect Action in Fig 4), corresponding to an object displacement. Extracting the identity of the displaced object, and its new location (with respect to the nearest landmark) allows the formation of an *Move(object, location)* action representation. Before imitating, the robot operates on this representation with a meaning-to-sentence construction in order to verify the action to the user, as in "Did you put the dog next to the rose?" It then asks the user to put things back as they were so that it can perform the imitation. At this point, the action is executed (Execute Action in Fig 4). In ten experimental runs the system performed correctly. This demonstrates (1) the ability of the system to detect the goals of user-generated actions based on visually perceived state changes, and (2) the utility of a common representation of action for perception, description and execution.

3.3 Experiment 3: A Cooperative Game

The cooperative game is similar to imitation, except that there is a sequence of actions (rather than just one), and the actions can be effected by either the user or the robot in a cooperative manner. In this experiment, the user responds to the system request and enters the "play" state. In what corresponds to the demonstration in Warneken et al. [17] the robot invites the user to start showing how the game works. The user then begins to perform a sequence of actions. For each action, the user specifies who does the action, i.e. either "you do this" or "I do this". The intentional plan is thus stored as a sequence of action-agent pairs, where each action is the movement of an object to a particular target location. In Fig 1, the resulting interleaved sequence is stored as the "We intention", i.e. an action sequence in which there are different agents for different actions. When the user is finished he says "play the game". The robot then begins to execute the stored intentional plan. During the execution, the "We intention" is decomposed into the components for the robot (Me Intention) and the human (You intention).

In one run, during the demonstration, the user said "I do this" and moved the horse from the lock location to the rose location. He then said "you do this" and moved the horse back to the lock location. After each move, the robot asks "Another move, or shall we play the game?". When the user is finished demonstrating the game, he replies "Play the game." During the playing of this game, the robot announced "Now user puts the horse by the rose". The user then performed this movement. The robot then asked the user "Is it OK?" to which the user replied "Yes". The robot then announced "Now robot puts the horse by the lock" and performed the action. In two experimental runs of different demonstrations, and 5 runs each of the two demonstrated games, the system performed correctly. This demonstrates that the system can learn a simple intentional plan as a stored action sequence in which the human and the robot are agents in the respective actions.

Action	User identifies agent	User Demonstrates Action	Ref in Figure 2
1.	I do this	Move dog from the lock to the rose	B
2.	You do this	Move the horse from the lion to the lock	B
3.	I do this	Move the dog from the rose to the hammer	C
4.	You do this	Move the horse from the lock to the rose	C
5.	You do this	Move the horse from the rose to the lion	D
6.	I do this	Move the dog from the hammer to the lock	D

Table 1. Cooperative "horse chase the dog" game specified by the user in terms of who does the action (indicated by saying) and what the action is (indicated by demonstration). Illustrated in Figure 2.

3.4 Experiment 4: Interrupting a Cooperative Game

In this experiment, everything proceeds as in experiment 3, except that after one correct repetition of the game, in the next repetition, when the robot announced "Now user puts the horse by the rose" the user did nothing. The robot asked "Is it OK" and during a 15 second delay, the user replied "no". The robot then said "Let me help you" and executed the move of the horse to the rose. Play then continued for the remaining move of the robot. This illustrates how the robot's stored representation of the action that was to be performed by the user allowed the robot to "help" the user.

3.5 Experiment 5: A More Complex Game

Experiment 3 represented the simplest behavior that could qualify as a cooperative action sequence. In order to more explicitly test the intentional sequencing capability of the system, this experiment replicates Exp 3 but with a more complex task, illustrated in Figure 2. In this game (Table 1), the user starts by moving0 the dog, and after each move the robot "chases" the dog with the horse, till they both return to their starting places.

As in Experiment 3, the successive actions are visually recognized and stored in the shared "We Intention" representation. Once the user says "Play the game", the final sequence is stored, and then during the execution, the shared sequence is decomposed into the robot and user components based on the agent associated with each action. When the user is the agent, the system invites the user to make the next move, and verifies (by asking) if the move was ok. When the system is the agent, the robot executes the movement. After each move the World Model is updated. As in Exp 3, two different complex games were learned, and each one "played" 5 times. This illustrates the learning by demonstration [31] of a complex intentional plan in which the human and the robot are agents in a coordinated and cooperative activity.

3.6 Experiment 6: Interrupting the Complex Game

As in Experiment 4, the objective was to verify that the robot would take over if the human had a problem. In the current experiment this capability is verified in a more complex setting. Thus, when the user is making the final movement of the dog back to the "lock" location, he fails to perform correctly, and indicates this to the robot. When the robot detects failure, it reengages the user with spoken language, and then offers to fill in for the user. This is illustrated in Figure 2H. This demonstrates the generalized ability to help that can occur whenever the robot detects the user is in trouble.

4. DISCUSSION

Significant progress has been made in identifying some of the fundamental characteristics of human cognition in the context of cooperative interaction, particularly with respect to social cognition [16-19]. Breazeal and Scassellati [4] investigate how perception of socially relevant face stimuli and object motion will both influence the emotional and attentional state of the system and thus the human-robot interaction. Scassellati [26] further investigates how developmental theories of human social cognition can be implemented in robots. In this context, Kozima and Yano [18] outline how a robot can attain intentionality – the linking of goal states with intentional actions to achieve those goals – based on innate capabilities including: sensory-motor function and a simple behavior repertoire, drives, an evaluation function, and a learning mechanism.

The abilities to observe an action, determine its goal and attribute this to another agent are all clearly important aspects of the human ability to cooperate with others. The current research demonstrates how these capabilities can contribute to the "social" behavior of learning to play a cooperative game, playing the game, and helping another player who has gotten stuck in the game, as displayed in 18-24 month children [29, 30]. While the primitive bases of such behavior is visible in chimps, its full expression is uniquely human [29, 30]. As such, it can be considered a crucial component of human-like behavior for robots.

The current research is part of an ongoing effort to understand aspects of human social cognition by bridging the gap between cognitive neuroscience, simulation and robotics [3, 9-14], with a focus on the role of language (see [20]). The experiments presented here indicate that functional requirements derived from human child behavior and neurophysiological constraints can be used to define a system that displays some interesting capabilities for cooperative behavior in the context of imitation. Likewise, they indicate that evaluation of another's progress, combined with a representation of his/her failed goal provides the basis for the human characteristic of "helping." This may be of interest to developmental scientists, and the potential collaboration between these two fields of cognitive robotics and human cognitive development is promising. The developmental cognition literature lays out a virtual roadmap for robot cognitive development [10, 28]. In this context, we are currently investigating the development of hierarchical means-end action sequences [27]. At each step, the objective will be to identify the behavior characteristic and to implement it in the most economic manner in this continuously developing system for human-robot cooperation.

At least two natural extensions to the current system can be considered. The first involves the possibility for changes in perspective. In the experiments of Warneken et al. the child watched two adults perform a coordinated task (one adult launching the block down the tube, and the other catching the block). At 24 months, the child can thus observe the two roles being played out, and then step into either role. This indicates a "bird's eye view" representation of the cooperation, in which rather than assigning "me" and "other" agent roles from the outset, the child represents the two distinct agents A and B for each action in the cooperative sequence. Then, once the perspective shift is established (by the adult taking one of the roles, or letting the child choose one) the roles A and B are assigned to me and you (or vice versa) as appropriate.

This actually represents a minimal change to our current system. First, rather than assigning the "you" "me" roles in the We Intention at the outset, these should be assigned as A and B. Then, once the decision is made as to the mapping of A and B onto robot and user, these agent values will then be assigned accordingly. Second, rather than having the user tell the robot "you do this" and "I do this" the vision system can be modified to recognize different agents who can be identified by saying their name as they act, or via visually identified cues on their acting hands.

The second issue has to do with inferring intentions. The current research addresses one cooperative activity at a time, but nothing prevents the system from storing multiple such intentional plans in a repertory (IntRep in Fig 1). In this case, as the user begins to perform a sequence of actions involving himself and the robot, the robot can compare this ongoing sequence to the initial subsequences of all stored sequences in the IntRep. In case of a match, the robot can retrieve the matching sequence, and infer that it is this that the user wants to perform. This can be confirmed with the user and thus provides the basis for a potentially useful form of learning for cooperative activity.

In conclusion, the current research has attempted to build and test a robotic system for interaction with humans, based on behavioral and neurophysiological requirements derived from the respective literatures. The interaction involves spoken language and the performance and observation of actions in the context of cooperative action. The experimental results demonstrate a rich set of capabilities for robot perception and subsequent use of cooperative action plans in the context of human-robot cooperation. This work thus extends the imitation paradigm into that of sequential behavior, in which the learned intentional action sequences are made up of interlaced action sequences performed in cooperative alternation by the human and robot. While many technical aspects of robotics (including visuomotor coordination and vision) have been simplified, it is hoped that the contribution to the study of imitation and cooperative activity is of some value.

5. ACKNOWLEDGEMENTS

I thank Mike Tomasello, Felix Warneken, Malinda Carpenter and Elena Lieven for useful discussions during a visit to the MPI EVA in Leipzig concerning shared intentions; and Giacomo Rizzolatti for insightful discussion concerning the neurophysiology of sequence imitation at the IEEE Humanoids meeting in Genoa 2006. This research is supported in part by the French Minister of Research under grant ACI-TTT, ACI-NIC and by the LAFMI.

6. REFERENCES

[1] Bekkering H, WohlschlagerA , Gattis M (2000) Imitation of Gestures in Children is Goal-directed, The Quarterly Journal of Experimental Psychology: Section A, 53, 153-164

[2] Billard A, Schaal (2006) Special Issue: The Brain Mechanisms of Imitation Learning, Neural Networks, 19(1) 251-338

[3] Boucher J-D, Dominey PF (2006) Programming by Cooperation: Perceptual-Motor Sequence Learning via Human-Robot Interaction, Proc. Simulation of Adaptive Behavior, Rome 2006.

[4] Breazeal C., Scassellati B., (2001) Challenges in building robots that imitate people, in: K. Dautenhahn, C. Nehaniv (Eds.), Imitation in Animals and Artifacts, MIT Press, Cambridge, MA,.

[5] Buchine G, Vogt S, Ritzl A, Fink GR, Zilles K, Freund H-J, Rizzolatti G (2004) Neural circuits Underlying Imitation Learning of Hand Actions: An Event-Related fMRI Study. Neuron, (42) 323-334.

[6] Carpenter M, Call Josep (2007) The question of 'what to imitate': inferring goals and intentions from demonstrations, in Chrystopher L. Nehaniv and Kerstin Dautenhahn Eds, *Imitation and Social Learning in Robots, Human sand Animals*, Cambridge Univerity Press, Cambridge.

[7] Cuijpers RH, van Schie HT, Koppen M, Erlhagen W, Bekkering H (2006) Goals and means in action observation: A computational approach, *Neural Networks* 19, 311-322,

[8] di Pellegrino G, Fadiga L, Fogassi L, Gallese V, Rizzolatti G (1992) Understanding motor events: a neurophysiological study. *Exp Brain Res.*;91(1):176-80.

[9] Dominey, P.F., (2003) Learning grammatical constructions from narrated video events for human–robot interaction. *Proceedings IEEE Humanoid Robotics Conference*, Karlsruhe, Germany

[10] Dominey PF (2005) Toward a construction-based account of shared intentions in social cognition. Comment on Tomasello et al. 2005, *Beh Brain Sci.* 28:5, p. 696.

[11] Dominey PF, Alvarez M, Gao B, Jeambrun M, Weitzenfeld A, Medrano A (2005) Robot Command, Interrogation and Teaching via Social Interaction, *Proc. IEEE Conf. On Humanoid Robotics 2005.*

[12] Dominey PF, Boucher (2005) Learning To Talk About Events From Narrated Video in the Construction Grammar Framework, *Artificial Intelligence*, 167 (2005) 31–61

[13] Dominey, P. F., Boucher, J. D., & Inui, T. (2004). Building an adaptive spoken language interface for perceptually grounded human–robot interaction. In *Proceedings of the IEEE-RAS/RSJ international conference on humanoid robots.*

[14] Dominey PF, Hoen M, Inui T. (2006) A neurolinguistic model of grammatical construction processing. *Journal of Cognitive Neuroscience*.18(12):2088-107.

[15] Ellwood CA (1901) The Theory of Imitation in Social Psychology *The American Journal of Sociology*, Vol. 6, No. 6 (May, 1901), pp. 721-741

[16] Fong T, Nourbakhsh I, Dautenhaln K (2003) A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42 3-4, 143-166.

[17] Goga, I., Billard, A. (2005), Development of goal-directed imitation, object manipulation and language in humans and robots. In M. A. Arbib (ed.), *Action to Language via the Mirror Neuron System*, Cambridge University Press (in press).

[18] Kozima H., Yano H. (2001) A robot that learns to communicate with human caregivers, in: *Proceedings of the International Workshop on Epigenetic Robotics*,.

[19] Lieberman MD (2007) Social Cognitive neuroscience: A Review of Core Processes, *Annu. Rev. Psychol.* (58) 18.1-18.31

[20] Lauria S, Buggmann G, Kyriacou T, Klein E (2002) Mobile robot programming using natural language. *Robotics and Autonomous Systems* 38(3-4) 171-181

[21] Mavridis N, Roy D (2006). Grounded Situation Models for Robots: Where Words and Percepts Meet. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*

[22] Nehaniv CL, Dautenhahn K eds. (2002) Imitation in Animals and Artifacts; MIT Press, Cambridge MA.

[23] Nehaniv CL, Dautenhahn K eds. (2007) *Imitation and Social Learing in Robots, Humans and Animals*, Cambridge University Press, Cambridge.

[24] Oztop E, Kawato M, Arbib M (2006) Mirror neurons and imitation: A computationally guided review. *Neural Networks*, (19) 254-271

[25] Rizzolatti G, Craighero L (2004) The Mirror-Neuron system, *Annu. Rev. Neuroscience* (27) 169-192

[26] Scassellati B (2002) Theory of mind for a humanoid robot, *Autonomous Robots*, 12(1) 13-24

[27] Schaal S, Ijspeert A, Billard A (2003) Computational Approaches to Motor Learning by Imitation; Phil Trans Royal Society London/ B, 358: 537-547.

[28] Sommerville A, Woodward AL (2005) Pulling out the intentional structure of action: the relation between action processing and action production in infancy. *Cognition*, 95, 1-30.

[29] Tomasello M, Carpenter M, Cal J, Behne T, Moll HY (2005) Understanding and sharing intentions: The origins of cultural cognition, *Beh. Brain Sc*;. 28; 675-735.

[30] Warneken F, Tomasello M (2006) Altruistic helping in human infants and young chimpanzees, *Science*, 311, 1301-1303

[31] Warneken F, Chen F, Tomasello M (2006) Cooperative Activities in Young Children and Chimpanzees, *Child Development*, 77(3) 640-663.

[32] Zöllner R., Asfour T., Dillman R.: Programming by Demonstration: Dual-Arm Manipulation Tasks for Humanoid Robots. *Proc IEEE/RSJ Intern. Conf on Intelligent Robots and systems (IROS 2004).*

Multiagent Collaborative Task Learning through Imitation

Sonia Chernova and **Manuela Veloso** [1]

Abstract. Learning through imitation is a powerful approach for acquiring new behaviors. Imitation-based methods have been successfully applied to a wide range of single agent problems, consistently demonstrating faster learning rates compared to exploration-based approaches such as reinforcement learning. The potential for rapid behavior acquisition from human demonstration makes imitation a promising approach for learning in multiagent systems. In this work, we present results from our single agent demonstration-based learning algorithm, aimed at reducing demonstration demand of a single agent on the teacher over time. We then demonstrate how this approach can be applied to effectively train a complex multiagent task requiring explicit coordination between agents. We believe that this is the first application of demonstration-based learning to simultaneously training distinct policies to multiple agents. We validate our approach with experiments in two complex simulated domains.

1 Introduction

Programming robots is a challenging problem due to sensor complexity, noise, and the non-deterministic effects of robot actions. To address this challenge, autonomous learning approaches have been developed that allow robots to learn task execution through interaction with the environment [14]. Most of these approaches, however, rely on a long trial-and-error experimental process that is impractical due to time constraints and physical wear on the robot. Learning in systems with multiple robots is further complicated by the complex interactions that can occur between distributed agents, such as communication via message passing, physical interaction and resource contention. To address these problems, natural and intuitive approaches must be developed that allow new skills to be taught to multiple of robots in a timely manner.

Learning from demonstration, a collaborative learning approach based on human-robot interaction, offers an alternative to exploration-based methods. The goal of this approach is to learn to imitate the behavior of a teacher by watching a demonstration of the task. Demonstration-based learning has been successfully applied to a variety of single agent learning problems [5, 8, 18, 28]; its fast learning rate compared to exploration-based learning methods, such as reinforcement learning, makes learning from demonstration a promising approach for multiagent systems.

In this work, we first present results of our single agent demonstration-based learning algorithm, the *confident execution* framework [10]. We then apply this framework to a collaborative multiagent domain, demonstrating its effectiveness in simultaneously training multiple robots to perform a joint task. Our learning framework aims to reduce each agent's demonstration demands on the teacher by allowing the agent to perform its task autonomously when it is confident about its actions, and request expert assistance at times of uncertainty. As a result, each agent operates with gradually increasing autonomy as the task is learned, relieving the teacher from repeated demonstrations of acquired behavior and allowing simultaneous supervision of multiple agents.

In the next section we discuss related work in the areas of demonstration and imitation learning, followed by a complete description of the confident execution learning framework in Section 3. In Section 4 we present experimental results demonstrating our approach in single and multi agent domains.

2 Related Work

Learning from demonstration is an interactive learning method in which the agent aims to imitate the behavior of an expert teacher. Demonstration-based methods have been successfully applied to a wide range of single agent learning problems.

Nicolescu and Mataric [17, 18] present a learning framework based on demonstration, generalization and teacher feedback, in which training is performed by having the robot follow a human and observe its actions. A high-level task representation is then constructed by analyzing the experience with respect to the robot's underlying capabilities. The authors also describe a generalization of the framework that allows the robot to interactively request help from a human in order to resolve problems and unexpected situations. This interaction is implicit as the agent has no direct method of communication; instead, it attempts to convey its intentions by communicating though its actions.

Lockerd and Breazeal [8, 15] demonstrate a robotic system where high-level tasks are taught through social interaction. In this framework, the teacher interacts with the agent through speech and visual inputs, and the learning agent expresses its internal state through emotive cues such as facial and body expressions to help guide the teaching process. The outcome of the learning is a goal-oriented hierarchical task model.

Bentivegna et al. [5, 6, 7] and Saunders et al. [25] present demonstration learning approaches using memory-based techniques. Both groups use the k-nearest neighbor (KNN) [16] algorithm to classify instances based on similarity to training examples, resulting in a policy mapping from sensory observations to actions. Our algorithm takes a similar approach by utilizing Gaussian mixture models for classification, but includes an interactive learning component similar to Nicolescu and Mataric. Inamura et al. [13] present a similar method based on Bayesian Networks [20] limited to a discretely-

[1] Computer Science Department, Carnegie Mellon University, email: soniac@cs.cmu.edu, veloso@cs.cmu.edu

valued feature set.

A handful of studies have also examined imitation in the context of multiagent systems. In the Ask For Help framework [11], reinforcement learning agents request advice from other similar agents in the environment. Help is requested when an agent is confused about what action to take, an event characterized by relatively equal quality estimates for all possible actions in a given state.

A similar approach is presented by Oliveira and Nunes [19], in which agents are able to select, exchange and incorporate advice from other agents, combining it with reinforcement learning to improve learning performance. The authors examine when and how agents should exchange advice, and which of an agent's teammates should be communicated with. Their results show that exchange of information can improve the average performance of learning agents, although it may reduce the exploration of the state space, preventing the optimal policy from being found in some cases.

Alissandrakis et al. [2, 3] present a general framework that enables a robotic agent to imitate another, possibly differently embodied, agent through observation. Using this framework, the authors demonstrate the transmission of skills between individuals in a heterogeneous community of software agents. Their results indicate that transmission of a behavior pattern through a chain of agents can be achieved despite differences in the embodiment of some agents in the chain. Additionally, the authors show that groups of mutually imitating agents are able to converge to a common shared behavior.

Price and Boutilier [21] present a multiagent system in which novice agents learn by passively observing other agents in the environment. Each learning agent is limited to observing the actions of others and no explicit teaching occurs. By observing a mentor, the reinforcement learning agent can extract information about its own capabilities in, and the relative value of, unvisited parts of the state space. However, the task of an observed agent may be so different that the observations provide little useful information for the learner, in which case direct imitation of this expert must be avoided by the algorithm.

The above methods study imitation from the perspective of a community of agents, where a single agent seeks advice from other members of its group. A different approach is taken in the study of coaching [23], where an external coach agent provides advice to a team of agents in order to improve their performance at a task. The coach has an external, often broader, view of the world and is able to provide advice to the agents, but not control them. The agents must decide how to incorporate the coach's advice into their execution. Riley [23] presents an approach for training the coach using imitation based on example executions.

Our approach differs from the presented techniques in that it enables a single human to simultaneously train multiple agents. The agents may be differently embodied, and may learn different policies and perform different tasks. In our proposed system, the human teacher is the only source of advice, providing demonstrations in the form of action commands.

3 The Confident Execution Framework

In this section, we present a summary of our confident execution learning framework which allows a single agent to learn a task policy from demonstration (for a more detailed description, please see [10]). We then describe how this framework can be applied to simultaneously training multiple robots to perform a joint task.

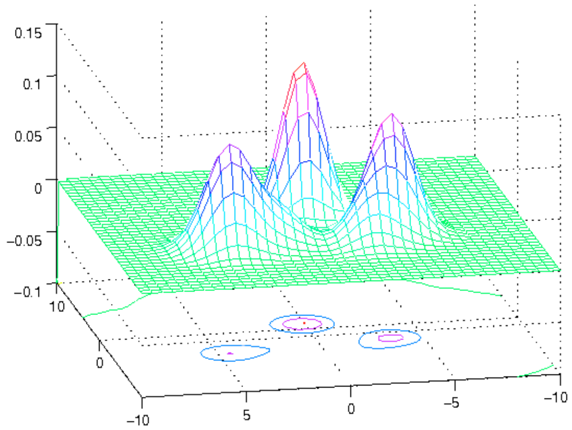


Figure 1. An example of a 2-dimensional Gaussian mixture model with three components. Contour lines below the GMM mark the one- and two-standard deviation ellipses.

3.1 Task Representation

Our approach utilizes the *learning by experienced demonstration* technique [18], in which the agent is fully under the expert's control while continuing to experience the task through its own sensors. During each training timestep, the agent records sensory observations about its environment and executes the action selected by the human expert. We assume the expert attempts to perform the task optimally, without necessarily succeeding.

Observations o are represented using an n-dimensional feature vector that can be composed of continuous or discrete values representing the state of the robot. The agent's actions a are bound to a finite set $\mathcal{A}$ of action primitives [4], which are the basic actions that can be combined together to perform the overall task. The goal of the system is to learn a policy $\pi : o \rightarrow \mathcal{A}$, mapping observations to action primitives. Each labeled training point (o, a) consists of an observation labeled by its corresponding expert-selected action.

During training, the algorithm separates all datapoints into classes based on their action label. A Gaussian mixture model (GMM), Figure 1, is then trained for each action class using the expectation-maximization (EM) algorithm [12]. We selected Gaussian mixture models for our approach due to previous successes of classification methods in demonstration learning [5, 25], and because GMMs provide a built-in measure of classification confidence. Their robustness to noise and ability to generalize and capture correlations between continuous features make GMMs a powerful tool for robotic data analysis.

Since a single action is often associated with a number of distinct states (the action *turn left* may be taken from several different locations), we use a separate Gaussian mixture to represent each action class. Components within the mixture represent the different state regions and the number of components is determined using cross-validation. New datapoints are classified by selecting the Gaussian mixture with the maximum likelihood. The output of the classification is the action represented by the selected GMM. Additionally, the model returns a confidence value representing the certainty of the classification based on the likelihood.

3.2 The Learning Process

Table 1 shows a pseudocode summary of the learning process. Learning begins with a non-interactive demonstration training phase during which each action of the robot is controlled by the expert through teleoperation. The algorithm uses training examples acquired from the demonstrations to generate a task model. Every time $maxNew$ additional training points are acquired, the algorithm updates the GMM based on the new data.

Additionally, the performance of the current learned policy is evaluated by comparing how closely it matches the behavior of the expert. Prior to updating the model with each new training point (o, a), the algorithm classifies observation o using the current model. It then compares the model-selected action to the demonstrated action a. Performing this comparison over a window of consecutive training points results in an estimate of the prediction accuracy of the model that relates how closely the policy matches the behavior of the expert.

The teacher performs non-interactive training until the model prediction accuracy is sufficiently high, as determined by the expert. At this point, learning transitions to the *confident execution* stage, during which the agent selects between autonomously executing its learned policy action and requesting help from the expert based on the classification confidence of the above model. The algorithm adjust the agent's autonomy by comparing the classification confidence to an autonomy threshold. Classification confidences greater than the threshold result in autonomous execution of the model-selected action, while confidences below the threshold cause the agent to pause its execution of the task and signal the teacher that a demonstration is needed.

Since the autonomy threshold value is continuous, our approach allows smooth adjustment of the autonomy level. This type of mechanism is referred to as adjustable, or sliding, autonomy and has been proven effective in a wide range of applications, from personal assistants [26] to space exploration [27]. Our algorithm combines learning with adjustable autonomy, resulting in an interactive teaching method that targets low confidence regions of the state space and reduces dependence on the human expert as the agent gains proficiency at its task. In the presented experiments, the human teacher manually sets the confidence threshold value that determines the level of autonomy. We are currently developing a technique for calculating this value automatically.

As the agent's model improves over time, the agent will encounter fewer observations with low classification confidence, resulting in fewer demonstration requests. Learning terminates when the agent is able to execute the task completely autonomously, or when the expert is satisfied with the performance of the model. The agent then deterministically executes the action selected by the model, regardless of the classification confidence. This mode of operation is typical of traditional learning approaches where the learned policy is always trusted once learning is complete.

3.3 Multiagent Approach

The confident execution learning framework is a promising approach for multiagent learning due to its fast learning rate compared to exploration-based methods such as reinforcement learning [10], and reduced demand on the expert over time. In this work, we examine how it can be directly applied to training multiple agents simultaneously.

In a multiagent setting, the expert's workload and teaching style differ depending on the degree of collaboration required between the agents. Domains with little collaboration allow each agent to operate with little regard for the actions of others, and training can be done independently for each agent. In such cases, it may be possible to introduce new agents at different times, resulting in a mixture of novice and expert agents to avoid overloading the expert at the beginning of the training stage. Domains that require greater collaboration between agents benefit from demonstration-based approaches because exploration over the joint action space of multiple robots is quite costly [9]. In these domains, it is beneficial to demonstrate the task to multiple collaborating agents at the same time.

Algorithm 3.1: THE LEARNING FRAMEWORK()

```
procedure INITIALTRAINING()
  $observation \leftarrow$ GETSENSORDATA()
  $expertAct \leftarrow$ GETEXPERTACTION()
  $(gmmAct, conf) \leftarrow$ CLASSIFY($observation$)
  $predAccuracy \leftarrow$ TRACKPRED($gmmAct, expertAct$)

  if $numNewDatapoints >$ maxNew :
    then UPDATEMODEL($observation, expertAct$)

  EXECUTEACTION($expertAct$)
  return ($predAccuracy$)

procedure CONFIDENTEXECUTION()
  $observation \leftarrow$ GETSENSORDATA()
  $(gmmAct, conf) \leftarrow$ CLASSIFY($observation$)

  if $conf > confThresh$ :
    then { EXECUTEACTION($gmmAct$)
    else { $expertAct \leftarrow$ GETEXPERTACTION()
           UPDATEMODEL($observation, expertAct$)
           EXECUTEACTION($expertAct$)
```

Table 1. Pseudocode overview of the learning framework.

Using our approach described in the previous section, each agent is able to learn its own individual policy regardless of the level of collaboration required. Our algorithm scales to an arbitrary number of robots without any modifications to the learning framework.

4 Experimental Results

We validate our approach using two simulated domains with continuous and multidimensional feature spaces.

4.1 Single Agent Driving Domain

In this section we present results of a fast and dynamic simulated car driving domain (Figure 2). In this domain, the agent takes the shape of a car that must be driven by the expert on a busy road. The speed of the car is fixed at 60 mph while all other cars move in their lanes at predetermined speeds between 20 and 40 mph. The learning agent can not change its speed, and must navigate between other cars to avoid collision. The agent is limited to three actions: remaining in the current lane, and shifting one lane to the left or right of the current position. The road has three normal lanes and a shoulder lane on both sides; the car is allowed to drive on the shoulder but can not go further off the road.

The environment is represented using four features, a distance to the nearest car in each of the three lanes and the current lane of the agent. The agent's lane is represented using a discrete value symbolizing the lane number. The distance features are continuously valued

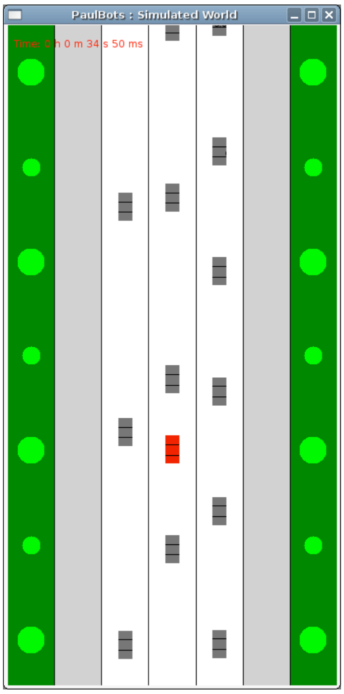


Figure 2. Screenshot of the driving simulator.

in the [-25,25] range; note that the nearest car in a lane can be behind the agent.

Demonstration of the task was performed by a human using a keyboard interface. Figure 3 shows the prediction accuracy of the model during the initial non-interactive training phase. Training was performed until the model reached 80% prediction accuracy over a 150-timestep window, which resulted in a demonstration length of 500 timesteps, or approximately 2.1 minutes. After transitioning to the confident execution phase, the expert completed the training after 150 demonstration timesteps when the model exhibited good performance. During the confident execution phase all demonstrations were done as sequences of ten consecutive moves to simplify the task of the expert due to the fast-paced nature of this domain.

The feature space of this domain is complex as the different action classes frequently overlap. Figure 4 shows a small sample of the data representing how the agent should drive in the middle lane. The data is split into two regions based on the relative position (in front or behind) of the nearest car in the agent's current lane. No samples appear in the 10 to -10 distance range along the Lane2 axis as the expert avoids collisions that would occur from having another car in such close proximity.

The final model consisted of 34 Gaussian components across three GMMs (one for each action class). The final policy was able to imitate the expert's driving style and navigate well in the complex driving domain. Since the algorithm aims to imitate the behavior of the expert, no 'true' reward function exists to evaluate the performance of a given policy. However, we present two domain-specific evaluation metrics that capture the key characteristics of the driving task.

Since the demonstrated behavior attempts to navigate the domain without collisions, our first evaluation metric is the number of collisions experienced under each policy. Collisions are measured as the percentage of the total timesteps that the agent spends in contact with another car. Always driving straight and colliding with every car in the middle lane results in a 30% collision rate.

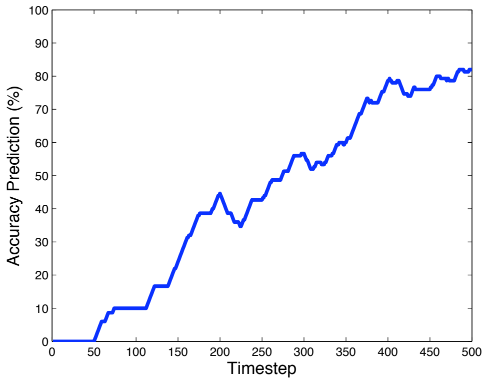


Figure 3. Prediction accuracy of the learned model over the non-interactive training phase using a window of 150 timesteps.

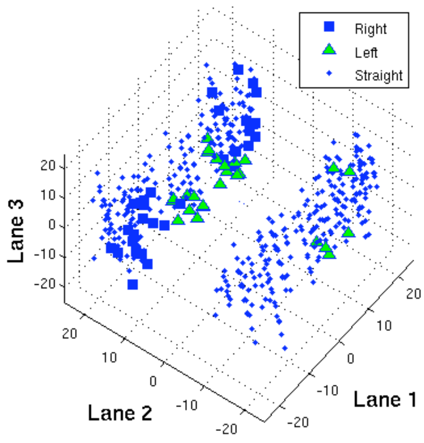


Figure 4. Driving training data representing the driving strategy used when the agent is in the middle lane. Graph axes represent distance to the nearest car in each of the three driving lanes.

Our second evaluation metric is the proportion of the time the agent spends in each lane over the course of a trial. This metric captures the driving preferences of the expert and provides an estimate of the similarity in driving styles. Each evaluation trial was performed for 1000 timesteps over an identical road segment.

Figure 5 compares the performance of the algorithm at different stages in the learning process using these two metrics. Each line in the figure represents a composite bar graph showing the percentage of total time spent by the agent in each lane. Collision percentages for each policy are reported to the right of the bar graphs. The bottom line in the figure shows the performance of the expert over the evaluation road segment (not used for training). We see that the expert successfully avoids collisions, and prefers to use the left three lanes, only rarely using the right lane and right shoulder.

The top five lines summarize the behavior of the agent during the non-interactive training phase. Training was stopped after every 100 training examples for evaluation. Initially the agent always remains in the center lane, accumulating a 30.4% collision rate in the process. As learning progresses, the agent learns to change lanes effectively, beginning to use all five available lanes after 500 demonstration instances, with a collision rate of only 1.3%. However, the agent's lane preference differs significantly from the expert as the agent spends most of its time driving on the right shoulder.

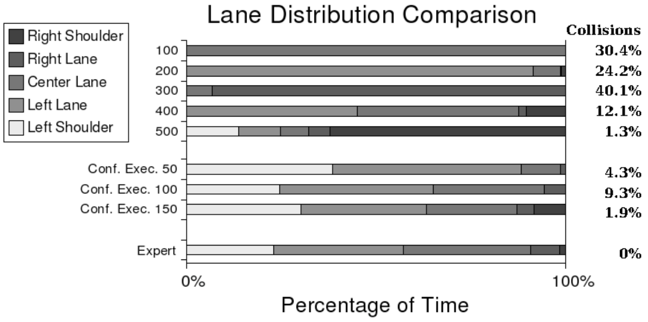


Figure 5. Policy performance comparison using lane distribution and collision evaluation metrics.

The three middle lines display performance during the confident execution phase at 50-timestep intervals. Similarity in lane preference improves over this final training phase, reaching final performance very similar to that of the expert. Additionally, our agent's performance is comparable to that learned using Inverse Reinforcement Learning by Abbeel et al. in [1]. For further evaluation of this domain, including empirical results demonstrating how adapting execution based on confidence focuses training on relevant areas of the domain and a comparison between confident execution and non-interactive demonstration, please see our previous work [10].

4.2 Multiagent Furniture Movers Domain

In this section we present a multiagent collaborative furniture movers domain, Figure 6. In this domain, two agents must move a long, heavy couch from one room to another through a narrow hallway and stairs. We assume that the agents hold opposite ends of the furniture piece throughout this task. Each agent uses six noisy local sensors to determine distances to nearby walls. Additionally, each agent is equipped with a stair sensor that reports a binary value representing the presence or absence of a staircase in the immediate vicinity. The complete feature vector for each agent consists of six continuous distance measurements, and two binary stair features, one for the agent's own location and one for its teammate's. Note that each agent only has a local view of the world, and its teammate's stair information is only updated via a special *communicate* action.

A total of six actions are available to the agents: *forward, back, left, right, communicate*, and *stair*. At each timestep, each agent executes an action based on its own individual policy, and their overall movement is determined by the joint action of both agents. Progress can only be made if the agents select complimentary actions; for example, pulling in opposite directions or attempting to rotate and push at the same time will have no effect on the overall position of the furniture piece. The *communicate* action has no special penalty associated with it, but it does not allow any other action to be activated during that cycle. Since the communicating agent remains stationary for that turn, it prevents any movement regardless of the action taken by the other agent (we assume the couch is too heavy for one agent to move on its own). All movements of the robots are discretized, and the domain can be completed optimally in 39 steps.

The staircase poses a special challenge in this domain, as it requires explicit coordination between the agents. Both agents must select the *stair* action to navigate over the stair segment successfully. However, the corridor is narrow, and the agents are forced to move one after the other instead of side-by-side. As a result, the rear agent is not able to sense when the front agent reaches the staircase. To successfully pass through this region, the front agent must communicate its stair data in order for the rear agent to recognize that the *stair* action is required. Similarly, once the front agent moves past the stairs, the rear agent must communicate its stair information to ensure that the front agent knows to continue executing the *stair* action.

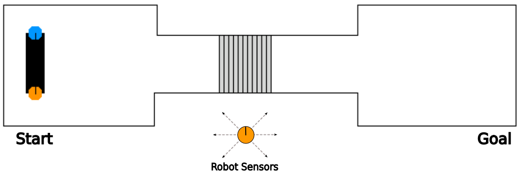


Figure 6. Screenshot of the furniture movers domain. Two agents must collaborate to move a couch from one room to another through a narrow hallway and stairs.

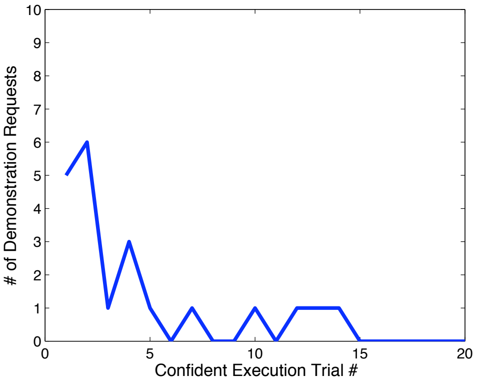


Figure 7. Total number of demonstration requests made by the agents during each trial of the confident execution training phase.

Since a single agent can not perform the task alone, both agents were trained to perform the task at the same time. Demonstrations were performed on an individual basis for each agent. During the confident execution stage, an agent requesting a demonstration waited for the teacher's response, while the other agent was free to continue its execution of the task. Note that in this task, the second agent is not able to make progress on its own due to the constraints of the domain, however, the algorithm places no restrictions upon this agent's actions.

We first evaluate the performance of our learning method using only the non-interactive demonstration technique, in which the agents have no autonomy and the expert performs exhaustive demonstrations of the task. We then present results using the complete confident execution framework. This comparison allows us to evaluate confident execution independently in the context of imitation learning.

Using only the non-interactive demonstration technique, the agents required four demonstrations of the complete domain, or a total of 156 examples per agent, to achieve 100% prediction accuracy and learn the optimal policy. This result confirms that learning from demonstration allows the agents to imitate the behavior of the expert from a small number of examples. Each agent learned its own, unique, policy; the final learned model for each agent consisted of six 8-dimensional Gaussian mixture models.

Confident execution was used to reduce the number of required demonstrations even further by eliminating demonstrations of already acquired behavior. Training was performed using non-interactive demonstration until both models reached 80% prediction accuracy over a window of 15 timesteps, resulting in a total of 65 demonstrations per agent. Under confident execution, the agents continued to perform the task, requesting assistance from the expert at times of uncertainty. Figure 7 shows the total number of demonstration requests made by both agents during each confident execution trial. The number of demonstration requests made decreases with training, until no further requests are made after the 14th learning trial. This resulted in an overall total of 86 demonstrations per agent, approximately half of the number of demonstrations required by the non-interactive method.

Finally, we compare the performance of our algorithm to reinforcement learning. Specifically, Q-learning with a non-deterministic update function was used the learn a policy for each agent. To simplify the task, all action combinations that did not have an effect (such as one agent moving forward, while the other moves back) were not taken into account. This approach was able to learn the optimal policy after 470 iterations, and a total of 58370 exploration steps. Table 2 summarizes the results of all three learning approaches. Note that reinforcement learning performs poorly in this domain because the state of the world is not fully observable as each agent does not know the action taken by its teammate. Partial observability makes this a very challenging problem [22], and a number of special approaches have been developed for dealing with this case [24]. We plan to evaluate and compare these approaches in future work.

Algorithm	**# Steps to Learn**
Non-Interactive Demonstration	156
Confident Execution	86
Reinforcement Learning	58370

Table 2. Comparison of the number of cycles required to learn the optimal policy in the furniture movers domain.

5 Conclusion

In this paper, we proposed imitation as an alternative to exploration-based methods for learning in multiagent systems. We demonstrated the effectiveness of this approach using our demonstration-based learning framework in a complex simulated multiagent domain. Using our technique, we were able to quickly and accurately train the agents to imitate a human demonstration of the task. Additionally, our results showed that the confident execution approach effectively reduces the workload of the expert, allowing training to scale to a greater number of agents.

REFERENCES

[1] Pieter Abbeel and Andrew Y. Ng, 'Apprenticeship learning via inverse reinforcement learning', in *International Conference on Machine learning*, New York, NY, USA, (2004). ACM Press.

[2] Aris Alissandrakis, Chrystopher L. Nehaniv, and Kerstin Dautenhahn, 'Synchrony and perception in robotic imitation across embodiments', in *IEEE International Symposium on Computatonal Intelligence in Robotics and Automation*, Kobe, Japan, (2003).

[3] Aris Alissandrakis, Chrystopher L. Nehaniv, and Kerstin Dautenhahn, 'Towards robot cultures?', *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, **5**(1), 3–44, (2004).

[4] R.C. Arkin, *Behavior-based robotics*, MIT Press, 1998.

[5] D. C. Bentivegna, C. G. Atkeson, and G. Cheng, 'Learning from observation and practice using primitives', *AAAI Fall Symposium Series, 'Symposium on Real-life Reinforcement Learning'*, (2004).

[6] D. C. Bentivegna, G. Cheng, and C. G. Atkeson, 'Learning from observation and from practice using behavioral primitives', *11th International Symposium of Robotics Research*, (2003).

[7] D. C. Bentivegna, A. Ude, C. G. Atkeson, and G. Cheng, 'Learning to act from observation and practice', *International Journal of Humanoid Robotics*, **1**(4), (2004).

[8] C. Breazeal, G. Hoffman, and A. Lockerd, 'Teaching and working with robots as a collaboration', in *AAMAS '04: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 1030–1037, Washington, DC, USA, (2004). IEEE Computer Society.

[9] Georgios Chalkiadakis and Craig Boutilier, 'Coordination in multiagent reinforcement learning: a bayesian approach', in *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pp. 709–716, New York, NY, USA, (2003). ACM Press.

[10] S. Chernova and M. Veloso, 'Confidence-based policy learning from demonstration using gaussian mixture models', in *Joint Conference on Autonomous Agents and Multi-Agent Systems*, (2007).

[11] Jeffery Allen Clouse, *On integrating apprentice learning and reinforcement learning*, Ph.D. dissertation, University of Massachisetts, Department of Computer Science, 1996. Director-Paul E. Utgoff.

[12] A.P. Dempster, N.M.Laird, and D.B. Rubin, 'Maximum likelihood from incomplete data via the em algorithm', *Journal of Royal Statistical Society*, **8**(1), (1977).

[13] T. Inamura, M. Inaba, and H. Inoue, 'Acquisition of probabilistic behavior decision model based on the interactive teaching method', in *Ninth International Conference on Advanced Robotics (ICAR)*, pp. 523–528, (1999).

[14] L.P. Kaelbling, M.L. Littman, and A.W. Moore, 'Reinforcement learning: A survey', *Journal of Artificial Intelligence Research*, **4**, 237–285, (1996).

[15] A. Lockerd and C. Breazeal, 'Tutelage and socially guided robot learning', in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, (2004).

[16] T. Mitchell, *Machine Learning*, McGraw Hill, 1997.

[17] M. N. Nicolescu and M. J. Mataric, 'Learning and interacting in human-robot domains', in *IEEE Transaction on Systems, Man and Cybernetics*, pp. 419–430, (2001).

[18] M. N. Nicolescu and M. J. Mataric, 'Natural methods for robot task learning: instructive demonstrations, generalization and practice', in *Second International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 241–248, New York, NY, USA, (2003). ACM Press.

[19] Eugnio Oliveira and Luis Nunes, *Learning by exchanging Advice*, Springer, 2004.

[20] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann.

[21] Bob Price and Craig Boutilier, 'Accelerating reinforcement learning through implicit imitation.', *J. Artif. Intell. Res. (JAIR)*, **19**, 569–629, (2003).

[22] D. Pynadath and M. Tambe, 'Multiagent teamwork: Analyzing the optimality and complexity of key theories and models', in *1st Conference on Autonomous Agents and Multiagent Systems*, (2002).

[23] Patrick Riley, *Coaching: Learning and Using Environment and Agent Models for Advice*, Ph.D. dissertation, Computer Science Dept., Carnegie Mellon University, 2005. CMU-CS-05-100.

[24] Maayan Roth, Reid Simmons, and Manuela Veloso, 'Reasoning about joint beliefs for execution-time communication decisions', in *The Fourth International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS)*, (2005).

[25] J. Saunders, C. L. Nehaniv, and K. Dautenhahn, 'Teaching robots by moulding behavior and scaffolding the environment', in *HRI '06: Proceeding of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pp. 118–125, New York, NY, USA, (2006). ACM Press.

[26] P. Scerri, D. Pynadath, and M. Tambe. Towards adjustable autonomy for the real world, 2003.

[27] M. Sierhuis, J. Bradshaw, A. Acquisti, R. Hoof, R. Jeffers, and A. Uszok. Human-agent teamwork and adjustable autonomy in practice, 2003.

[28] W. D. Smart and L. P. Kaelbling, 'Effective reinforcement learning for mobile robots', in *IEEE International Conference on Robotics and Automation*, (2002).

Echo State Network Applied to a Robot Docking Task

Xavier Dutoit and **Davy Sannen** and **Marnix Nuttin**[1]

Abstract. Reservoir Computing (RC) is a new technique which allows to use complex recurrent neural networks while keeping the training complexity low. We apply here RC as a high-level controller for a robot which has to perform a docking task. The RC method presents two main advantages. The task can be taught in a black-box approach, using only learning by imitation. The explicit dependence from situations to actions does not need to be coded. And RC requires only training simple readouts which can be guaranteed to find a local minimum.

1 INTRODUCTION

When controlling a robot, one wants the robot to be intelligent and autonomous. This means that the robot has to be able to decide actions by itself in a environment which is constantly changing. Moreover, as the sensors are not perfect, the robot has a noisy or even inconsistent perception of this environment.

In order to solve those problems, a lot of work has been done in the field of robotic control. This work can be divided in different categories (see [13] for a more complete description):

- **Reactive Control**: The robot has no memory but just makes a mapping from situations to action. This is simple to implement and fast to execute, but the number of tasks it can perform is rather limited.
- **Deliberative Control**: Here some more complex processing is involved, and the robot has a memory, so it can associate an action to a given situation with a given past. This allows to deal with more complex task, but requires more hardware and computation time.
- **Hybrid Control**: This approach is a trade-off between the two preceding techniques, and can allow to combine their advantages.
- **Behaviour-based Control**: As the name says, the robot has a set of behaviour. Depending on the situation, it can choose which behaviour to execute. This allows to be more flexible.

We intend to solve here a non-reactive task and will use deliberative control. However, deliberative architectures usually need explicit coding. We will rather use here another approach which would allow to solve the task in a more natural way, without coding the explicit dependency from situation to actions. Instead, in our approach, it is possible to train the robot by imitation.

To do so, we use Artificial Neural Networks (ANNs). They allow to process inputs in a nonlinear and adaptive way. Unlike classical approaches, there is no need to know in advance how to solve the task: neural networks can show an ability to learn by themselves, when provided a good set of examples. They can then generalize from this set of examples. Moreover, they are typically able to deal with noisy or inconsistent data (see for instance [25]). This altogether makes them very interesting for robotic applications. More precisely, we will consider ANNs in the framework of Reservoir Computing (RC).

2 RESERVOIR COMPUTING

The basic idea of RC is to input the data into a big recurrent network and then to train some simple readouts to extract useful information, while the network itself remains unchanged. RC has been introduced by [17] with the *Liquid State Machines* (LSMs), where the network consist of spiking neurons, and [10] with the *Echo State Networks* (ESNs), where the network consists of sigmoidal neurons. It can also be compared to the results from [24] when they studied the weight dynamics of a recurrent neural network and to the *Backpropagation-Decorrelation* algorithm [26].

The part of the network which is not trained can be seen as a reservoir of functions, and the output neurons as readouts that can extract the main features from this reservoir.

When the input is presented to the reservoir, it is in fact projected into a high dimensional and highly dynamic space. This is similar to a kernel method (see e.g. [6] for a review), and has the advantage over classical kernel methods that it can include time.

A great advantage of RC is then that we can apply simple readout functions to the reservoir, like linear discriminants, which are simple to train and can be guaranteed to find a global optimum in the offline case.

The power of reservoir computing has potentially no limit: any task can be solved as long as the desired features are present in the reservoir. On the other hand, a drawback is that the features have to be presents in the reservoir, which is not always the case, and it is typically hard to know in advance how to design a reservoir in order to make it capture those features. But if it manages to have those features, it requires absolutely no prior knowledge about the task to be solved, whereas with other approaches, some hard-coding of time-dependent actions has to be made.

We will here focus more particularly on ESNs. They are simpler to implement and simulate than LSMs, as they use classical (sigmoidal) neurons whereas LSMs use spiking neurons interconnected by synapses with a weight and a delay.

2.1 Applications of Reservoir Computing

RC has been applied with promising results in several domains, like:

Speech recognition In [28], an LSM has been trained to recognize spoken digits. The LSM has shown a good robustness against noise. It is interesting to see that, amongst 3 different pre-processing techniques of the sound, the most biological model, the Lyon Passive Ear[16], has lead to the best results.

[1] Katholieke Universiteit Leuven, Belgium, email:{xavier.dutoit,davy.sannen,marnix.nuttin}@mech.kuleuven.be

Movement prediction In [5], an LSM has been trained to predict the movement of a ball with real-world images. It was able to predict the movement reliably up to 200 ms ahead. However, the results depend on a good choice of the parameters of the liquid.

The XOR problem and real liquid [9] used a real liquid excited by electric motors and whose image was recorded by a web-cam. They trained it to simulate a XOR logic gate and to distinguish between the spoken digits 'one' and 'zero' and showed good performances and good robustness against noise.

Real-time obstacle avoidance [4] used a LSM implemented in real-time to control a small robot and make it avoid obstacles. The learning was done by demonstration.

Arm control [14] used a LSM to control a robot arm in a biologically inspired way. The arm was trained to reach different target points. It was a first implementation of a closed-loop system controlled by neural microcircuits.

2.2 Learning by imitation

If we control a robot with reservoir computing, as nothing is programmed beforehand, it has to learn the task. A very appealing way is to make it learn by imitation. It consists of showing the robot a desired behaviour in order to make it learn to exhibit the same behaviour afterward, when the same situation is presented (see for instance [2] for a review).

Learning by imitation is very appealing because of its conceptual simplicity when compared to other methods. Typically, it is often much simpler to show a robot what to do by doing it ourselves than to program it. It has also the advantage that it does not necessarily require concrete knowledge about the robotic domain: a person who does not know how a robot has to be programmed can still show some tasks to the robot. This advantage is interesting, especially if we consider the application of domestic robots, where anybody could teach a robot a given task in this way. It is very appealing for cooperation between human and robot and for real-world learning applications [1], [3], [8], [18], [19], [23].

Moreover, learning by imitation is a natural way to teach and learn for human beings and animals. It is very commonly observed amongst monkeys, for instance, and in fact it is the reason of the name *aping*.

3 THE EXPERIMENT

3.1 Related work

We try here to apply the technique of RC to learn a docking task. ESNs have already been applied to control task (see for instance [21, 22, 20]). However, the previous applications generally use the ESN as a low-level controller of which the goal is to output the motor command based on a desired trajectory. In our approach, we first present a set of trajectories to learn, but then, during the testing phase, the ESN has to decide the trajectory based on the sonar input only.

The docking task has also already been solved with a behaviour-based approach [15]. However, when using a behaviour-based approach, the task needs to somehow be segmented in the different manoeuvres the robot will have to make. On the other hand, with RC, the raw data is fed to the reservoir, without any preprocessing or prior knowledge involved.

This task can also be solved using planners [27],[7]. However, we are interested here in a more adaptive and flexible approach, as we think it might exhibit some interesting features in the long run.

3.2 Goal

The robot must perform a docking task, i.e. it must first go backward and then go forward and turn left (cf. Fig. 1). It starts around the point indicated by 'Start', oriented towards the positive direction of the x axis. In one zone, the shaded area, there are some points where the robot will go twice, and thus be twice in the same situation, but with different desired outputs. So the task is not a purely reactive task, it features time dependency.

A run is considered as successful if the robot first goes back enough to enter the shaded zone (cf. Fig. 1) and then reaches the goal area.

3.3 Setup

We first simulated a robot based on the PIONEER-1 robots. It has 7 sonars placed symmetrically at its front (see Fig. 1), at the angles of ±90,±30,±15 and 0 degrees (0 being the forward direction of the robot).

The robot can give discrete commands out of two independent sets:

- Linear velocity: go forward or go backward
- Rotational velocity: turn-left, do-not-turn or turn-right

A step forward or backwards corresponds to a distance of 100 mm, a turn left or right to an angle of $\frac{\pi}{8}$.

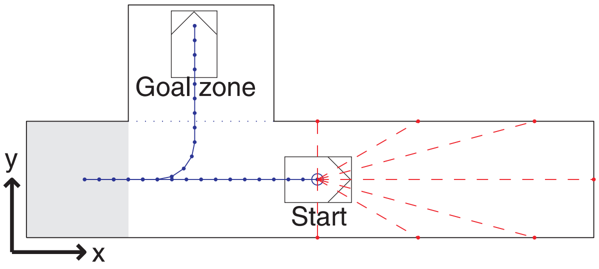


Figure 1. The Docking Task. The robot starts at the point indicated by *Start*; it moves along the solid line, each dot representing a step; the shaded area represent the contradictory zone; the dotted lines indicate the sonars for the robot at the starting position.

3.4 Training data

The network is trained by demonstration. 60 different trajectories were created by a human supervisor (20 from the original starting point, and 40 from other random points in the environment), and they were then randomly shifted to create new trajectories. In total, the training set consists of 494 trajectories.

For each trajectory, we record the 7 sonar data. In the simulation environment, those data do not get any noise. Each sonar returns the distance (in [mm]) to the nearest wall.

For each trajectory, the sonar data were fed to the network and the 3 readouts were trained to map the states of the network onto the desired motor commands (see section 4.5 for the detailed training procedure).

4 THE ECHO STATE NETWORK

The ESN considered here consists of one input layer, one reservoir, and one output layer.

There are n_i neurons in the input layer, n_r neurons in the reservoir and n_o neurons in the output layer.

In the present task, $n_i = 7$ (the 7 sonars inputs), and $n_o = 3$ (we use 3 outputs to command the two velocities, the mapping from outputs to commands is described below, section 4.3)

4.1 Input

At each time step t, the input vector $\mathbf{i}(t)$ is multiplied by a input weight matrix $\mathbf{W}_I$, of size $n_r \times n_i$, and fed to the reservoir.

4.2 Reservoir

The reservoir, consisting of n_r neurons, is described by a connection matrix $\mathbf{W}$, of size $n_r \times n_r$, and at each time step by a state vector $\mathbf{s}(t)$. This state vector is all zero at the beginning and is updated according to the following equation:

$$\mathbf{s}(t{+}1) = f\Big(m{\cdot}\big(\mathbf{W}_I{\cdot}\mathbf{i}(t){+}\mathbf{W}{\cdot}\mathbf{s}(t)\big){+}(1{-}m){\cdot}\mathbf{s}(t)\Big) \quad \forall\, t > 0 \quad (1)$$

$$\mathbf{s}(0) = 0$$

where f can be any linear or non-linear function (here we use a sigmoidal function, the hyperbolic tangent), and m ($0 \leq m \leq 1$) is a parameter tuning the dynamic of the reservoir.

4.3 Output

Each readout r is a linear discriminant, described by a weight vector $\mathbf{W}_r$. The output of the network O_r at time t is given by:

$$O_r(t) = \mathbf{W}_r \cdot \bar{\mathbf{s}}(t) \qquad (2)$$

where $\bar{\mathbf{s}}(t)$ is the state vector augmented with a bias term:

$$\bar{\mathbf{s}}(t) = \begin{bmatrix} \mathbf{s}(t) \\ 1 \end{bmatrix}$$

In our case, there are 3 readouts, one for the linear velocity V, two for the rotational velocity R. The actual commands are:

$$V(t) = \begin{cases} +1 & \text{if } O_1(t) > 0 \\ -1 & \text{otherwise} \end{cases} \qquad (3)$$

$$R(t) = \begin{cases} +1 & \text{if } O_2(t) - O_3(t) > \Theta \\ -1 & \text{if } O_3(t) - O_2(t) > \Theta \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

Θ being a threshold factor, determined experimentally.

4.4 Network creation

The reservoir is created at random, according to the following parameters:

n_i	the number of inputs
n_r	the size of the reservoir
c_i	the input connection fraction
c_r	the reservoir connection fraction
i_w	the weights distribution of the inputs
m	the memory parameter

The input connection matrix $\mathbf{I}$ is a $n_r \times n_i$ matrix with a proportion c_i of non-zero weights. Those non-zero weights take on their values uniformly in i_w (in our case, $i_w = \{-0.1; +0.1\}$).

The reservoir connection matrix $\mathbf{C}$ is a $n_r \times n_r$ matrix with a proportion c_r of non-zero weights. Those non-zero weights take on their values out of a 0-mean gaussian distribution with variance 1.

Once generated, this connection matrix is rescaled: it is divided by its spectral radius (so that its spectral radius is 1 after rescaling). This rescaling allows to stand at the limit of the *echo state property* [10].

4.4.1 Effect of the memory parameter m

The parameter m allows to have leaky neurons, i.e. neurons which have a certain memory. Indeed, if $m < 1$, at each time step, a neuron will have as net input (i.e. before applying the non-linear function) the net input from other neurons multiplied by m and $(1 - m)$ times its own delayed input. In the absence of external input, the activity level of a given neuron exponentially decays with a time constant of $\frac{1}{m}$[time steps].

Now concerning the echo state property, one has to note that we can rewrite equation 1 as:

$$\mathbf{s}(t+1) = f\left(\widetilde{\mathbf{W}}_I \cdot \mathbf{i}(t) + \widetilde{\mathbf{W}} \cdot \mathbf{s}(t)\right) \quad \forall\, t > 0$$

where $\widetilde{\mathbf{W}}_I = m \cdot \mathbf{W}_I$ and $\widetilde{\mathbf{W}} = m \cdot \mathbf{W} + (1 - m) \cdot \mathbf{I}$ ($\mathbf{I}$ being the identity matrix).

As $\mathbf{W}$ has a spectral radius equal to one, i.e. all its eigenvalues are smaller or equal to one, $\widetilde{\mathbf{W}}$ has all its eigenvalues smaller or equal to m, and its spectral radius equal to m. So the echo state property is guaranteed for $m < 1$, and we stand at the limit of this property when $m = 1$.

4.5 Training

The training set consist of a set of n_t vector of size n_i, and of a set of n_t associated desired output pairs $(\hat{\mathcal{V}}(t), \hat{\mathcal{R}}(t))$. For each sample t, we define the 3 desired output $(\hat{O}_1(t), \hat{O}_2(t), \hat{O}_3(t))$ as follows:

$$\hat{O}_1(t) = \hat{\mathcal{V}}(t)$$

$$\hat{O}_2(t) = \begin{cases} +1 & \text{if } \hat{R}(t) = +1 \\ -1 & \text{otherwise} \end{cases}$$

$$\hat{O}_3(t) = \begin{cases} +1 & \text{if } \hat{\mathcal{R}}(t) = -1 \\ -1 & \text{otherwise} \end{cases}$$

Now to do the actual training, the network is fed with the n_t input samples, and we collect the augmented states in a matrix $\mathbf{S}$:

$$\mathbf{S} = [\bar{\mathbf{s}}(1)\, \bar{\mathbf{s}}(2)\, \ldots\, \bar{\mathbf{s}}(n_t)]$$

The readouts are computed by solving the following equation in the least square sense:

$$\mathbf{W}_r \cdot \mathbf{S} = \hat{\mathbf{O}}_r \qquad r = 1, 2, 3$$

where $\hat{\mathbf{O}}_r$ is the vector containing the desired output for all the n_t samples:

$$\hat{\mathbf{O}}_r = [\hat{O}_r(1)\,\hat{O}_r(2)\,\ldots\,\hat{O}_r(n_t)]$$

The actual commands are then computed according to (2), (3) and (4).

4.6 Training error

The training error is the proportion of wrong commands over the training set, defined as:

$$E_T = \sum_t e(t)$$

where:

$$e(t) = \begin{cases} 0 & \text{if } V(t) = \hat{V}(t) \text{ and } R(t) = \hat{R}(t) \\ 1 & \text{if } V(t) \neq \hat{V}(t) \text{ and } R(t) \neq \hat{R}(t) \\ 0.5 & \text{otherwise} \end{cases}$$

4.7 Testing error

Once the network was trained, it was tested starting from 10 different point chosen randomly around the original starting point according to a normal distribution of mean 0 and of variance 200 mm on the x axis and 100 mm on the y axis. To avoid the robot starting too close to a wall or outside the world, the starting point was limited to be no more than 500 mm and 300 mm away from the starting point, on the x and y axis resp.

The testing error is the proportion of trajectories which did not fulfill the success criterion (see above, section 3.2).

5 RESULTS

We applied here an ESN approach to teach a robot to perform a docking task. Several reservoirs were created randomly, without any programming of the task beforehand, and were trained by demonstration to reproduce the training runs. By training only 3 linear discriminants, it is possible to achieve an average success rate of 76 % on testing (see Fig. 2 for examples of successful trajectories). Some of the networks managed to perform the task successfully in all the cases tested.

In the present experiment, as the task is time-dependent, an important point is the memory of the reservoir. So far, there exist little methodology or measure of the memory of a given reservoir[11, 12]. However, we can say that the memory roughly depends on two parameters: the reservoir size n_r and the memory scale m.

The n_r controls the memory on a global scale. When all parameters stay constant, a bigger reservoir will mean that there exist potentially longer loops inside the reservoir, and the input will thus have longer echoes.

On the other hand, m controls the memory on a local scale: the smaller m is, the longer is the memory of a given neuron, as a past input will have an exponentially decreasing effect for a longer time. However, m also scales down the global spectral radius, thus changing the memory on a global scale as well.

The general results for those two parameters are shown in Fig. 3.

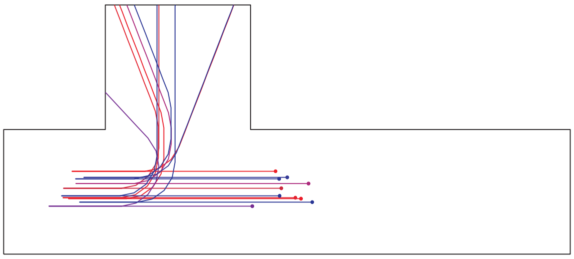


Figure 2. Example of trajectories with different starting points around the original starting point

5.1 Reservoir size

If we take a closer look at the effect on n_r (see Fig. 4), we see that in the present experiment a bigger reservoir leads to better results. This is because it allows to have more memory, but also because a bigger reservoir is likely to capture more features from the input, and thus is more likely to capture the relevant features for the task.

5.2 Memory scale

If we now look at the effect of m, the memory scale, we see that a smaller m, corresponding to a longer memory, produces on average better results. However, for the testing error, there is a lower limit under which the test error starts to increase again.

One can notice than even with a badly scaled memory, we can still have around 40 % of the networks which succeed to perform the task. This shows that it is not required to know in advance what are the memory requirements of the task in order to be able to perform it successfully.

There is also a second noticeable point: in the training samples, the time spent in the contradictory input zone (shaded area in Fig. 1) was around 7 steps. So the robot saw twice the roughly same input, first at a given time and then 7 time steps later. So we can roughly say that the memory requirements for this task is about 7 time steps, i.e. a robot must have a memory spanning at least 7 time steps in order to perform the task. But we can see that with $m = 0.25$, i.e. when each neuron has a time scale of 4 steps, there were still around 35 % of the networks which performed the task successfully. So even when the memory is badly scaled, it is possible to sometimes succeed in performing the task. This shows that a reservoir can exhibit on a global scale a behaviour on a time scale larger than the local time scale of any of its element.

5.3 Extension to more realistic environment

We then applied the task in the environment of the Saphira robot simulator. This means that there was some noise in the input and in the output, i.e. the sonars data were noisy and the commands were not perfectly executed. Moreover, in this environment, the robot has now a radius of about 200 mm and so it has less margin to manoeuver. So the sonars data were all getting subtracted 200 before being fed to the robot.

It succeeded both in the simulation environment and with a real robot (Fig. 7), and an example of successful trajectories is shown in Fig. 6.

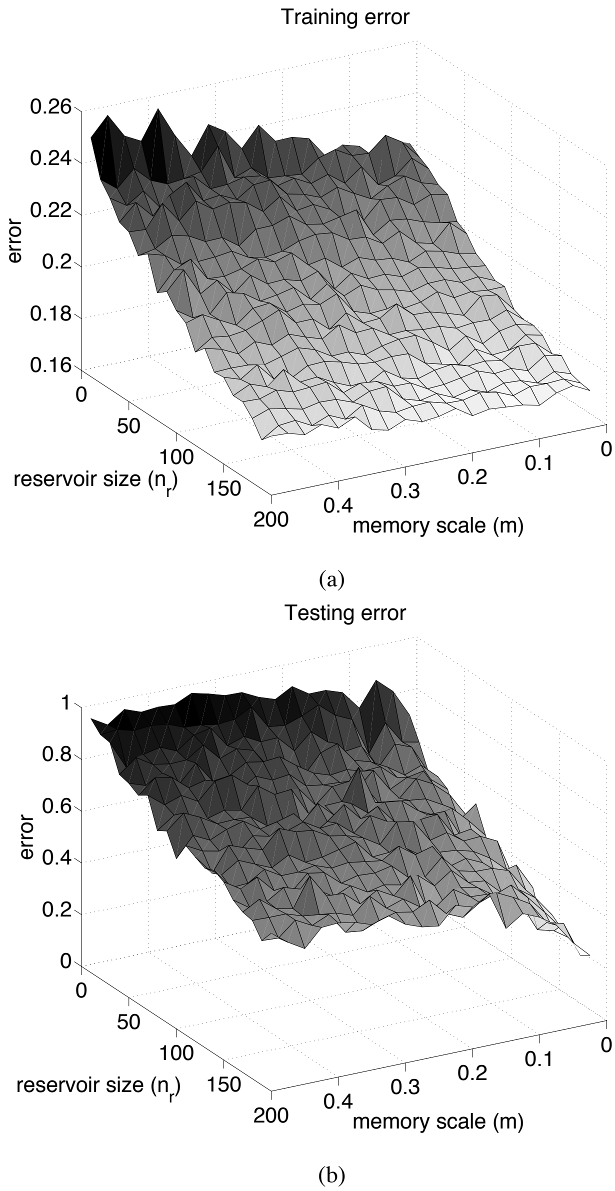


Figure 3. General view of the effect of n_r and m: (a) training error, (b) testing error

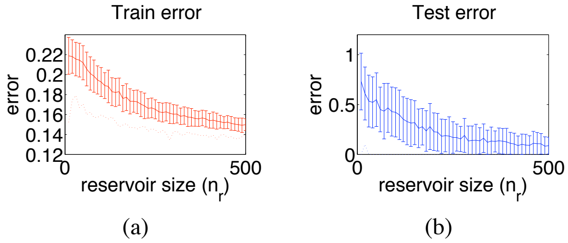


Figure 4. Effect of the reservoir size, n_r: (a) training error, (b) testing error. The solid line represents the average error with the standard deviation, the dotted line represents the minimum error.(out of 100 simulations, $m = 0.02$)

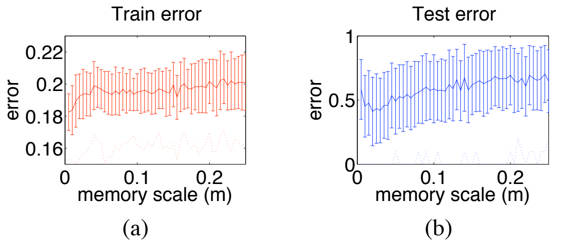


Figure 5. Effect of the memory scale, m: (a) training error, (b) testing error. The solid line represents the average error with the standard deviation, the dotted line represents the minimum error.(out of 100 simulations, $n_r = 100$)

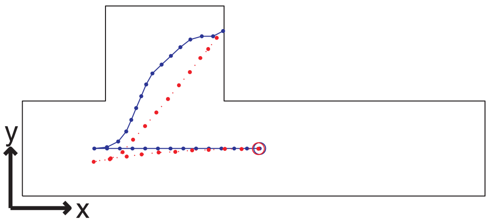


Figure 6. Example of the application in the Saphira environment: solid line: simulation environment, dotted line: real world.

6 CONCLUSION

We considered here an application of an emerging technique: reservoir computing. To the best of our knowledge, it is the first time RC is applied on a high-level to a navigation problem such as this docking task. With RC, it has been possible to successfully perform a control task in the simulated world, as well as in the real world, where the sensor readings were noisy, and the commands were not perfectly executed.

This RC technique allows a simple training. Indeed, it is sufficient to generate some training examples and to show them to the network. Thus we do not need to know the explicit correspondence from input to output, which is typically hard to know. Moreover, we can also use a very simple training algorithm, which is guaranteed to find an optimal solution (in the least-square sense): the training consists of a matrix inversion, which can be computationally expensive, but is straightforward and guarantees to find the global optimum. A drawback is that this method can only be implemented offline. The RC approach also allows to be flexible, and even though some parameters have to be tuned and tested, it is possible to perform the desired task even with badly scaled parameters. Thus we think that this technique is promising for robot task control.

ACKNOWLEDGEMENTS

This work was partially funded by FWO Flanders project G.0317.05, and by the Belgian government under the Inter-University Attraction Poles, Office of the Prime Minister, IAP-AMS.

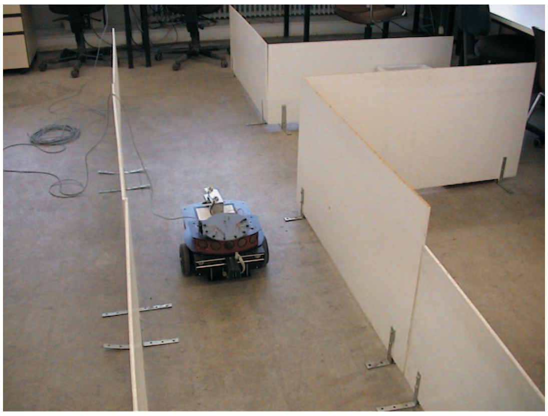


Figure 7. The real world setup

REFERENCES

[1] P. Baker and Y. Kuniyoshi, 'Robot see, robot do: an overview of robot imitation', *AISB96 Workshop on Learning in Robots and Animals*, 3–11, (1996).

[2] A. Billard and R. Siegwart, 'Robot learning by demonstration', *Robotics and Autonomous Systems*, **47**(2-3), 65–67, (2004).

[3] E. Burdet and M. Nuttin, 'Learning complex tasks using a stepwise approach', *Journal of Intelligent and Robotic Systems*, **24**, 43–69, (1999).

[4] H. Burgsteiner, 'Training networks of biological realistic spiking neurons for real-time robot control', *Proceedings of the EANN 2005*.

[5] H. Burgsteiner, M. Kröll, A. Leopold, and G. Steinbauer, 'Movement prediction from real-world images using a liquid state machine', in *Proceedings of the 18th Internal Conference IEA/AIE*. Springer-Verlag, Berlin, Germany, (2005).

[6] C. Campbell, 'Kernel methods: a survey of current techniques', *Neurocomputing*, **48**, 63–84, (2002).

[7] E. Demeester, M. Nuttin, D. Vanhooydonck, and H. Van Brussel, 'Fine motion planning for shared wheelchair control: Requirements and preliminary experiments', *Int. Conf. on Adv. Robotics*, 1278–1283, (2003).

[8] R. Dillmann, O. Rogalla, M. Ehrenmann, R. Zollner, and M. Bordegoni, 'Learning robot behaviour and skills based on human demonstration and advice: the machine learning paradigm', *9th International Symposium of Robotics Research*, (1999).

[9] C. Fernando and S. Sojakka, 'Pattern recognition in a bucket', in *Proc. of the ECAL 2003*, (1998).

[10] H. Jaeger, 'The "echo state" approach to analysing and training reccurent neural networks', Technical Report GMD 148, German National Research Center for Information Technology, (2001).

[11] H. Jaeger, 'Short-term memory in echo state networks', Technical report, (2002).

[12] H. Jaeger, 'A tutorial on training recurrent neural networks, covering bppt, rtrl, ekf and the "echo state network" approach', Technical report, (2002).

[13] C. Jones and M. Matarić, 'Behavior-based coordination in multi-robot systems', *Autonomous Mobile Robots: Sensing, Control, Decision-Making, and Applications*, (2005).

[14] P. Joshi and W. Maass, 'Movement generation and control with generic neural microcircuits', *Proceedings of BIO-ADIT*, (2004).

[15] M. Kasper, G. Fricke, K. Steurenagel, and E. von Puttkamer, 'A behavior-based mobile robot architecture for learning from demonstration', in *Robotics and Autonomous Systems*, volume 34, pp. 153–164, (2001).

[16] R.F. Lyon, 'A computational model of filtering, detection and compression in the cochlea', in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, (1982).

[17] W. Maass, T. Natschläger, and H. Markram, 'Real-time computing without stable states: A new framework for neural computation based on perturbations', *Neural Computation*, **14**(11), 2531–2560, (2002).

[18] M. Nuttin, *Learning Approaches to Robotics Manufacturing: Contributions and Experiments*, Ph.D. dissertation, K.U.Leuven, 1998.

[19] M. Nuttin and H. Van Brussel, 'Learning assembly operations: A case study with real-world objects', *Studies in Informatics and Control*, **3**, 205–221, (1996).

[20] M. Oubbatti, *Neural Dynamics for Mobile Robot Adaptive Control*, Ph.D. dissertation, Universität Stuttgart, 2006.

[21] P.G. Ploeger, A. Arghir, T. Günther, and R. Hosseiny, 'Echo state networks for mobile robot modeling and control', *Proceedings of the ROBOCUP 2003*.

[22] M. Salmen and P.G. Ploeger, 'Echo state networks used for motor control', *Proceedings of the ICRA 2005*.

[23] S. Schaal, A. Ijspeert, and A. Billard, 'Computational approaches to motor learning by imitation', *Philosophical Transactions of the Royal Society B: Biological Sciences*, **358**(1431), 537–547, (2003).

[24] U.D. Schiller and J.J. Steil, 'On the weights dynamics of reccurent learning', in *ESANN'2003 Proceedings*, pp. 73–78, (2003).

[25] J.W. Shavlik, R.J. Mooney, and G.G. Towel, 'Symbolic and neural learning algorithms: An experimental comparison', *Machine Learning*, **6**(2), 111–144, (1991).

[26] J.J. Steil, 'Backpropagation-decorrelation: online reccurent learning with o(n) complexity', in *Proc. IJCNN*, volume 1, pp. 843–848, (2004).

[27] J. Vandorpe, *Navigation techniques for the mobile robot LiAS*, Ph.D. dissertation, K.U.Leuven, 1997.

[28] D. Verstraeten, B. Schrauwen, D. Stroobandt, and J. Van Campenhout, 'Isolated word recognition with the liquid state machine: a case study', *preprint submitted to Elsevier Science*, (2005).

Can Motionese Tell Infants and Robots “What to Imitate”?

Yukie Nagai [1] and **Katharina J. Rohlfing** [2]

Abstract. An open question in imitating actions by infants and robots is how they know “what to imitate.” We suggest that parental modifications in their actions, called *motionese*, can help infants and robots to detect the meaningful structure of the actions. Parents tend to modify their infant-directed actions, e.g., put longer pauses between actions and exaggerate actions, which are assumed to help infants to understand the meaning and the structure of the actions. To investigate how such modifications contribute to the infants’ understanding of the actions, we analyzed parental actions from an infant-like viewpoint by applying a model of saliency-based visual attention. Our model of an infant-like viewpoint does not suppose any a priori knowledge about actions or objects used in the actions, or any specific capability to detect a parent’s face or his/her hands. Instead, it is able to detect and gaze at salient locations, which are standing out from the surroundings because of the primitive visual features, in a scene. The model thus demonstrates what low-level aspects of parental actions are highlighted in their action sequences and could attract the attention of young infants and robots. Our quantitative analysis revealed that motionese can help them (1) to receive immediate social feedback on the actions, (2) to detect the initial and goal states of the actions, and (3) to look at the static features of the objects used in the actions. We discuss these results addressing the issue of “what to imitate.”

1 INTRODUCTION

Imitation learning is a promising approach for robotics researchers to enable their robots to autonomously acquire new skills from humans [21, 31]. It allows robots to learn new behaviors by first observing human movements and then reproducing them by mapping into their motor commands. It consequently reduces the efforts of designers in developing robots’ behaviors. In addition to these engineering benefits, the research on imitation learning leads us to the deeper understanding of human intelligence [2]. Human infants, even neonate [25, 26], are able to imitate actions. In the course of their development, infants can reproduce actions and the goal of actions shown by another person. The ability to imitate is moreover discussed as a route to their further cognitive development, e.g., the differentiation of the self and other, the understanding of other’s intention, and the use of language [9]. Thus, to investigate the mechanism for imitation learning from a constructivist viewpoint allows us to uncover human intelligence [2].

There are some advantages in robot imitation, however, we still have an open question of how robots know “what to imitate” and “how to imitate.” Nehaniv and Dautenhahn [28, 29] discussed these two fundamental issues in robot imitation. Breazeal and Scassellati [7, 8] also pointed out the issues and reported the current techniques used in robot systems. When a robot attempts to imitate a human action or a sequence of his/her actions to achieve a goal-oriented task, it has to first detect the movements of the person and then determine which movements are relevant to the task. A robot without any a priori knowledge about the task does not know which actions of the person are important and necessary for the task, while he/she sometimes produces not only actions directly related to the task but also unrelated ones. It is also required to detect the initial and goal states of the actions and the objects involved in the actions so that a robot can imitate the sequence of the actions not only at a trajectory level but also at a goal level. These problems are stated as the issue of “what to imitate,” and several approaches have been proposed from different perspectives (e.g., [4, 6, 10, 11, 34]).

Another issue to be solved in robot imitation is how a robot knows “how to imitate.” A robot that tries to imitate human actions has to be able to transform the observed actions of a person into its motor commands so as to reproduce the same actions or to achieve the same goal of the actions. A difficulty in transforming the actions is that a robot cannot access to the somatosensory information of the person and is thereby unable to directly map the actions into the motor commands. Moreover, the body structure of a robot is usually different from the person’s, which makes the problem more difficult. These issues are called “how to imitate” and have been investigated from various approaches (e.g., [1, 3, 4, 10]).

In addressing these issues from a standpoint of cognitive developmental robotics [2], we suggest that parental modifications in their infant-directed actions can help robots as well as infants to imitate the actions [12, 30]. When infants attempt to imitate actions presented by their parents, they also face the same problems: “what to imitate” and “how to imitate.” Although infants are supposed to have little semantic knowledge about actions as robots do, they are surprisingly able to imitate the actions. They are skillful in processing a stream of ongoing activity into meaningful actions and organizing the individual actions around ultimate goals [33]. We thus consider that parental actions aid infants solving “what to imitate” and “how to imitate.” It is known that parents tend to modify their actions when interacting with their infants (e.g., [5, 30]). They, for example, put longer and more pauses between their movements, repeat the same movements, and exaggerate their movements when interacting with infants compared to when interacting with adults. Such modifications, called *motionese*, are suggested to aid infants structuring the actions and understanding the meaning of the actions. However, we do not know yet how it actually affects and contributes to the infants’ understanding of the actions. Because the current researches have analyzed motionese only from an adult’s viewpoint, i.e., they focused

[1] Bielefeld University, Germany, email: yukie@techfak.uni-bielefeld.de

[2] rohlfing@techfak.uni-bielefeld.de

only on the actions relevant to a task, it is still unclear what aspects of parental actions would be attended to by infants and how they help infants to understand and imitate the actions.

We analyze motionese from an infant-like viewpoint and discuss how it can help infants and robots to detect "what to imitate." Our model of an infant-like viewpoint does not suppose any a priori knowledge about actions or objects used in the actions. It does not know which parental actions are relevant to a task, what the goal of the task is, or what objects are involved in the task. Furthermore, it is not equipped with any specific ability to detect a parent's face or his/her hands. Instead, it is able to detect and gaze at outstanding locations in a scene. To simulate such a capability of visual attention, we adopt a model of saliency-based visual attention [16, 17] inspired by the behaviors and the neural mechanism of primates. A salient location in this model is defined as a location which locally stands out from the surroundings because of its color, intensity, orientation, flicker, and motion [16]. It thus can demonstrate what low-level aspects of parental actions are highlighted in their action sequences and could attract the attention of young infants and robots. We analyze motionese with the model and discuss the results toward solving the issue of "what to imitate."

The rest of this paper is organized as follows. In Section 2, we summarize the current evidences of motionese from psychological and computational studies. In Section 3, we introduce the model of saliency-based visual attention and describe the benefits of using it for the analysis of motionese. Next, we show analytical experiments of motionese in Section 4, and discuss the experimental results in Section 5. Finally, we conclude with future directions in Section 6.

2 PARENTAL MODIFICATIONS IN INFANT-DIRECTED INTERACTIONS

It is well known that parents significantly alter the acoustic characteristics of their speech when talking to infants (e.g., [19]). They, for example, raise the overall pitch of their voice, use wider pitch, slow the tempo, and increase the stress. These phenomena, called *motherese*, are suggested to have the effects of attracting the attention of infants and providing easily structured sentences to infants, which consequently facilitates their language learning.

In contrast to motherese, motionese is phenomena of parental modifications in their actions. Parents tend to modify their actions when interacting with infants so that they maintain the attention of infants and highlight the structure and the meaning of the actions as in motherese. Brand et al. [5] revealed that mothers altered their actions when demonstrating the usage of novel objects to their infants. They videotaped mothers' interactions first with an infant and then with an adult, and manually coded them on eight dimensions: the proximity to the partner, the interactiveness, the enthusiasm, the range of the motion, the repetitiveness, the simplification, the punctuation, and the rate. Their results comparing the infant-directed interactions (IDI) and adult-directed interactions (ADI) revealed significant differences in the first six dimensions out of the eight (higher rates in IDI than in ADI). Masataka [22] focused on a signed language and found that deaf mothers also altered their signed language. He observed deaf mothers when interacting with their deaf infants and when interacting with their deaf adult friends, and analyzed the characteristics of their signs. His comparison indicated that, when interacting with infants, deaf mothers significantly slowed the tempo of signs, frequently repeated the same signs, and exaggerated each sign. His further experiments showed that such modifications in a signed language attracted greater attention of both deaf and hearing infants [23, 24]. Gogate et al. [14] investigated the relationship between maternal gestures and speech in a object-naming task. They asked mothers to teach their infants novel words by using distinct objects and observed how the mothers used their gestures along with their speech. Their results showed that mothers used the target words more often than non-target words in temporal synchrony with the motion of the objects. They thus suggested that maternal gestures likely highlighted the relationship between target words and objects, of which effects were demonstrated in their further experiment [13]. Iverson et al. [18] also revealed that maternal gestures tended to co-occur with speech, to refer to the immediate context, and to reinforce the message conveyed in speech in daily mother-infant interactions. Their analysis moreover showed positive relationships between the production of maternal gestures and the verbal and gestural productions and the vocabulary size of infants.

In contrast to the former studies, in which motionese was manually coded, Rohlfing and her colleagues [12, 30] applied a computational technique to evaluate motionese. They adopted a 3D body tracking system [32], which was originally developed for human-robot interactions, to detect the trajectory of a parent's hand when he/she was demonstrating a stacking-cups task to his/her infant first and then to an adult. Their quantitative analysis revealed that parents put longer and more pauses between actions and decomposed a rounded movement into several linear movements in IDI compared with in ADI. They suggested with these results that motionese can help infants and robots to detect the meaning of actions. This approach is very attractive for robotics researchers because their model can be immediately implemented into robots and enables them to leverage the advantages of motionese in imitation learning. However, it is still an open question how robots know "what to imitate." Although their study as well as the former studies showed that parents modify their task-relevant actions so as to be easily understood, robots as well as young infants do not know which parents' actions are relevant to a task. To address this problem, we apply a model of saliency-based visual attention to the analysis of motionese.

3 SALIENCY-BASED VISUAL ATTENTION

3.1 Architecture of model

To analyze motionese from an infant-like viewpoint, i.e., without any a priori knowledge about actions or objects used in the actions, we adopt a model of saliency-based visual attention [16, 17]. The model, inspired by the behavior and the neuronal mechanism of primates, can simulate the attention shift of humans when they see natural scenes. Humans are able to rapidly detect and gaze at salient locations in their views. A salient location here is defined as a location which locally stands out from the surroundings because of its color, intensity, orientation, flicker, and motion [16]. For example, when we see a white ball in a green field, we can rapidly detect and look at the ball because of its outstanding color, intensity, and orientation. When a dot is moving left while a number of dots moving right, the former dot will be tracked visually because of its distinguished motion. The model of saliency-based visual attention imitates such a primal but adaptable attention mechanism of humans.

Figure 1 shows the overview of the model used in our experiment. This is the same as the model proposed in [16] excepting the absence of the mechanism of "inhibition of return," which inhibits the saliency of locations that have been gazed at. It means that our model determines attended locations frame by frame independently. The model works as follows:

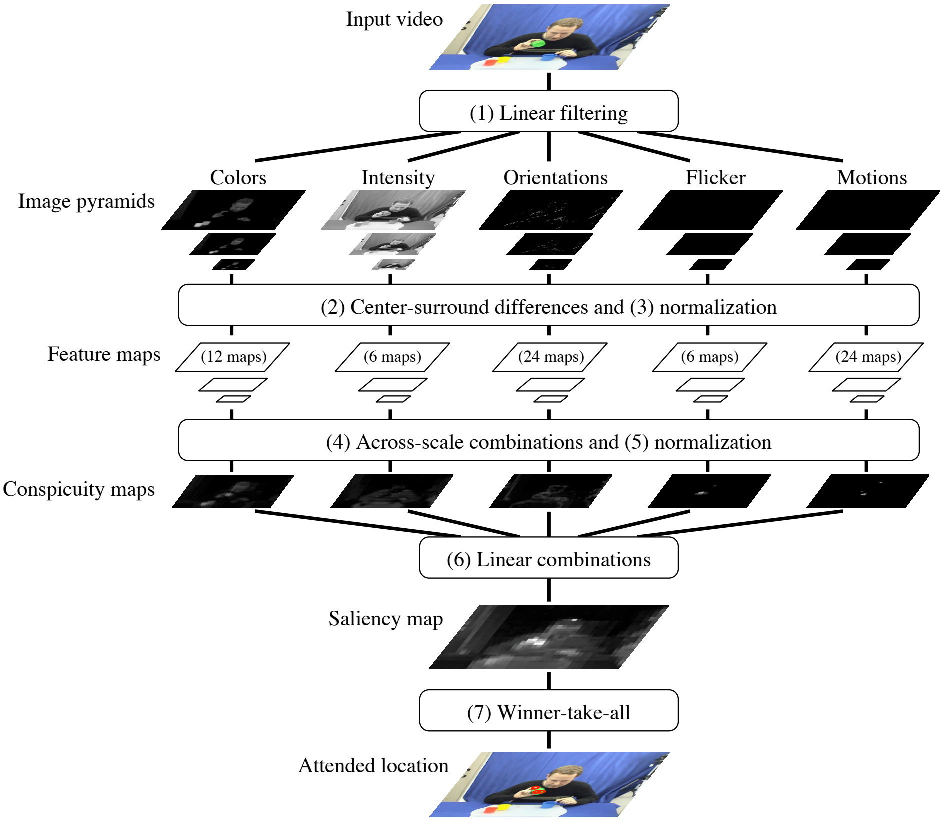


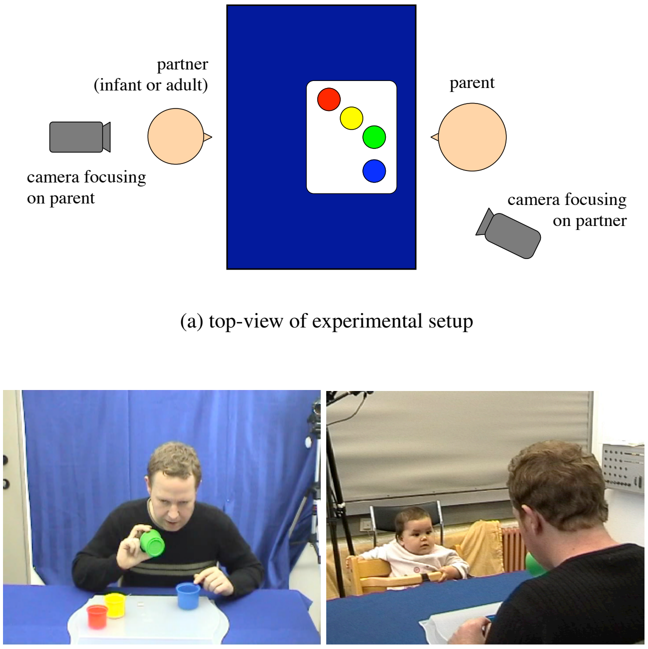
Figure 1. A model of saliency-based visual attention, which was revised from original one proposed in [17]

1. Five visual features (colors, intensity, orientations, flicker, and motions) are first extracted by linearly filtering a frame of an input video, and then image pyramids with different scales are created.
2. The differences between a center-fine scale and a surround-coarser scale image are calculated to detect how much each location stands out from the surroundings.
3. The center-surround differences are normalized to first eliminate modality-dependent differences and then globally promote maps containing a few conspicuous locations while globally suppressing maps containing numerous conspicuous peaks. The results are called feature maps.
4. The feature maps are combined through the across-scale addition to get together the different scales into one map.
5. The combined maps are normalized again to obtain conspicuity maps.
6. The conspicuity maps of the five features are linearly summed into a saliency map.
7. Finally, the most salient locations in the saliency map are selected as the attended locations in the frame.

In our analysis, image locations of which saliency were higher than the maximum × 0.9 in each frame were selected as the attended locations. That is, not only one location but several locations could be attended to in a frame. Refer to [16, 17] for more detail explanations of the processing.

3.2 Benefit of applying model to analysis of motionese

Applying the model to the analysis of motionese enables us to reveal what visual features of parental actions are highlighted in their action streams and could attract the attention of young infants and robots. Over the first year of life, infants semantic knowledge of actions, such as environmental, social, and psychological constraints on their organization and structure, is quite limited in comparison to adults. Thus, infants do not clearly understand the meaning or the structure of the actions when they see the actions for the first time. They also have limited information about objects, e.g., what objects are involved in the actions and what the initial and goal states of the objects are. Instead, they are certainly able to detect and gaze at salient locations in their views. For example, when colorful toys are shown to infants (usually, infants' toys have bright colors like yellow, red, and blue), they will look at the toys because of their salient colors. When a parent moves his/her hand to grasp and manipulate the toys, the hand as well as the toys will attract the attention of infants. Assuming only perceptual saliency, a parent's face can also attract the infants' attention because both of its static visual features and of its movement caused by his/her smiling and talking. Note that a parent's face and his/her hands can be attended to as salient locations without supposing any specific capability to detect their features or even skin color. We aim at evaluating how much meaningful structures of parental actions are detected without any knowledge about actions, objects, or humans, and how they can contribute to solving the problem of "what to imitate."



(a) top-view of experimental setup

(b) image frame of video focusing on parent, which was used as input to saliency-based attention model

(c) image frame of video focusing on infant

Figure 2. Experimental setup and sample image frames of videos

4 ANALYSIS OF MOTIONESE WITH SALIENCY-BASED ATTENTION MODEL

4.1 Method

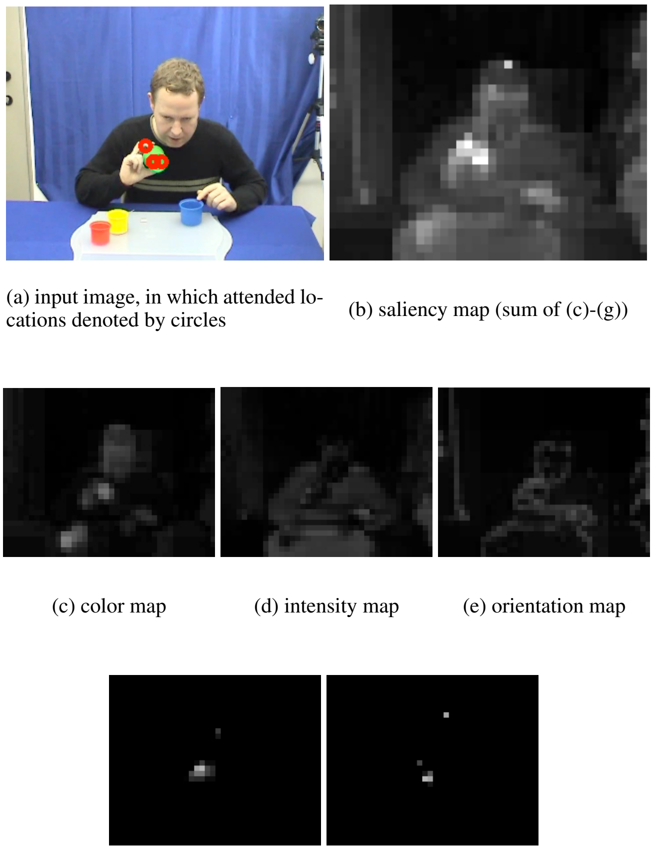
We analyzed the videotaped data used in [30]. In contrast to [30], in which only the task-related parental actions were analyzed, we dealt with all visual features in the scenes.

4.1.1 Subjects

Subjects were 15 parents (7 fathers and 8 mothers) of preverbal infants at the age of 8 to 11 months ($M = 10.56$, $SD = 0.89$). We chose this age because infants start to imitate simple means-end actions such as acting on one object to obtain another [33] and to show the understanding of goal-directed actions at 6 months of age [20].

4.1.2 Procedure

Parents were instructed to demonstrate a stacking-cups task to an interaction partner while explaining him/her how to do it. The interaction partner was first their infants and then an adult. **Figure 2** (a) illustrates the top-view of the experimental setup, and (b) and (c) show sample image frames of cameras which were set behind a parent and a partner and focused on each of them. The stacking-cups task was to sequentially pick up the green, the yellow, and the red cups and put them into the blue one on the white tray.



(a) input image, in which attended locations denoted by circles

(b) saliency map (sum of (c)-(g))

(c) color map

(d) intensity map

(e) orientation map

(f) flicker map

(g) motion map

Figure 3. Example of saliency map equally summing up five conspicuity maps and attended locations

4.1.3 Analysis

We analyzed videos recording the parents' actions as shown in Figure 2 (b). The videos were input to the model of saliency-based visual attention, and image locations with high saliency were detected as the attended locations frame by frame. **Figure 3** shows how the attended locations were determined in a frame: (a) shows an input image (320 × 256 [pixels]), in which three attended locations are denoted by red circles, and (b) shows the saliency map of the scene (40 × 32 [pixels]), which sums up the five conspicuity maps: (c) the color, (d) the intensity, (e) the orientation, (f) the flicker, and (g) the motion maps. The view of the maps corresponds to the input image, and the brightness of the pixels represents the degree of saliency, i.e., white means high saliency while black means low. In the example, the father was showing the green cup to his infant by shaking it, and therefore the cup and his right hand were attended to by the model. The color map extracted the green, the yellow, and the red cups as well as the father's face and hands as salient locations, while the intensity map detected the white tray and the father's black cloth. The orientation map detected the father's face, his hands, and the contour of the tray because of their rich edges. The flicker and the motion maps extracted the father's right hand with the green cup because of their movement. As a result, the saliency map, which equally summed up the five conspicuity maps, detected the three highly salient locations in the scene (see Figure 3 (a)). Note that our model selected the locations of which saliency was higher than the maximum × 0.9 in each frame, which allows us to evaluate the general tendency of parental actions. Through our experiment, the blue cup was not salient due to the blue background.

4.2 Results

4.2.1 Proportion of attended locations

We first compared how often a parent's face, his/her hands, and the cups were attended to by the model in IDI and in ADI. The attended locations were automatically classified using the predefined colors and positions of the targets. The results were compared separately in three time periods: before, during, and after the task. The start and the end of the task were defined when a parent picked up the first cup and when he/she put down the final cup into the blue one, respectively. The length of the periods before and after the task was 2 [sec].

Figures 4, **5**, and **6** show the results for the periods before, during, and after the task. In each graph, the horizontal axis denotes the label of the subjects, and the vertical axis denotes the proportion at which (a) a parent's face, (b) his/her hands, and (c) the cups were attended to over the period. When an attended location was at none of them, e.g., at a parent's cloth and at the tray, it was counted as (d) the others. The means and the standard deviations are listed in **Table 1**.

Table 1. Proportions of attended locations

		IDI		ADI	
		M	SD	M	SD
before task	parent's face	0.070	0.104	0.049	0.047
	parent's hands	0.583	0.171	0.521	0.192
	cups	0.289	0.145	0.196	0.185
	others	0.216	0.184	0.356	0.220
during task	parent's face	0.040	0.038	0.019	0.017
	parent's hands	0.680	0.150	0.715	0.127
	cups	0.448	0.117	0.433	0.112
	others	0.089	0.088	0.089	0.083
after task	parent's face	0.085	0.103	0.154	0.117
	parent's hands	0.484	0.311	0.475	0.239
	cups	0.306	0.198	0.180	0.123
	others	0.230	0.232	0.270	0.176

Before task: The non-parametric test (the Wilcoxon test) revealed significant differences in the proportion of attention on the cups (Figure 4 (c); $Z = -2.045,\ p < 0.05$) and in that on the others ((d); $Z = -1.988,\ p < 0.05$). It indicates that the cups attracted more attention in IDI than in ADI, and that the others were less attended to in IDI than in ADI.

During task: The non-parametric test revealed a significant difference in the proportion of attention on a parent's face (Figure 5 (a); $Z = -2.556,\ p < 0.05$). It also showed a statistical trend in the proportion of attention on parent's hands ((b); $Z = -1.817,\ p = 0.069$). A parent's face attracted much more attention in IDI than in ADI while his/her hands attracted less attention in IDI than in ADI.

After task: The non-parametric test revealed a statistical trend in the proportion of attention on a parent's face (Figure 6 (a); $Z = -1.874,\ p = 0.061$). The parametric t-test showed a trend in the proportion of attention on the cups ((c); $t(14) = 1.846$, $p = 0.086$). These results suggest that a parent's face was attended to in ADI more than in IDI, and that the cups were attended to in IDI more than in ADI.

4.2.2 Contribution of static features to saliency of objects

We next analyzed how much the static visual features of the cups contributed to their saliency in IDI and in ADI. Here the static features include the color, the intensity, and the orientation while the motion features include the flicker and the motion. The sum of the degrees of saliency derived from the static features was compared between IDI and ADI.

Figure 7 shows the contribution rate of the static features to the saliency of the cups (a) before, (b) during, and (c) after the task. **Table 2** lists the means and the standard deviations. The non-parametric test (the Wilcoxon test) revealed significant differences in the contribution rates before the task (Figure 7 (a); $Z = -2.040,\ p < 0.05$) and during the task ((b); $Z = -3.045,\ p < 0.05$). It indicates that in the two time periods the static features much more contributed to the saliency of the cups in IDI than in ADI.

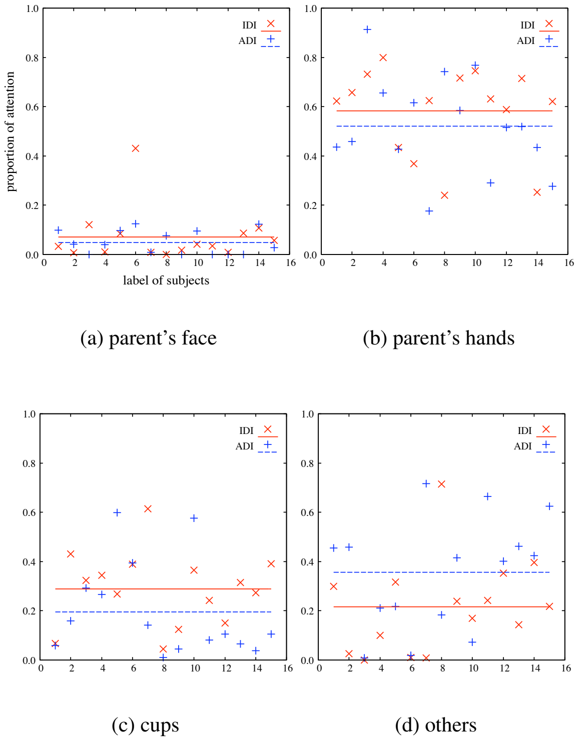
Table 2. Contribution of static features to saliency of cups

	IDI		ADI	
	M	SD	M	SD
before task	0.461	0.331	0.240	0.267
during task	0.256	0.203	0.090	0.100
after task	0.650	0.349	0.421	0.405

5 DISCUSSIONS

Our first focus of analysis revealed that a parent's face attracted much more attention in IDI than in ADI during the task while it attracted less attention in IDI than in ADI after the task. A reason is that the parents in IDI often talked to and smiled at their infants when demonstrating the task. They commented on each action while executing it, tried to maintain the infants' attention by addressing them verbally, and tried to get the infants interested in the task by showing emotional expressions. These behaviors caused movements on the parents' faces and made them more salient than others (see **Figure 8** (a)). By contrast, in ADI the parents rarely talked to or smiled at the adult partner during the task but explained the task after finishing it. Thus, their faces attracted more attention after the task. The result that the parents' hands were more attended to in ADI than in IDI during the task also indicates that their faces did not often move compared to their hands. We suggest from these results that parents give their infants immediate feedback on their actions, which helps infants to detect what actions are important and relevant to the task.

Our further analysis focusing on the objects involved in the task revealed that the objects were more salient in IDI than in ADI before and after the task. The saliency emerged because the parents interacting with their infants tended to put longer pauses before and after the task. While many of the parents in ADI started the task without checking whether the adult partner looked at the task-relevant locations, in IDI, they looked at the infants first and then started the task after confirming the infants' attention on the cups (see Figure 8 (b)). They also tried to attract the infants' attention on the cups by shaking them before the task. The result that the other locations attracted less attention in IDI than in ADI before the task also indicates that the parents made much effort to attract the attention of infants on the



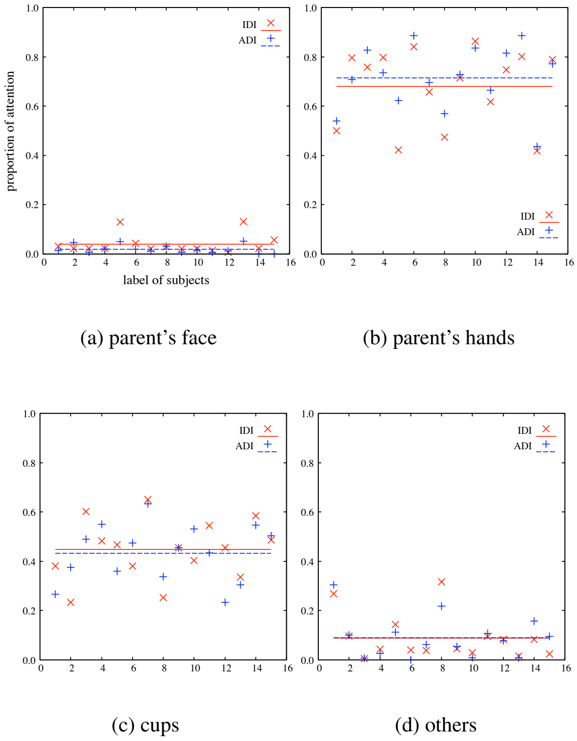
(a) parent's face

(b) parent's hands

(c) cups

(d) others

Figure 4. Proportions of attended locations before task (2 [sec])



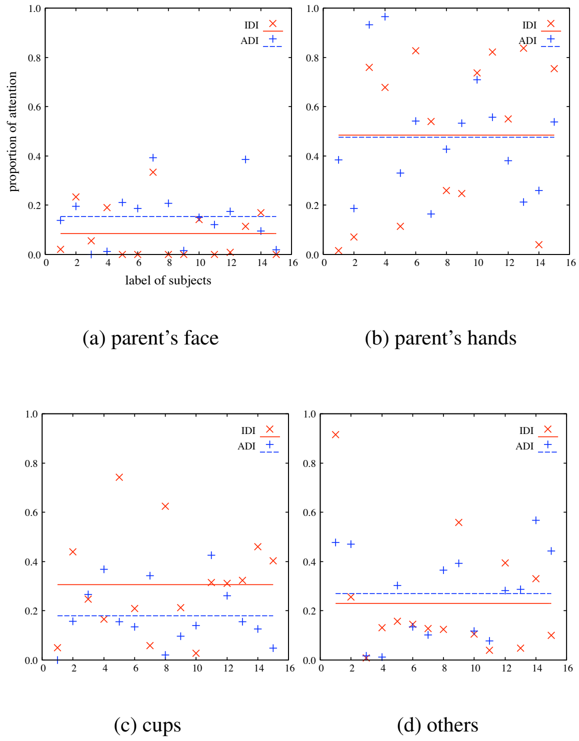
(a) parent's face

(b) parent's hands

(c) cups

(d) others

Figure 5. Proportions of attended locations during task



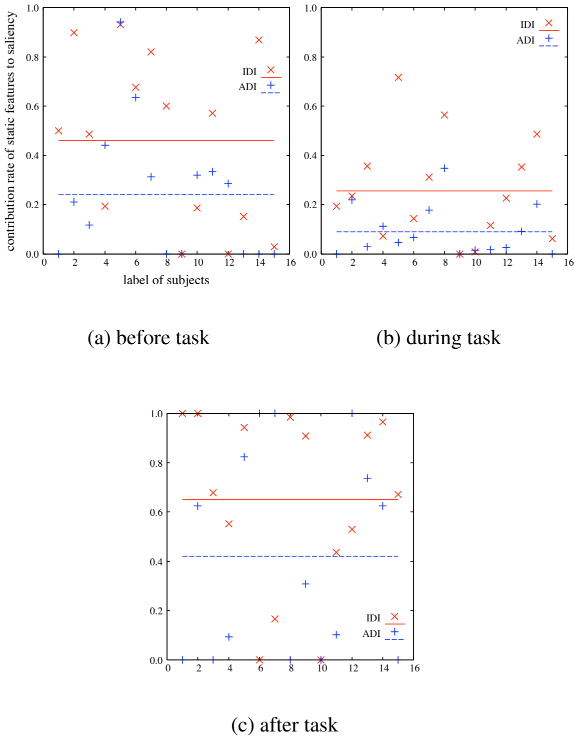
(a) parent's face

(b) parent's hands

(c) cups

(d) others

Figure 6. Proportions of attended locations after task (2 [sec])

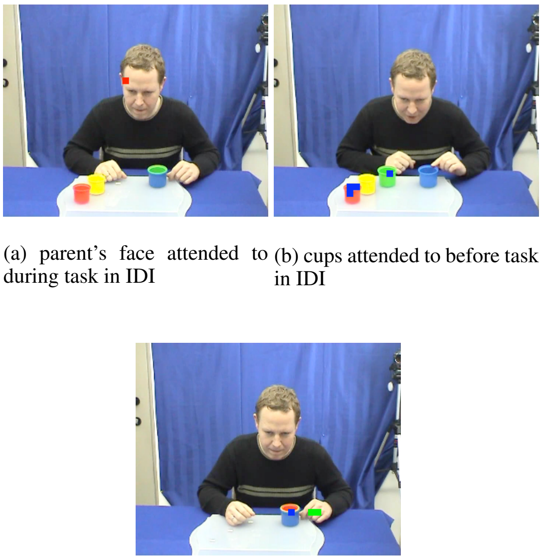


(a) before task

(b) during task

(c) after task

Figure 7. Contribution of static features to saliency of cups



(a) parent's face attended to during task in IDI

(b) cups attended to before task in IDI

(c) cups attended to after task in IDI

Figure 8. Examples of attended locations, which are indicated by a red, a green, or a blue box if they are on a parent's face, on his/her hands, or on the cups, respectively

task-related locations. In addition, the parents in IDI tended to stop their movement and look at the infants for a while after the task (see Figure 8 (c)) while the parents in ADI continued to move and commented a lot on the task. They likely showed the goal state of the task to the infants. We therefore suggest that parents aid their infants detecting the initial and goal states of the actions by inserting longer pauses before and after the task.

Our analysis on the contribution of the static features to the saliency of the objects showed that the features of the color, the intensity, and the orientation of the cups contributed much more to their saliency in IDI than in ADI. When the cups are attended to as salient locations, two reasons are considered: motion and static visual features. In IDI the saliency of the cups was derived not only from their movement but also from their intrinsic features, i.e., the color, the intensity, and the orientation, while in ADI the saliency was mostly came from their movement. The reason is that the parents in IDI often stopped their movement during the demonstration of the task and tried to attract the infants' attention not on their hands' motion but on the cups they were holding. Thus, the cups were attended to as salient locations because of their intrinsic features. We suggest with these results that parental actions help infants to detect the static features of the objects, which consequently enables them to better perceive the physical structure of the objects.

Although these findings are already very significant, some results are considered to be improved. Our analysis, for example, found a trend but did not reveal a statistically significant difference between the proportions of attention on the cups in IDI and in ADI after the task. Before the experiment, we hypothesized that the cups would attract much more attention in IDI than in ADI after the task as before the task. The reason why the cups were not so salient after the task is the blue background. In the goal state, all of the green, the yellow, and the red cups were put in the blue one, which means only the blue one was visible. Thus, the blue cup in the blue background was not detected as a salient location. We will therefore analyze other tasks using other colored objects to evaluate our hypothesis.

The position of the camera which recorded parents' actions also can be optimized. The camera was set higher than the head position of infants so that the view of the camera was not occluded by the infants. This position caused less saliency of the parents' faces because they always looked down to gaze at infants. We will thus change the position of the camera so that we can analyze motionese from a real infant viewpoint.

6 CONCLUSION

Our analysis on parental actions using a saliency-based attention model revealed that motionese can help infants (1) to receive immediate social feedback on the actions, (2) to detect the initial and goal states of the objects used in the actions, and (3) to look at the static features of the objects. In imitation learning, immediate feedback on the actions may allow infants to detect what actions are important and should be imitated. To look at the initial and goal states of the objects may be helpful in understanding the intention of the actions and in imitating the actions not only at the trajectory level but also at the goal level. To attend to the static features of the objects may also help infants to perceive the structure and the configuration of the objects. Therefore, all these results indicate that parental actions contribute to highlight the meaningful structures of the actions. We conclude that motionese can help infants to detect "what to imitate" and that the saliency-based attention model enables a robot to leverage these advantages in its imitation learning.

In contrast to current studies on robot imitation, in which a robot was given the knowledge about task-related actions and/or the goal of actions, our analysis showed that motionese enables a robot to detect these features autonomously. The model of saliency-based visual attention could highlight them in the sequences of parental actions. However, to solve the problem of "what to imitate," we still need to answer the following question. Which characteristics of actions, i.e., the trajectory or the goal of actions, should be imitated? We intend to further analyze motionese with respect to this problem.

We will also address the issue of "how to imitate." A robot that attempts to imitate human actions has to know how to transform the human movement into its own movement. To approach this problem, we propose a simple mapping from human movement detected in a robot's vision to the motion primitives of the robot represented in its somatic sense is enough to make the robot roughly imitate the actions [15, 27]. The motion primitives are designed with a set of neurons that are responsible to different motion directions while human movement is also detected and represented with neurons that are responsible to different motion directions [27]. We will develop such a mechanism and evaluate together with the attention model if they enable robots to imitate human actions by leveraging motionese.

REFERENCES

[1] Aris Alissandrakis, Chrystopher L. Nehaniv, and Kerstin Dautenhahn, 'Action, state and effect metrics for robot imitation', in *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication*, (2006).

[2] Minoru Asada, Karl F. MacDorman, Hiroshi Ishiguro, and Yasuo Kuniyoshi, 'Cognitive developmental robotics as a new paradigm for the design of humanoid robots', *Robotics and Autonomous Systems*, **37**, 185–193, (2001).

[3] Aude Billard, 'Learning motor skills by imitation: A biologically inspired robotic model', *Cybernetics and Systems: An International Jounal*, **32**, 155–193, (2001).

[4] Aude G. Billard, Sylvain Calinon, and Florent Guenter, 'Discriminative and adaptive imitation in uni-manual and bi-manual tasks', *Robotics and Autonomous Systems*, **54**(5), 370–384, (2006).

[5] Rebecca J. Brand, Dare A. Baldwin, and Leslie A. Ashburn, 'Evidence for 'motionese': modifications in mothers' infant-directed action', *Developmental Science*, **5**(1), 72–83, (2002).

[6] Cynthia Breazeal and Brian Scassellati, 'A context-dependent attention system for a social robot', in *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pp. 1146–1151, (1999).

[7] Cynthia Breazeal and Brian Scassellati, 'Challenges in building robots that imitate people', in *Imitation in Animals and Artifacts*, eds., K. Dautenhahn and C. L. Nehaniv, 363–389, MIT Press, (2002).

[8] Cynthia Breazeal and Brian Scassellati, 'Robots that imitate humans', *Trends in Cognitive Sciences*, **6**(11), 481–487, (2002).

[9] J. Gavin Bremner, *Infancy*, Blackwell Publishers Limited, 1994.

[10] Sylvain Calinon, Florent Guenter, and Aude Billard, 'Goal-directed imitation in a humanoid robot', in *Proceedings of the International Conference on Robotics and Automation*, (2005).

[11] Yiannis Demiris and Gillian Hayes, 'Imitation as a dual-route process featuring predictive and learning components: a biologically-plausible computational model', in *Imitation in Animals and Artifacts*, eds., K. Dautenhahn and C. L. Nehaniv, 321–361, MIT Press, (2002).

[12] Jannik Fritsch, Nils Hofemann, and Katharina Rohlfing, 'Detecting 'when to imitate' in a social context with a human caregiver', in *Proceedings of the ICRA Workshop on Social Mechanisms of Robot Programming by Demonstration*, (2005).

[13] Lakshmi J. Gogate and Lorraine E. Bahrick, 'Intersensory redundancy and 7-month-old infants' memory for arbitrary syllable-object relations', *Infancy*, **2**(2), 219–231, (2001).

[14] Lakshmi J. Gogate, Lorraine E. Bahrick, and Jilayne D. Watson, 'A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures', *Child Development*, **71**(4), 878–894, (2000).

[15] Verena V. Hafner and Yukie Nagai, 'Imitation behaviour evaluation in human robot interaction', in *Proceedings of the 6th International Workshop on Epigenetic Robotics*.

[16] L. Itti, N. Dhavale, and F. Pighin, 'Realistic avatar eye and head animation using a neurobiological model of visual attention', in *Proceedings of the SPIE 48th Annual International Symposium on Optical Science and Technology*, pp. 64–78, (2003).

[17] Laurent Itti, Christof Koch, and Ernst Niebur, 'A model of saliency-based visual attention for rapid scene analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(11), 1254–1259, (1998).

[18] Jana M. Iverson, Olga Capirci, Emiddia Longobardi, and M. Cristina Caselli, 'Gesturing in mother-child interactions', *Cognitive Development*, **14**, 57–75, (1999).

[19] Joseph L. Jacobson, David C. Boersma, Robert B. Fields, and Karen L. Olson, 'Paralinguistic features of adult speech to infants and small children', *Child Development*, **54**, 436–442, (1983).

[20] I. Kiraly, B. Jovanovic, W. Prinz, G. Aschersleben, and G Gergely, 'The early origins of goal attribution in infancy', *Consciousness and Cognition*, **12**, 752–769, (2003).

[21] Yasuo Kuniyoshi, Masayuki Inaba, and Hirochika Inoue, 'Learning by watching: Extracting reusable task knowledge from visual observation of human performance', *IEEE Transactions on Robotics and Automation*, **10**, 799–822, (1994).

[22] Nobuo Masataka, 'Motherese in a signed language', *Infant Behavior and Development*, **15**, 453–460, (1992).

[23] Nobuo Masataka, 'Perception of motherese in a signed language by 6-month-old deaf infants', *Developmental Psychology*, **32**(5), 874–879, (1996).

[24] Nobuo Masataka, 'Perception of motherese in japanese sign language by 6-month-old hearing infants', *Developmental Psychology*, **34**(2), 241–246, (1998).

[25] Andrew N. Meltzoff and M. Keith Moore, 'Imitation of facial and manual gestures by human neonates', *Science*, **198**, 75–78, (1977).

[26] Andrew N. Meltzoff and M. Keith Moore, 'Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms', *Developmental Psychology*, **25**(6), 954–962, (1989).

[27] Yukie Nagai, 'Joint attention development in infant-like robot based on head movement imitation', in *Proceedings of the Third International Symposium on Imitation in Animals and Artifacts*, pp. 87–96, (2005).

[28] Chrystopher L. Nehaniv and Kerstin Dautenhahn, 'Of hummingbirds and helicopters: An algebraic framework for interdisciplinary studies of imitation and its applications', in *Interdisciplinary Approaches to Robot Learning, World Scientific Series in Robotics and Intelligent Systems*, eds., J. Demiris and A. Birk, volume 24, (2000).

[29] Chrystopher L. Nehaniv and Kerstin Dautenhahn, 'Like me? measures of correspondence and imitation', *Cybernetics and Systems: An International Journal*, **32**, 11–51, (2001).

[30] Katharina J. Rohlfing, Jannik Fritsch, Britta Wrede, and Tanja Jungmann, 'How can multimodal cues from child-directed interaction reduce learning complexity in robot?', *Advanced Robotics*, **20**(10), 1183–1199, (2006).

[31] Stefan Schaal, 'Is imitation learning the route to humanoid robots?', *Trends in Cognitive Science*, **3**, 233–242, (1999).

[32] Joachim Schmidt, Jannik Fritsch, and Bogdan Kwolek, 'Kernel particle filter for real-time 3d body tracking in monocular color images', in *Proceedings of the Automatic Face and Gesture Recognition*, pp. 567–572, (2006).

[33] J. A. Sommerville and A. L. Woodward, 'Pulling out the intentional structure of action: the relation between action processing and action production in infancy', *Cognition*, **95**, 1–30, (2005).

[34] Ales Ude, Curtis Man, Marcia Riley, and Christopher G. Atkeson, 'Automatic generation of kinematic models for the conversion of human motion capture data into humanoid robot motion', in *Proceedings of the First IEEE-RAS International Conference on Humanoid Robots*, (2000).

A Theoretical Consideration on Robotic Imitation of Human Action According to Demonstration plus Suggestion

Masao Yokota [1]

Abstract. The Mental Image Directed Semantic Theory (MIDST) has proposed an omnisensory mental image model and its description language $\boldsymbol{L_{md}}$ intended to facilitate intuitive human-system interaction such that happens between non-expert people and home robots. The most remarkable feature of $\boldsymbol{L_{md}}$ is its capability of formalizing both temporal and spatial event concepts on the level of human/robotic sensations. This paper presents a brief sketch of $\boldsymbol{L_{md}}$ and a theoretical consideration on robotic imitation of human action driven by human suggestion interpreted in $\boldsymbol{L_{md}}$, controlling the robotic attention mechanism efficiently.

1 INTRODUCTION

Robotic or artificial imitation is one kind of machine learning on human actions and there have been reported a considerable number of studies on imitation learning from human actions demonstrated without any verbal hint [e.g., 1-3]. In this case, it is extremely difficult for a robot to understand which part of human demonstration is significant or not because there are too many things to attend to as it is. That is, it is an important issue where the attention of the observer should be focused on when a demonstrator performs an action. Whereas there have been several proposals to control attention mechanisms efficiently in such top-down ways as guided by the prediction or strategy based on sensory data and knowledge of goals or tasks [e.g., 4, 5, 14], they are not realistic when a large number of actions must be imitated distinctively with various speeds, directions, trajectories, etc.

The author has been working on integrated multimedia understanding for intuitive human-robot interaction, that is, interaction between non-expert or ordinary people and home robots, where natural language is the leading information medium for their intuitive communication [6, 12]. For ordinary people, natural language is the most important because it can convey the exact intention of the sender to the receiver due to its syntax and semantics common to its users, which is not necessarily the case for another medium such as gesture or so. Therefore, the author believes that it is most desirable to realize robotic imitation aided by human verbal suggestion where robotic attention to human demonstration is efficiently controllable based on semantic understanding of the suggestion.

For such a purpose, it is essential to develop a systematically computable knowledge representation language (KRL) as well as representation-free technologies such as neural networks for processing unstructured sensory/motory data. This type of language is indispensable to *knowledge-based* processing such as *understanding* sensory events, *planning* appropriate actions and *knowledgeable* communication with ordinary people in natural language, and therefore it needs to have at least a good capability of representing spatiotemporal events that correspond to humans'/robots' sensations and actions in the real world.

Most of conventional methods have provided robotic systems with such quasi-natural language expressions as 'move(*Velocity*, *Distance*, *Direction*)', 'find(*Object*, *Shape*, *Color*)' and so on for human instruction or suggestion, uniquely related to computer programs to deploy sensors/ motors [e.g., 7, 8]. In association with robotic imitation intended here, however, these expression schemas are too linguistic or coarse to represent and compute sensory/motory events in an integrated way.

The Mental Image Directed Semantic Theory (MIDST) [9] has proposed a model of human attention-guided perception yielding omnisensory images that inevitably reflect certain movements of the focus of attention of the observer (FAO) scanning certain matters in the world. More analytically, these omnisensory images are associated with spatiotemporal changes (or constancies) in certain attributes of the matters scanned by FAO and modeled as temporally parameterized "loci in attribute spaces", so called, to be formulated in a formal language $\boldsymbol{L_{md}}$. This language has already been implemented on several types of computerized intelligent systems [e.g., 10, 12].

This paper presents a brief sketch of the formal language $\boldsymbol{L_{md}}$ and a theoretical consideration on robotic imitation of human demonstrated action aided by human suggestion interpreted as semantic expression in $\boldsymbol{L_{md}}$. The most remarkable feature of $\boldsymbol{L_{md}}$ is its capability of formalizing spatiotemporal matter concepts grounded in human/robotic sensation while the other similar KRLs are designed to describe the logical relations among conceptual primitives represented by lexical tokens [e.g., 11]. In $\boldsymbol{L_{md}}$ expression are hinted what and how should be attended to in human action as analogy of human FAO movement and thereby the robotic attention can be controlled in a top-down way.

2 A BRIEF SKETCH OF L_{md}

An attribute space corresponds with a certain measuring instrument just like a barometer, thermometer or so and the loci represent the movements of its indicator. For example, the moving black triangular object shown in Figure 1 is assumed to be perceived as the loci in the three attribute spaces, namely, those of 'Location', 'Color' and 'Shape' in the observer's brain.

[1] Fukuoka Institute of Technology, Japan, email: yokota@fit.ac.jp

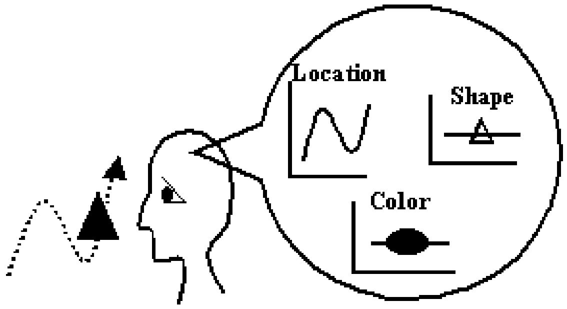


Figure1. Mental image model

Such a locus is to be articulated by “Atomic Locus” with an ***absolute*** time-interval [t_i, t_f] ($t_i< t_f$) as depicted in Figure 2 (up) and formulated as (1).

$L(x,y,p,q,a,g,k)$ (1)

This formula is called ‘Atomic Locus Formula’ whose first two arguments are often referred to as ‘Event Causer (EC)’ and ‘Attribute Carrier (AC)’, respectively. A logical combination of atomic locus formulas defined as a well-formed formula (i.e., wff) in predicate logic is called simply ‘Locus Formula’. The intuitive interpretation of (1) is given as follows, where ‘matter’ refers to ‘object’ or ‘event’ largely.

> ***“Matter ‘x’ causes Attribute ‘a’ of Matter ‘y’ to keep (p=q) or change ($p \neq q$) its values temporally ($g=G_t$) or spatially ($g=G_s$) over a time-interval, where the values ‘p’ and ‘q’ are relative to the standard ‘k’.”***

When $g=G_t$ and $g=G_s$, the locus indicates monotonic change or constancy of the attribute in time domain and that in space domain, respectively. The former is called ‘temporal event’ and the latter, ‘spatial event’. For example, the motion of the ‘bus’ represented by S1 is a temporal event and the ranging or extension of the ‘road’ by S2 is a spatial event whose meanings or concepts are formulated as (2) and (3), respectively, where A_{12} denotes ‘Physical Location’. These two formulas are different only at ‘Event Type (i.e., g)’.

(S1) The bus runs from Tokyo to Osaka.

$(\exists x,y,k)L(x,y,Tokyo,Osaka,A_{12},\mathbf{G_t},k)\wedge bus(y)$ (2)

(S2) The road runs from Tokyo to Osaka.

$(\exists x,y,k)L(x,y,Tokyo,Osaka,A_{12},\mathbf{G_s},k)\wedge road(y)$ (3)

The author has hypothesized that the difference between temporal and spatial event concepts can be attributed to the relationship between the Attribute Carrier (AC) and the Focus of the Attention of the Observer (FAO) [9]. To be brief, it is assumed that the FAO is fixed on the whole AC in a temporal event but *runs* about on the AC in a spatial event. According to this assumption, as shown in Figure 3, the *bus* and the FAO move together in the case of S1 while the FAO solely moves along the *road* in the case of S2.

Any locus in a certain Attribute Space can be formalized as a combination of atomic locus formulas and, so called, tempo-logical connectives, among which the most frequently used are ‘Simultaneous AND (Π)’ and ‘Consecutive AND (•)’ as appear in the conceptual definition (4) of the English verb ‘fetch’ depicted in Figure 2 (down).

$(\lambda x,y)fetch(x,y)\leftrightarrow(\lambda x,y)(\exists p_1,p_2,k)L(x,x,p_1,p_2,A_{12},G_t,k)\bullet$
$((L(x,x,p_2,p_1,A_{12},G_t,k)\Pi L(x,y,p_2,p_1,A_{12},G_t,k))\wedge x\neq y\wedge p_1\neq p_2$ (4)

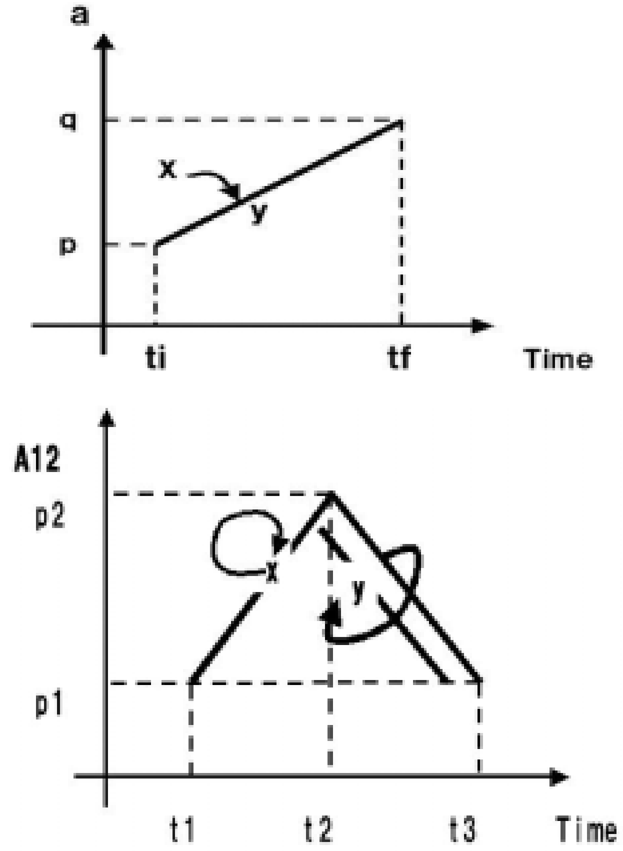


Figure 2. Atomic Locus (up) and Locus of ‘fetch’ (down)

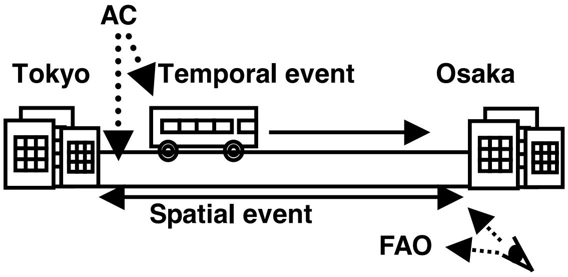


Figure 3. FAO movements and Event types

In order for explicit indication of time duration, ‘Empty Event (EE)’ denoted by ‘ε’ is introduced by the definition (5) with the attribute ‘Time Point (A34)’. According to this scheme, the duration [t_a, t_b] of an arbitrary locus χ can be expressed as (6).

$\varepsilon([t_1,t_2])\leftrightarrow(\exists x,y,g,k)\ L(x,y,t_1,t_2,A34,g,k)$ (5)

$\chi\ \Pi\ \varepsilon([t_a, t_b])$ (6)

All the same way, an object concept is also defined and expressed in L_{md} as a combination of potential events on its properties and its relations with others. For example, the conceptual descriptions of ‘rain’, ‘wind’ and ‘air’ can be given as (7)-(9), reading ‘Rain is water attracted from the sky by the earth, makes an object wetter, is pushed an umbrella to by a human,…,’ ‘Wind is air, affects the direction of rain,… ,’ and ‘Air has no shape, no taste, no vitality, …,’ respectively.

$$(\lambda x)\mathrm{rain}(x)\leftrightarrow(\lambda x)(\exists x_1,x_2,\ldots)L(\_,x,x_1,x_1,A_{41},G_t,\_)$$
$$\prod L(\mathrm{Earth},x,\mathrm{Sky},\mathrm{Earth},A_{12},G_t,\_)\prod L(x,x_2,p,q,A_{25},G_t,\_)$$
$$\prod L(x_3,x_4,x,x,A_{19},G_t,x_3)\wedge\mathrm{water}(x_1)$$
$$\wedge\mathrm{object}(x_2)\wedge\mathrm{human}(x_3)\wedge\mathrm{umbrella}(x_4)\wedge(p<q)\ldots \quad (7)$$
$$(\lambda x)\mathrm{wind}(x)\leftrightarrow(\lambda x)(\exists x_1,x_2,\ldots)L(\_,x,x_1,x_1,A_{41},G_t,\_)$$
$$\wedge\mathrm{air}(x_1)\wedge(L(x,x_2,p,q,A_{13},G_t,\_)\wedge\mathrm{rain}(x_2)\ldots \quad (8)$$
$$(\lambda x)\mathrm{air}(x)\leftrightarrow(\lambda x)(\ldots\wedge L^*(\_,x,/,/,A_{11},G_t,\_)\wedge\ldots\wedge$$
$$L^*(\_,x,/,/,A_{29},G_t,\_)\wedge\ldots\wedge L^*(\_,x,/,/,A_{39},G_t,\_)\wedge\ \ldots) \quad (9)$$

Hereafter, for simplicity of ***L_{md}*** expression, the special symbols ‘*’, ‘_’and ‘/’ are often employed to represent ‘always’, ‘something (or some value)’ and ‘nothing (no value)’ as defined by (10)-(12), respectively.

$$X^*\leftrightarrow(\forall[p,q])X\ \Pi\ \varepsilon([p,q]) \quad (10)$$
$$L(\ldots,\_,\ldots)\leftrightarrow(\exists\omega)L(\ldots,\omega,\ldots) \quad (11)$$
$$L(\ldots,/,\ldots)\leftrightarrow{\sim}(\exists p)\ L(\ldots,\omega,\ldots) \quad (12)$$

Table 1 shows about 50 attributes extracted exclusively from English and Japanese words of common use contained in certain thesauri [9]. Most of them (i.e., A01-A45) correspond to the sensory receptive fields in human brains. For example, those marked with ‘*’ in this table can be associated to the sense ‘sight’. Correspondingly, six categories of standards shown in Table 2 have been extracted that are necessary for representing relative values of each attribute in Table 1. ***These tables show that ordinary people live their casual lives, attending to tens of attributes of the matters in the world to cognize them in comparison with several kinds of standards.***

Table 1. List of attributes

Code	Attribute [Property†] (*words/phrases concerned*)
*A01	PLACE OF EXISTE NCE [N] (*happen, perish*)
*A02	LENGTH [S] (*long, shorten, close, away*)
*A03	HEIGHT [S] (*high, lower*)
*A04	WIDTH [S] (*widen, narrow*)
*A05	THICKNESS [S] (*thick, thin*)
*A06	DEPTH1 [S] (*deep, shallow*)
*A07	DEPTH2 [S] (*deep, concave*)
*A08	DIAMETER [S] (*across, in diameter*)
*A09	AREA [S] (*square meters, acre*)
*A10	VOLUME [S] (*litter, gallon*)
*A11	SHAPE [N] (*round, triangle*)
*A12	PHYSICAL LOCATION [N] (*move, stay*)
*A13	DIRECTION [N] (*turn, wind, left*)
*A14	ORIENTATION [N] (*orientate, command*)
*A15	TRAJECTORY [N] (*zigzag, circle*)
*A16	VELOCITY [S] (*fast, slow*)
*A17	MILEAGE [S] (*far, near*)
A18	STRENGTH OF EFFECT [S] (*strong, powerful*)
A19	DIRECTION OF EFFECT [N] (*pull, push*)
A20	DENSITY [S] (*dense, thin*)
A21	HARDNESS [S] (*hard, soft*)
A22	ELASTICITY [S] (*elastic, flexible*)
A23	TOUGHNESS [S] (*fragile, stiff*)
A24	TACTILE FEELING [S] (*rough, smooth*)
A25	HUMIDITY [S] (*wet, dry*)
A26	VISCOSITY [S] (*oily, watery*)
A27	WEIGHT [S] (*heavy, light*)
A28	TEMPERATURE [S] (*hot, cold*)
A29	TASTE [N] (*sour, sweet, bitter*)
A30	ODOUR [N] (*pungent, sweet*)
A31	SOUND [N] (*noisy, silent, loud*)
*A32	COLOR [N] (*red, white*)
A33	INTERNAL SENSATION [N] (*tired, hungry*)
A34	TIME POINT [S] (*o’clock, elapse*)
A35	DURATION [S] (*hour, minute, long, short*)
A36	NUMBER [S] (*ten, quantity, number*)
A37	ORDER [S] (*first, last*)
A38	FREQUENCY [S] (*sometimes, frequent*)
A39	VITALITY [S] (*alive, dead, vivid*)
A40	SEX [S] (*male, female*)
A41	QUALITY [N] (*make, destroy*)
A42	NAME [V] (*name, token*)
A43	CONCEPTUAL CATEGORY [V] (*mammal*)
*A44	TOPOLOGY [V] (*in, out, touch*)
*A45	ANGULARITY [S] (*sharp, dull, right angle*)
B01	WORTH [N] (*improve, praise, deny, alright*)
B02	LOCATION OF INFORMATION [N] (*tell, hear*)
B03	EMOTION [N] (*like, hate*)
B04	BELIEF VALUE [S] (*believe, trust*)
	…………………………..

†S: scalar value, N: non-scalar value. *Attributes concerning the sense of sight.

Table 2. List of standards

Categories	Remarks
Rigid Standard	Objective standards such as denoted by measuring *units* (meter, gram, etc.).
Species Standard	The *attribute value ordinary* for a species. A *short train* is ordinarily longer than a *long pencil*.
Proportional Standard	‘*Oblong*’ means that the width is greater than the height at a physical object.
Individual Standard	*Much* money for one person can be too *little* for another.
Purposive Standard	One room large enough for a person’s *sleeping* must be too small for his *jogging*.
Declarative Standard	The origin of an order such as ‘next’ must be declared explicitly just as ‘next *to him*’.

3 INTELLIGENT SYSTEM IMAGES-M

3.1 System configuration

The intelligent system IMAGES-M [e.g., 10, 12] is assumed to be the main intelligence of the robot intended here. As shown in Figure 4, IMAGES-M is one kind of expert system equipped with five kinds of user interfaces for multimedia communication, that is, Sensory Data Processing Unit (SDPU), Speech Processing Unit (SPU), Picture Processing Unit (PPU), Text Processing Unit (TPU), and Action Data Processing Unit (ADPU) besides Inference Engine (IE) and Knowledge Base (KB). Each processing unit in collaboration with IE performs mutual conversion between each type of information medium and locus formulas.

IMAGES-M is a language-centered intelligent system in order to facilitate intuitive interaction between humans and robots. For comprehensible communication with humans, robots must understand natural language *semantically* and *pragmatically*. Here, as shown in Figure 5, semantic understanding means associating symbols to conceptual images of matters (i.e., objects or events), and pragmatic understanding means anchoring symbols to real matters by unifying conceptual images with perceptual images.

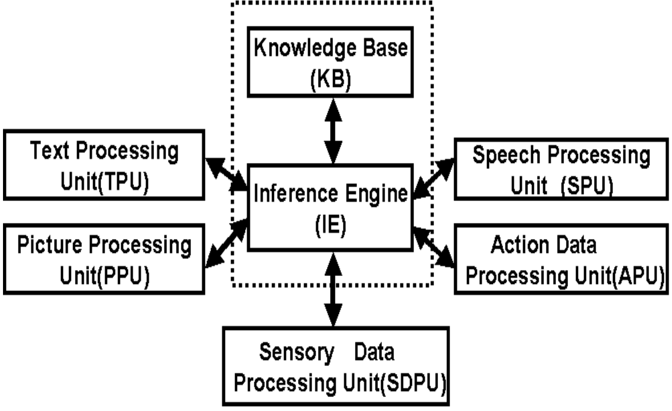

Knowledge Base (KB)
Text Processing Unit(TPU)
Inference Engine (IE)
Speech Processing Unit (SPU)
Picture Processing Unit(PPU)
Action Data Processing Unit(APU)
Sensory Data Processing Unit(SDPU)

Figure 4. Configuration of IMAGES-M

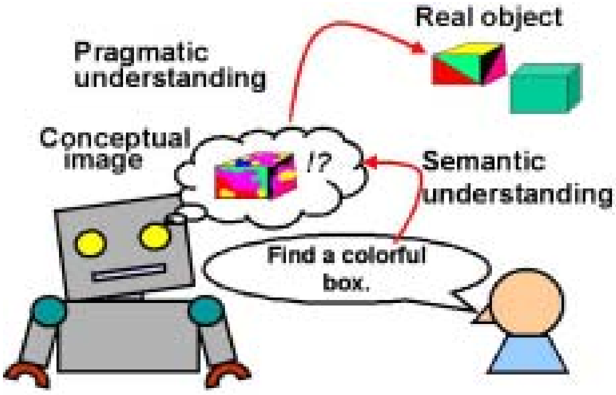


Figure 5. Semantic and pragmatic understanding

3.2 Semantic understanding

As shown in Figure 6, natural language expression (i.e, surface structure) and $\boldsymbol{L_{md}}$ expression (i.e., conceptual structure) are mutually translatable through surface dependency structure by utilizing syntactic rules and word meaning descriptions [9].

A word meaning description M_w is defined by (13) as a pair of 'Concept Part (C_p)' and 'Unification Part (U_p)'.

$M_w \leftrightarrow [C_p{:}U_p]$ (13)

The C_p of a word W is a locus formula about properties and relations of the matters involved such as shapes, colors, functions, potentialities, etc while its U_p is a set of operations for unifying the C_ps of W's syntactic governors or dependents. For example, the meaning of the English verb 'carry' can be given by (14).

$[(\exists x,y,p_1,p_2)\ L(x,x,p_1,p_2,A12,Gt,\_)\Pi$
$L(x,y,p_1,p_2,A12,Gt,\_)\wedge x\neq y\wedge p_1\neq p_2{:}ARG(Dep.1,x);$
$ARG(Dep.2,y);]$ (14)

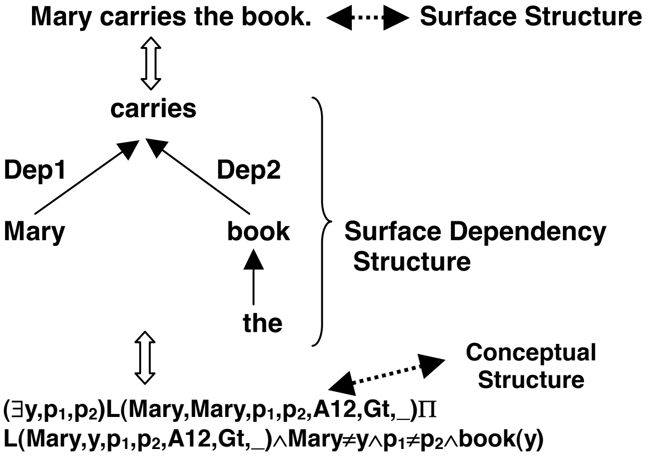

Mary carries the book.
Surface Structure
carries
Dep1
Dep2
Mary
book
the
Surface Dependency Structure
Conceptual Structure
(∃y,p1,p2)L(Mary,Mary,p1,p2,A12,Gt,_)Π
L(Mary,y,p1,p2,A12,Gt,_)∧Mary≠y∧p1≠p2∧book(y)

Figure 6. Mutual conversion between natural language and $\boldsymbol{L_{md}}$

(Input)

With the long red stick Tom precedes Jim.

(Output)

Tom with the long red stick goes before Jim goes.
Jim goes after Tom goes with the long red stick.
Jim follows Tom with the long red stick.
Tom carries the long red stick before Jim goes.
…………………

Figure 7. Paraphrasing as semantic understanding by IMAGES-M

The U_p above consists of two operations to unify the first dependent (Dep.1) and the second dependent (Dep.2) of the current word with the variables x and y, respectively. Here, Dep.1 and Dep.2 are the 'subject' and the 'object' of 'carry', respectively. Therefore, the surface structure '*Mary carries a book*' is translated into the conceptual structure (15) via the surface dependency structure shown in Figure 6.

$(\exists y,p_1,p_2)L(Mary,Mary,p_1,p_2,A12,Gt,\_)\Pi$
$L(Mary,y,p_1,p_2,A12,Gt,\_)\wedge Mary\neq y\wedge p_1\neq p_2\wedge book(y)$ (15)

For another example, the meaning description of the English preposition 'through' is also given by (16).

$[(\exists x,y,p_1,z,p_3,g,p_4)(\underline{L(x,y,p_1,z,A12,g,\_)}\bullet$
$L(x,y,z,p_3,A12,g,\_))\Pi\ L(x,y,p_4,p_4,A13,g,\_)\wedge p_1\neq z\wedge z\neq p_3$
$:ARG(Dep.1,z);\ IF(Gov=Verb)\rightarrow PAT(Gov,(1,1));$
$IF(Gov=Noun)\rightarrow ARG(Gov,y);]$ (16)

The U_p above is for unifying the C_ps of the very word, its governor (Gov, a verb or a noun) and its dependent (Dep.1, a noun). The second argument (1,1) of the command PAT indicates the underlined part of (13) and in general (i,j) refers to the partial formula covering from the ith to the jth atomic formula of the current C_p. This part is the pattern common to both the C_ps to be unified. This is called 'Unification Handle (U_h)' and when missing, the C_ps are to be combined simply with '$\wedge$'.

Therefore the sentences S3, S4 and S5 are interpreted as (17)-(19), respectively. The underlined parts of these formulas are the results of PAT operations. The expression (20) is the C_p of the adjective 'long' implying 'there is some value greater than some standard of 'Length (A02)' which is often simplified as (20').

(S3) The train runs through the tunnel.

$(\exists x,y,p_1,z,p_3,p_4)(\underline{L(x,y,p_1,z,A12,Gt,\_)}\bullet$
$L(x,y,z,p_3,A12,Gt,\_))\Pi\ L(x,y,p_4,p_4,A13,Gt,\_)$
$\wedge p_1\neq z\ \wedge z\neq p_3\wedge \mathrm{train}(y)\ \wedge \mathrm{tunnel}(z)$ (17)

(S4) The path runs through the forest.

$(\exists x,y,p_1,z,p_3,p_4)(\underline{L(x,y,p_1,z,A12,Gs,\_)}\bullet$
$L(x,y,z,p_3,A12,Gs,\_))\Pi\ L(x,y,p_4,p_4,A13,Gs,\_)$
$\wedge p_1\neq z\ \wedge z\neq p_3\ \wedge \mathrm{path}(y)\ \wedge \mathrm{forest}(z)$ (18)

(S5) The path through the forest is long.

$(\exists x,y,p_1,z,p_3,x_1,q,p_4,k_1)$
$(L(x,y,p_1,z,A12,Gs,\_)\bullet L(x,y,z,p_3,A12,Gs,\_))$
$\Pi\ L(x,y,p_4,p_4,A13,Gs,\_)\ \wedge L(x_1,y,q,q,A02,Gt,k_1)$
$\wedge p_1\neq z\wedge z\neq p_3\wedge q>k_1\wedge \mathrm{path}(y)\wedge \mathrm{forest}(z)$ (19)
$(\exists x_1,y_1,q,k_1)L(x_1,y_1,q,q,A02,Gt,k_1)\wedge q>k_1$ (20)
$(\exists x_1,y_1,k_1)L(x_1,y_1,\mathrm{Long},\mathrm{Long},A02,Gt,k_1)$ (20')

The process above is completely reversible except that multiple natural expressions as paraphrases can be generated by TPU in IMAGES-M as shown in Figure 7 because such event patterns as shown in Figure 2 are sharable among multiple word concepts. This is one of the most remarkable features of MIDST and is also possible between different languages as understanding-based translation [10, 12].

3.3 Pragmatic understanding

An event expressed in $\boldsymbol{L_{md}}$ is compared to a movie film recorded through a floating camera because it is necessarily grounded in FAO's movement over the event. For example, it is not the 'path' but the 'FAO' that 'sinks' in S6 or 'rises' in S7. Therefore, such expressions refer to the same scene pragmatically in spite of their appearances, whose semantic descriptions are given as (21) and (22), respectively, where 'A_{13}', '↑' and '↓' refer to the attribute 'Direction', and its values 'upward' and 'downward', respectively. This fact is generalized as '***Postulate of Reversibility of a Spatial Event*** (PRS)' belonging to people's intuitive knowledge about geography, and the conceptual descriptions (21) and (22) are called ***equivalent in the PRS***.

(S6) The path sinks to the brook.

$(\exists x,y,p,z)L(x,y,p,z,A_{12},G_s,\_)\Pi L(x,y,\downarrow,\downarrow,A_{13},G_s,\_)$
$\wedge \mathrm{path}(y)\ \wedge \mathrm{brook}(z)\ \wedge p\neq z$ (21)

(S7) The path rises from the brook.

$(\exists x,y,p,z)L(x,y,z,p,A_{12},G_s,\_)\Pi L(x,y,\uparrow,\uparrow,A_{13},G_s,k_2)$
$\wedge \mathrm{path}(y)\ \wedge \mathrm{brook}(z)\ \wedge p\neq z$ (22)

For another example of spatial event, Figure 8 (up) concerns human perception of the formation of multiple distinct objects, where FAO runs along an imaginary object so called 'Imaginary Space Region (ISR)'. This spatial event can be verbalized as S8 using the preposition 'between' and formulated as (22), corresponding also to such concepts as 'row', 'line-up', etc. Any type of topological relation between two objects is also to be formulated by employing an ISR. For example, S9 is translated into (23) or (23'), where '*In*', and '*Cont*' are the values 'inside' and 'contains' of the attribute 'Topology (A44)' represented by 3x3 matrices at the Sandard of '9-intersection model (*IM*)' [13], where '*In'* and *'Cont'* are the transposes each other.

(S8) □ is between Δ and ○.

$(\exists y,p)(L(\_,y,\Delta,\square,A_{12},G_s,\_)\bullet L(\_,y,\square,\circ,A_{12},G_s,\_))\Pi$
$L(\_,y,p,p,A_{13},G_s,\_)\ \wedge ISR(y)$ (22)

(S9) □ is in the room.

$(\exists x,y)L(\_,x,y,\square,A12,Gs,\_)\Pi L(\_,x,\mathit{In},\mathit{In},A44,Gt,\mathit{IM})$
$\wedge ISR(x)\wedge \mathrm{room}(y)$ (23)
$(\exists x,y)L(\_,x,\square,y,A12,Gs,\_)\Pi L(\_,x,\mathit{Cont},\mathit{Cont},A44,Gt,\mathit{IM})$
$\wedge ISR(x)\wedge \mathrm{room}(y)$ (23')

For more complicated examples, consider S10 and S11. The underlined parts are deemed to refer to some events neglected in time and in space, respectively. These events correspond with skipping of FAOs and are called 'Temporal Empty Event' and 'Spatial Empty Event', denoted by 'ε_t' and 'ε_s' as Empty Events with $g=G_t$ and $g=G_s$ at (5), respectively. Their concepts are described as (24) and (25), where 'A_{15}' and 'A_{17}' represent the attribute 'Trajectory' and 'Mileage', respectively. From the viewpoint of pragmatic understanding, the formula (25) can refer to such a spatial event depicted as the still picture in Figure 8 (down) while (24), a temporal event to be recorded as a movie.

(S10) The *bus* runs 10km straight east from A to B, and *<u>after a while</u>*, at C it meets the street with the sidewalk.

$(\exists x,y,z,p,q)(L(\_,x,A,B,A_{12},G_t,\_)\Pi$
$L(\_,x,0,10km,A_{17},G_t,\_)\Pi L(\_,x,\mathrm{Point},\mathrm{Line},A_{15},G_t,\_)\Pi$
$L(\_,x,\mathrm{East},\mathrm{East},A_{13},G_t,\_))\bullet\varepsilon_t\bullet(L(\_,x,p,C,A_{12},G_t,\_)$
$\Pi L(\_,y,q,C,A_{12},G_s,\_)\Pi L(\_,z,y,y,A_{12},G_s,\_))$
$\wedge \mathrm{bus}(x)\wedge \mathrm{street}(y)\wedge \mathrm{sidewalk}(z)\wedge p\neq q$ (24)

(S11) The *road* runs 10km straight east from A to B, and *<u>after a while</u>*, at C it meets the street with the sidewalk.

$(\exists x,y,z,p,q)(L(\_,x,A,B,A_{12},G_s,\_)\Pi$
$L(\_,x,0,10km,A_{17},G_s,\_)\Pi L(\_,x,\mathrm{Point},\mathrm{Line},A_{15},G_s,\_)\Pi$
$L(\_,x,\mathrm{East},\mathrm{East},A_{13},G_s,\_))\bullet\varepsilon_s\ \bullet(L(\_,x,p,C,A_{12},G_s,\_)$
$\Pi L(\_,y,q,C,A_{12},G_s,\_)\Pi L(\_,z,y,y,A_{12},G_s,\_))$
$\wedge \mathrm{road}(x)\wedge \mathrm{street}(y)\wedge \mathrm{sidewalk}(z)\wedge p\neq q$ (25)

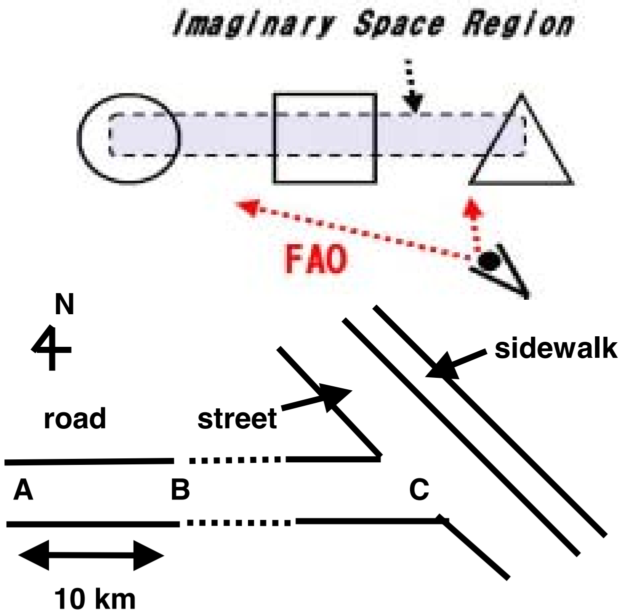
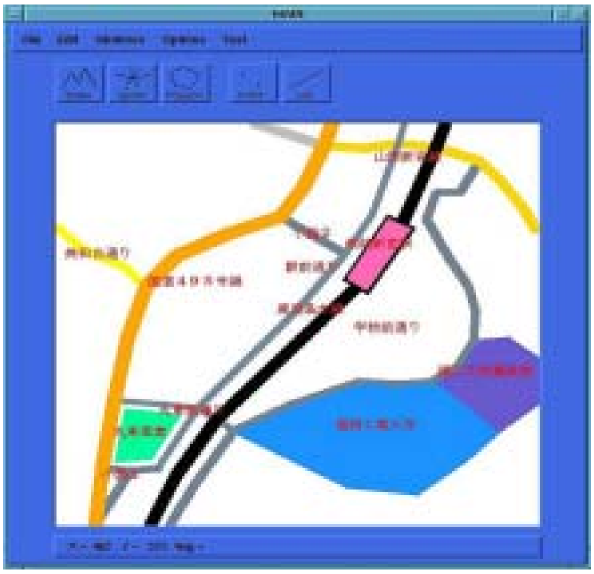


Figure 8. Complicated spatial events: 'row' (up) and 'example of road map' (down)



(a) A map generated from a locus formula by IMAGES-M

H: How does the national road run?
S: It extends between Pref. A and Pref. C via Pref. B.
H: Where does the bus go from the rail way station A?
S: It reaches the town D.
H: What is between the buildings A and B?
S: The railway D.
H: Where do the street A and the road B meet?
S: At the crossing C.
H: Where do the street A and the road B separate?
S: At the crossing C.

(b) Q-A on the map (a) by human (H) and IMAGES-M (S)

Figure 9. Cross-media operations as pragmatic understanding

Figures 9 (b) shows an example of question-answering on the real map (a) between a human and IMAGES-M [6, 10, 12], where the map is a pictorial interpretation of a locus formula by PPU. The system understood the query texts pragmatically by anchoring them to the map as a model of the real world, utilizing effectively several kinds of intuitive postulates such as PRS, as a matter of course, where distinction between temporal and spatial events is crucially important.

4 IMITATION GUIDED BY SUGGESTION

4.1 Definition

As shown in Figures 10 and 11, robotic imitation intended here is defined as a human-robot interaction where a human presents a robot a pair of demonstration and suggestion that is the expression of his/her intention and it behaviouralizes its conception, namely, the result of semantic and pragmatic understanding of the suggestion.

The processes shown in Figures 10 and 11 can be formalized as follows, where the pair of $\mathbf{P_i}$ and $\mathbf{Def_i}$ is called 'Conception' for the i-th imitation and denoted by $\mathbf{C_i}$.

$\mathbf{Int_i} \Rightarrow \mathbf{T_i}, \mathbf{D_i}$
$\mathbf{T_i}, \mathbf{K_L} \Rightarrow \mathbf{S_i}$
$\mathbf{D_i}, \mathbf{K_D} \Rightarrow \mathbf{Per_i}$
$\mathbf{S_i}, \mathbf{Per_i}, \mathbf{K_D} \Rightarrow \mathbf{P_i}, \mathbf{Def_i}$ (= $\mathbf{C_i}$)
$\mathbf{P_i}, \mathbf{Def_i}, \mathbf{K_D} \Rightarrow \mathbf{I_i}$

, where

$\mathbf{Int_i}$: The i-th intention by the human,
$\mathbf{T_i}$: The i-th suggestion by the human,
$\mathbf{S_i}$: Result of semantic understanding of the i-th suggestion,
$\mathbf{K_L}$: Linguistic knowledge in the robot,
$\mathbf{D_i}$: The i-th demonstration by the human,
$\mathbf{K_D}$: Domain-specific knowledge in the robot at the i-th session,
$\mathbf{Per_i}$: Perception of the i-th demonstration,
$\mathbf{P_i}$: Result of pragmatic understanding of the i-th suggestion,
$\mathbf{Def_i}$: Default specification for the i-th imitation,
$\mathbf{I_i}$: The i-th imitation by the robot,
$\Rightarrow$: Conversion process (e.g., inference, translation).

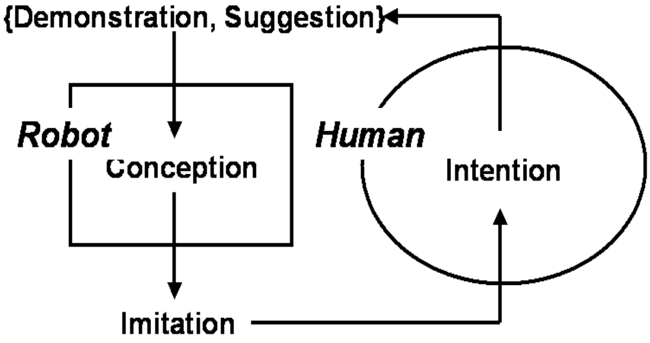


Figure 10. Imitation as human-robot interaction

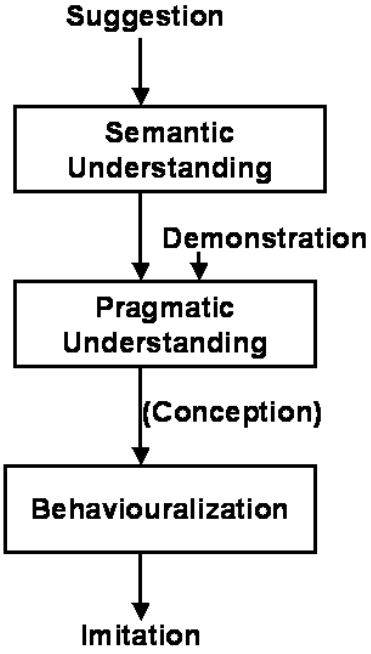


Figure 11. Imitation guided by suggestion

4.2 Theoretical simulation

As shown in Figure 10, it is assumed that there is a feedback loop between a human and a robot in order for the human to improve his/her previous suggestion or demonstration and for the robot to correct its previous imitation. For example, consider the scenario presented below and depicted in Figure12.

Scenario :

Robby is an intelligent humanoid robot and Tom is his user. Robby is called by Tom and enters Tom's room. This is Robby's first visit there. Robby sees Tom leftward and the brown pillar forward (, but doesn't see the green box or the yellow table). After a while, Tom tells Robby "Imitate me to my demonstration and suggestion."……

Here is described a theoretical simulation of the robotic imitation driven by the top-down control of the attention mechanism, which is almost that of problem finding/solving in the filed of AI [6, 12].

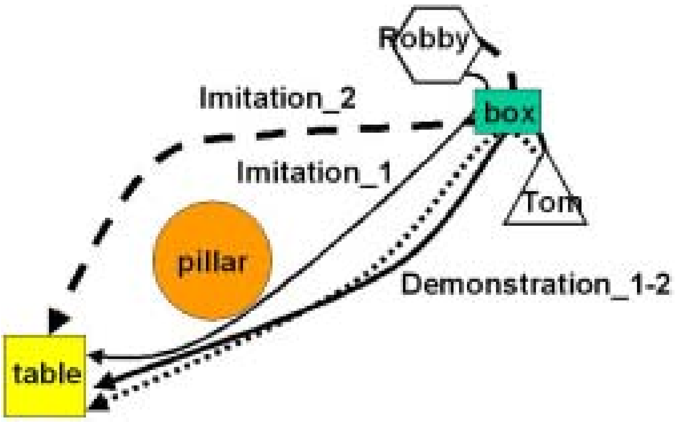


Figure 12. Tom's demonstrations and Robby's imitations

The sequence of the events assumed to happen is as follows.

[Robby's Perception of the initial situation, ***Sit₀***]

$\boldsymbol{Sit_0}\leftrightarrow L(\_,O_{21},Brown,Brown,A_{32},G_{t},\_)\Pi$
$L(\_,O_{22},Robby,Tom,A_{12},G_{s},\_)\Pi$
$L(\_,O_{22},Lw_{21},Lw_{21},A_{13},G_{s},Robby)\Pi$
$L(\_,O_{23},Robby,O_{21},A_{12},G_{s},\_)\Pi$
$L(\_,O_{23},Fw_{21},Fw_{21},A_{13},G_{s},Robby)$
$\wedge pillar(O_{21})\wedge ISR(O_{22})\wedge ISR(O_{23})$

Robby's perception of the situation (i.e., the underlined part of the scenario) is still rough due to its economical working mode that is to be specified by each Standard (or precision). The attributes A_{32} *and* A_{13} *are 'Color' and 'Direction', respectively. The values* Fw_{21} *and* Lw_{21} *stand for 'forward' and 'leftward' viewed from Robby as designated at the Standard, respectively.*

[Tom's Intention_1, ***Int₁***]

$\boldsymbol{Int_1}\leftrightarrow\underline{L(Robby,Robby,O_{11},O_{13},A_{12},G_{t},\_)\Pi}$
$\underline{L(Robby,O_{11},Robby,Robby,A_{12},G_{t},\_)\Pi}$
$\underline{(L(\_,O_{14},Tom,O_{11},A_{12},G_{s},\_)\bullet L(\_,O_{14},O_{11},Robby,A_{12},G_{s},\_))\Pi}$
$\underline{L(\_,O_{14},D_{11},D_{11},A_{13},G_{s},\_)}\Pi L(Robby,Robby,V_{11},V_{11},A_{16},G_{t},\_)\Pi L$
$(\_,O_{15},Robby,O_{12},A_{12},G_{s},\_)\Pi L(Robby,O_{15},Dis,Dis,A_{44},G_{t},\_)$
$\wedge box(O_{11})\wedge pillar(O_{12})\wedge table(O_{13})\wedge ISR(O_{14})\wedge ISR(O_{15})$

This formula implies that Tom wants Robby to carry the box between them to the table at a certain 'Velocity(A_{16})*',* V_{11} *without touching the pillar on the way, where '*O_{11}*' and '*O_{13}*' as the values of* A_{12} *represent their locations at each time point, and '*D_{11}*' is the direction to the box and Robby viewed from Tom.*

Tom is conscious that every attribute value to specify Robby's action is essentially vague but he believes that it should be imitated within certain tolerance associated with each Standard. The values **Dis** *and* **Meet** *stand for 'disjoint' and 'meet (or touch)' in Topology*(A_{44})*, respectively.*

<SESSION_1>

[Tom's Suggestion_1, ***T₁*** and Demonstration_1, ***D₁***]

$\boldsymbol{Int_1\Rightarrow T_1, D_1}$
$\boldsymbol{T_1}\leftrightarrow$ "Go to the table with the box between us like this."
$\boldsymbol{D_1}\leftrightarrow$ Figure 12

Tom decides to verbalize only the underlined part of Intention_1, ***Int₁*** *saliently with the belief that the rest can be included in his demonstration. Tom converts (or translates)* ***Int₁*** *into* ***T₁*** *and* ***D₁****.*

[Robby's Semantic_Understanding_1, ***S₁***]

$\boldsymbol{T_1, K_L\Rightarrow S_1}$
$\boldsymbol{S_1}\leftrightarrow(\exists\, x_1,x_2,x,y,z,p)L(x_2,x_2,y,x,A_{12},G_{t},\_)\Pi$
$L(x_2,y,x_2,x_2,A_{12},G_{t},\_)\Pi\ (L(\_,z,x_2,y,A_{12},G_{s},\_)\bullet$
$L(\_,z,y,x_1,A_{12},G_{s},\_))\Pi L(\_,z,p,p,A_{13},G_{s},\_)$
$\wedge x_2\neq x\wedge x_2\neq y\wedge box(y)\wedge table(x)\wedge ISR(z)$
$\wedge person\_1(x_1)\wedge person\_2(x_2)$

Robby interprets ***T₁*** *into* ***S₁****. The variable* 'x' *or* 'y' *is not yet anchored to the 'real table' or the 'real box' in the real environment because Robby has not perceived them yet. The predicates* 'person_1' *and* 'person_2' *refer to the first person (I) and the second person (You) and are to be pragmatically understood as 'Tom' and 'Robby', respectively.*

[Robby's Pragmatic_Understanding_1, ***P₁*** and Default_1, ***Def₁***]

$\boldsymbol{D_1\Rightarrow Per_1}$
$\boldsymbol{S_1, Per_1, K_D\Rightarrow P_1, Def_1}$
$\boldsymbol{P_1}\leftrightarrow L(Robby,Robby,O_{24},O_{25},A_{12},G_{t},\_)\Pi$
$L(Robby,O_{24},Robby,Robby,A_{12},G_{t},\_)\Pi$
$(L(\_,O_{26},Robby,O_{25},A_{12},G_{s},\_)\bullet L(\_,O_{26},O_{25},Tom,A_{12},G_{s},\_))\Pi$
$L(\_,O_{26},Lw_{21},Lw_{21},A_{13},G_{s},\_)\wedge box(O_{24})\wedge table(O_{25})\wedge ISR(O_{26})$
$\boldsymbol{Def_1}\leftrightarrow L(Robby,Robby,1m/sec,1m/sec,A_{16},G_{t},\_)\wedge\ldots$

The 'Location (A_{12})*' is attended to according to* ***S₁****.* ***Per₁*** *makes Robby aware that the words 'box' and 'table' should be anchored to the 'green object* O_{24}*' and the 'yellow object* O_{25}*' behind the pillar in the real environment, respectively. Robby conceives that he should approach to the table at his certain Standard.* ***Def₁*** *is inferred from* ***Per₁*** *and* ***K_D*** *as the default specification for the attributes not explicit in* ***T₁****.*

[Robby's Imitation_1, ***I₁***]

$\boldsymbol{P_1, Def_1, K_D\Rightarrow I_1}$
$\boldsymbol{I_1}\leftrightarrow$Figure12

Robby imitates ***D₁*** *according to* ***P₁****,* ***Def₁*** *and* ***K_D****.*

----- Resetting the situation to the initial situation ***Sit₀*****-----**

<SESSION_2>

[Tom's Suggestion_2, ***T₂*** and Demonstration_2, ***D₂***]

$\boldsymbol{I_1\Rightarrow PI_1}$
$\boldsymbol{Int_1, \sim PI_1\Rightarrow Int_2}$
$\boldsymbol{Int_2\Rightarrow T_2, D_2}$
$\boldsymbol{T_2}\leftrightarrow$"Don't touch the pillar."
$\boldsymbol{D_2}\leftrightarrow$ Figure 12

Tom perceives ***I₁*** *as* ***PI₁****. He denies* ***PI₁*** *and creates* ***Int₂*** *followed by* ***T₂*** and ***D₂****.*

[Robby’s Semantic_Understanding_2, $\boldsymbol{S_2}$]

$\boldsymbol{T_2}, \boldsymbol{K_L} \Rightarrow \boldsymbol{S_2}$

$\boldsymbol{S_2} \leftrightarrow (\exists x)L(\_,y,Robby,O_{21},A_{12},G_s,\_)$
$\Pi{\sim}L(Robby,x,Dis,Meet,A_{44},G_t,\_)\wedge ISR(x)\wedge pillar(O_{21})$

Robby gets aware that his imitation has been denied at the change of attribute ‘Topology (A$_{44}$)*’ from* ‘**Dis**joint*’ to* ‘**Meet***’.*

[Robby’s Pragmatic_Understanding_2, $\boldsymbol{P_2}$ and Default_2, $\boldsymbol{Def_2}$]

$\boldsymbol{D_2} \Rightarrow \boldsymbol{Per_2}$

$\boldsymbol{S_2}, \boldsymbol{Per_2}, \boldsymbol{K_D} \Rightarrow \boldsymbol{P_2}, \boldsymbol{Def_2}$

$\boldsymbol{P_2} \leftrightarrow \boldsymbol{P_1}\wedge \underline{L(\_,O_{27},Robby,O_{21},A_{12},G_s,\_)}\Pi$
$\underline{L(Robby,O_{27},Dis,Dis,A_{44},G_t,\_)\wedge pillar(O_{21})\ \wedge ISR(O_{27})}$

$\boldsymbol{Def_2} \leftrightarrow L(Robby,Robby,\ 1m/sec,\ 1m/sec,A_{16},G_t,\_)\wedge\ldots$

According to $\boldsymbol{S_2}$*, the ‘Location* (A$_{12}$)*’ of Robby and the pillar and their ‘Topology* (A$_{44}$)*’ are especially attended to, and the underlined part is conceived in addition to* $\boldsymbol{P_1}$*. No special attention is paid to the other attributes unmentioned yet.*

[Robby’s Imitation_2, $\boldsymbol{I_2}$]

$\boldsymbol{P_2}, \boldsymbol{Def_2}, \boldsymbol{K_D} \Rightarrow \boldsymbol{I_2}$

$\boldsymbol{I_2} \leftrightarrow$ Figure 12

-----Resetting the situation to the initial situation $\boldsymbol{Sit_0}$-----

<SESSION_3>

[Tom’s Suggestion_3, $\boldsymbol{T_3}$ and Demonstration_3, $\boldsymbol{D_3}$]

$\boldsymbol{I_2} \Rightarrow \boldsymbol{PI_2}$

$\boldsymbol{Int_2}, \boldsymbol{{\sim}PI_2} \Rightarrow \boldsymbol{Int_3}\ (\leftrightarrow \boldsymbol{Null})$

$\boldsymbol{Int_3} \Rightarrow \boldsymbol{T_3}, \boldsymbol{D_3}$

$\boldsymbol{T_3} \leftrightarrow$“Alright.”

$\boldsymbol{D_3} \leftrightarrow \boldsymbol{Null}$

Tom fails to deny $\boldsymbol{PI_2}$ *and comes to have no other intention (*$\boldsymbol{Int_3} \leftrightarrow \boldsymbol{Null}$*). That is, Tom is satisfied by* $\boldsymbol{I_2}$ *and only tells Robby* “Alright.”

[Robby’s Semantic_Understanding_3, $\boldsymbol{S_3}$]

$\boldsymbol{T_3}, \boldsymbol{K_L} \Rightarrow \boldsymbol{S_3}$

$\boldsymbol{S_3} \leftrightarrow (\exists x,y,k)L(x,y,1,1,B_{01},G_t,k)\wedge person(x)$

Tom gets aware that something ‘y’ *has evaluated by some person* ‘x’ *as perfect* ‘1’ *at ‘Worth (*B$_{01}$*)’ with a certain Standard* ‘k’.

[Robby’s Pragmatic_Understanding_3, $\boldsymbol{P_3}$ and Default_3, $\boldsymbol{Def_3}$]

$\boldsymbol{S_3}, \boldsymbol{Per_3}, \boldsymbol{K_D} \Rightarrow \boldsymbol{P_3}, \boldsymbol{Def_3}$

$\boldsymbol{P_3} \leftrightarrow L(Tom,\boldsymbol{I_2},1,1,B_{01},G_t,Tom)\wedge person(Tom)$

$\boldsymbol{Def_3} \leftrightarrow L(Robby,\ \boldsymbol{I_3},/,/,A_{01},G_t,\_)$

Finally, Robby pragmatically conceives that Tom is satisfied by $\boldsymbol{I_2}$ *at Tom’s Standard and believes that the next imitation,* $\boldsymbol{I_3}$ *is not needed to take ‘Place of Existence* (A$_{01}$)*’.*

[Robby’s Imitation_3, $\boldsymbol{I_3}$]

$\boldsymbol{P_3}, \boldsymbol{Def_3}, \boldsymbol{K_D} \Rightarrow \boldsymbol{I_3}$

$\boldsymbol{I_3} \leftrightarrow$ Null

Finally, no more imitation is performed.

-----End of all the sessions-----

5 TOP-DOWN CONTROL BASED ON $\boldsymbol{L_{md}}$

5.1 Attention mechanism

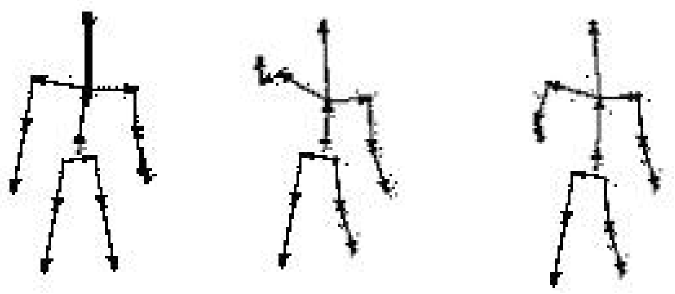
As mentioned above, the semantic understanding of human verbal suggestion makes a robot abstractly (i.e., conceptually) aware which matters and attributes involved in human demonstration should be attended to, and its pragmatic understanding provides the robot with concrete idea of real matters with real attribute values significant for imitation. More exactly, semantic understanding in $\boldsymbol{L_{md}}$ of human suggestion enables the robot to control its attention mechanism in such a top-down way that focuses the robot’s attention on the significant attributes of the significant matters involved in human demonstration. Successively, in order for pragmatic understanding in $\boldsymbol{L_{md}}$ of human suggestion, the robot is to select the appropriate sensors corresponding with the suggested attributes and make them run on the suggested matters so as to pattern after the movements of human FAO implied by the locus formulas yielded in semantic understanding. ***That is to say in short,*** $\boldsymbol{L_{md}}$ ***expression suggests a robot what and how should be attended to in human demonstration and its environment.***

For example, consider such a suggestion as S12 presented to a robot by a human. In this case, unless the robot is aware of the existence of a certain box between the stool and the desk, such semantic understanding of the underlined part as (26) and such a semantic definition of the word ‘box’ as (27) are very helpful for it. The attributes A$_{12}$ (Location), A$_{13}$ (Direction), A$_{32}$ (Color), A$_{11}$ (Shape) and the spatial event on A$_{12}$ in these $\boldsymbol{L_{md}}$ expressions indicate that the robot has only to activate its vision system in order to search for the box from the stool to the desk during the pragmatic understanding. That is, the robot can attempt to understand pragmatically the words of objects and events in an integrated top-down way.

(S12) Avoid <u>the green box between the stool and the desk</u>.

$$(\exists x_1,x_2,x_3,x_4,p)(L(\_,x_4,x_1,x_2,A_{12},G_s,\_)\bullet L((\_,x_4,x_2,x_3,A_{12},G_s,\_))\Pi$$
$$L(\_,x_4,p,p,A_{13},G_s,\_)\Pi L(\_,x_2,Green,Green,A_{32},G_t,\_)$$
$$\wedge stool(x_1)\wedge box(x_2)\wedge desk(x_3)\wedge ISR(x_4) \qquad (26)$$

$$(\lambda x)box(x)\leftrightarrow(\lambda x)L(\_,x,Hexahedron,Hexahedron,A_{11},G_t,\_)$$
$$\wedge container(x) \qquad (27)$$



(1) Data at t$_1$ **(2)** Data at t$_2$ **(3)** Data at t$_3$

Figure 13. Graphical interpretations of real motion data

Tom moved the right arm.
Tom raised the right arm.
Tom bent the right arm.
……………

(a) Text for motion data from t$_1$ to t$_2$.

……………
Tom lowered the right arm.
Tom stretched the right arm and simultaneously lowered the right arm.
……………

(b) Text for motion data from t$_2$ to t$_3$.

Figure 14. Texts generated from real motion data

This top-down control of attention mechanism enables IMAGES-M can take in real human motion data through the motion capturing system in SDPU. For example, Figure 13 shows graphical interpretations of the real motion data taken in at the time point t_1, t_2 and t_3. These real data were translated via $\boldsymbol{L_{md}}$ into such texts as shown in Figure 14 by TPU. In this case, IMAGES-M's attention was guided by the suggestion S13 below.

(S13) Move your right arm like this.

5.2 Utilization of domain-specific knowledge

The linguistic knowledge $\mathbf{K_L}$ is employed exclusively for semantic understanding, consisting of syntactic and semantic rules and dictionaries. On the other hand, the domain-specific knowledge $\mathbf{K_D}$ is employed for pragmatic understanding and behaviouralization, containing all kinds of knowledge pieces acquired so far concerning the robot, the human and their environment. For example, the human body can be described in a computable form using locus formulas. That is, the structure of the human body is one kind of spatial event where the body parts such as head, trunk, and limbs extend spatially and connect with each other. The expressions (28) and (29) are examples of these descriptions in $\boldsymbol{L_{md}}$, reading that an arm extends from a hand to a shoulder and that a wrist connects a hand and a forearm, respectively.

$$(\lambda x)\mathrm{arm}(x)\leftrightarrow(\lambda x)(\exists y_1,y_2)L(\_,x,y_1,y_2,A_{12},G_s,\_)$$
$$\wedge\mathrm{shoulder}(y_1)\wedge\mathrm{hand}(y_2) \qquad (28)$$
$$(\lambda x)\mathrm{wrist}(x)\leftrightarrow(\lambda x)(\exists y_1,y_2,y_3,y_4)(L(\_,y_1,y_2,x,A_{12},G_s,\_)\bullet$$
$$L(\_,y_1,x,y_3,A_{12},G_s,\_))\wedge\mathrm{body\text{-}part}(y_1)\wedge\mathrm{forearm}(y_2)$$
$$\wedge\mathrm{hand}(y_3) \qquad (29)$$

These descriptions are necessary for the robot to understand human action and text well enough to obtain an appropriate conception, eliminating such an anomalous one as is represented by S14 in a top-down way.

(S14) The left arm moved away from the left shoulder and
the left hand.

Each of such human's/robot's motions (M_k) as 'walk' and 'bow' is given as an ordered set of its standardized characteristic snapshots (S_k) called 'Standard Motion' and defined by (30). In turn, a family (F_X) of S_ks is called 'Family of Standard Motions' and defined by (31), where the suffix 'X' refers to 'human (X=H)' or 'robot (X=R)'. The families F_H and F_R are contained in $\mathbf{K_D}$ and their members are employed for the default motions, namely, motions not specified in human suggestion or demonstration, during pragmatic understanding.

$$S_k=\{M_{kS}, \ldots, M_{kE}\} \qquad (30)$$
$$F_X=\{S_1, S_2, \ldots, M_N\} \qquad (31)$$

For example, the $\boldsymbol{L_{md}}$ expression of human walking in default is given by (32), reading that a human moves by his/her legs making his/her shape change monotonically from Walk_S to Walk_E.

$$(\exists x,y,p_1,p_2,q_1,q_2)\ L(\_,y,x,x,A_{01},G_t,\_)\Pi$$
$$L(y,x,q_1,q_2,A_{12},G_t,\_)\Pi\ L(x,x,\mathrm{Walk}_S,\mathrm{Walk}_E,A_{11},G_t,F_H)$$
$$\wedge q_1\neq q_2\wedge\mathrm{human}(x)\wedge\mathrm{legs}(y) \qquad (32)$$

For another example, the $\boldsymbol{L_{md}}$ expression (33) is for the robotic motion of head shaking in default, reading that a robot affects its head in the Orientation (A_{14}), making its shape change monotonically from $\mathrm{Shake\_head}_S$ to $\mathrm{Shake\_head}_E$. The shape values are given in a computable form general enough to reconstruct any human/robot motion in 3D graphics or so. Figure 15 shows an example of its interpretation in 3D graphics by PPU in IMAGES-M, which is also an example of cross-media translation from the text 'The robot shakes its head' into the animation.

$$(\exists x,y,p_1,p_2)L(\_,y,x,x,A_{01},G_t,\_)\Pi L(x,y,p_1,p_2,A_{14},G_t,\_)\Pi$$
$$L(x,x,\mathrm{Shake\_head}_S,\mathrm{Shake\_head}_E,A_{11},G_t,F_R)$$
$$\wedge\mathrm{robot}(x)\wedge\mathrm{head}(y) \qquad (33)$$

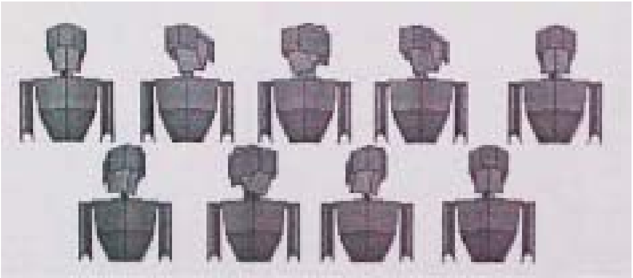


Figure 15. 3D animation of 'The robot shakes its head.'

5.3 Behaviouralization

The process for behaviouralization is to translate a conception (i.e., $\mathbf{C_i}$) into an imitation (i.e., $\mathbf{I_i}$) as a appropriate sequence of control codes for certain sensors or actuators in the robot to be decoded into a real behaviour by SDPU or ADPU in IMAGES-M. For this purpose, there are needed two kinds of core procedures so called 'Locus formula paraphrasing' and 'Behaviour chain alignment' as detailed below.

5.3.1 Locus formula paraphrasing

The attributes listed in Table 1 are essentially for human sensors or actuators and therefore the locus formula as $\mathbf{C_i}$ should be translated into its equivalent concerning the attributes specific to the robot's. For example, an atomic locus of the robot's 'Shape (A_{11})' specified by the human should be paraphrased into a set of atomic loci of the 'Angularity (A_{45})' of each joint in the robot. For another example, 'Velocity (A_{16})' for the human into a set of change rates in 'Angularity (A_{45})' over 'Duration (A_{35})' (i.e., A_{45}/A_{35}) of the robot's joints involved. These knowledge pieces are called 'Attribute Paraphrasing Rules (APRs)' [10] and contained in $\mathbf{K_D}$.

5.3.2 Behaviour chain alignment

Ideally, the atomic loci in the conception $\mathbf{C_i}$ (original or paraphrased) should be realized as the imitation $\mathbf{I_i}$ in a perfect correspondence with an appropriate chain of sensor or actuator deployments. Actually, however, such a chain as a direct translation of $\mathbf{C_i}$ must often be aligned to be feasible for the robot due to the situational, structural or functional differences between the human and the robot. For example of situational difference, in the simulation above, the robot must interpolate the travel from its initial location to the green box and the action to pick up the box. On the other hand, for example of structural or functional difference, consider the case of imitation by a non-humanoid robot. Figure 16 shows the action by a dog-shaped robot (SONY) to the suggestion 'Walk and wave your left hand.' The robot pragmatically understood the suggestion as '*I* walk and wave *my* left *foreleg*' based on the knowledge piece that only forelegs can be waved' and behaviouralized its conception as 'I walk *BEFORE* sitting down *BEFORE* waving my left foreleg' but not as 'I walk, *SIMULTANEOUSLY* waving my left foreleg', in order not to fall down.

The procedure here [6, 12] is based on the conventional AI, where a problem is defined as the difference or gap between a 'Current State' and a 'Goal State' and a task as its cancellation. Here, the term 'Event' is preferred to the term 'State' and 'State' is defined as static 'Event' which corresponds to a level locus. On this line, the robot needs to interpolate some transit event X_T between the two events, 'Current Event (X_C)' and 'Goal Event (X_G)' as (34).

$$X_C \bullet X_T \bullet X_G \qquad (34)$$

According to this formalization, a problem X_P can be defined as $X_T \bullet X_G$ and a task can be defined as its realization and any problem is to be detected by the unit of atomic locus. For example, employing such a postulate as (35) implying 'Continuity in attribute values', the event X in (36) is to be inferred as (37).

$$L(x,y,p_1,p_2,a,g,k) \bullet L(z,y,p_3,p_4,a,g,k) .\supset. p_3=p_2 \qquad (35)$$

$$L(x,y,q_1,q_2,a,g,k) \bullet X \bullet L(z,y,q_3,q_4,a,g,k) \qquad (36)$$

$$L(z',y,q_2,q_3,a,g,k) \qquad (37)$$

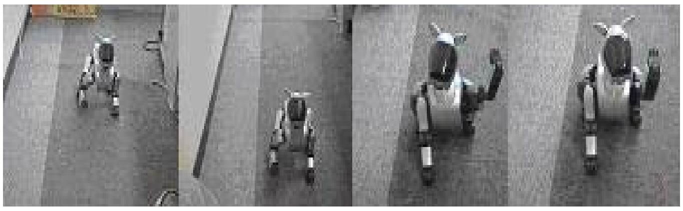


Figure 16. Robot's action to 'Walk and wave your left hand'

6 DISCUSSION AND CONCLUSION

The key contribution of this paper is the proposal of a novel idea of robotic imitation driven by semantic representation of human suggestion, where are hinted in the formal language L_{md} what and how should be attended to in human action as analogy of human FAO movement and thereby the robotic attention can be controlled in a top-down way. Without such a control, a robot is to simultaneously attend to tens of attributes of every matter involved in human action as shown in Table 1. This is not realistic, considering the difficulties in autonomous robotic vision understanding today. The author has a good perspective for the proposed theory of robotic imitation based on his previous work utilizing L_{md} for robot manipulation by text [6, 12]. This is one kind of cross-media operation via intermediate L_{md} representation [e.g., 6, 10, 12]. At my best knowledge, there is no other theory or system that can perform cross-media operations in such a seamless way as ours. This is due to the descriptive power of L_{md} enabling systematic organization and computation of spatiotemporal knowledge including sensation and action. Our future work will include establishment of learning facilities for automatic acquisition of word concepts from sensory data and multimodal interaction between humans and robots under real environments in order to realize the robotic imitation proposed here.

ACKNOWLEDGEMENTS

This work was partially funded by the Grants from Computer Science Laboratory, Fukuoka Institute of Technology and Ministry of Education, Culture, Sports, Science and Technology, Japanese Government, numbered 14580436 and 17500132.

REFERENCES

[1] A.Billard, 'Learning motor skills by imitation: biologically inspired robotic model', ***Cybernetics and Systems***, **32**, 155–193, (2000).

[2] A.Alissandrakis, C.L.Nehaniv, and K.Dautenhahn, 'Imitating with ALICE: Learning to imitate corresponding actions across dissimilar embodiments', ***IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans***, **32-4**, 482–496, (2003).

[3] J.Nakanishi, J.Morimoto, G.Endo, G.Cheng, S.Schaal, and M.Kawato, 'Learning from demonstration and adaptation of biped locomotion', ***Robotics and Autonomous Systems***, **47**(2-3), 79–81, (2004).

[4] J.M.Wolfe, 'Visual search in continuous, naturalistic stimuli', ***Vision Research***, **34**, 1187–1195, (1994).

[5] Y.Demiris and B.Khadhouri, 'Hierarchical attentive multiple models for execution and recognition of actions', ***Robotics and Autonomous Systems***, 54, 361–369, (2006).

[6] M.Yokota, 'Towards a universal language for distributed iIntelligent robot networking', ***Proc. of 2006 IEEE International Conference on Systems, Man and Cybernetics***, Taipei, Taiwan, (Oct., 2006).

[7] S.Coradeschi and A.Saffiotti, 'An introduction to the anchoring problem', ***Robotics and Autonomous Systems***, **43**, 85–96, (2003).

[8] E.Drumwright, V.Ng-Thow-Hing, and M.J.Mataric´, 'Toward a vocabulary of primitive task programs for humanoid robots', ***Proc. of International Conference on Development and Learning*** (*ICDL*), Bloomington,IN, (May, 2006).

[9] M.Yokota, 'An approach to integrated spatial language understanding based on Mental Image Directed Semantic Theory', ***Proc. of 5th Workshop on Language and Space***, Bremen, Germany, (Oct., 2005).

[10] M.Yokota and G.Capi, 'Cross-media operations between text and picture based on Mental Image Directed Semantic theory', ***WSEAS Trans. on INFORMATION SCIENCE and APPLICATIONS***, Issue 10, 2, 1541–1550, (2005).

[11] J.F. Sowa, ***Knowledge Representation: Logical, Philosophical, and Computational Foundations***, Brooks Cole Publishing Co., Pacific Grove, CA, (2000).

[12] M.Yokota, 'Integrated multimedia understanding for ubiquitous intelligence based on Mental Image Directed Semantic Theory', ***Handbook on Mobile and Ubiquitous Computing Innovations and Perspectives***, American Scientific Publishers, (in press).

[13] M.Egenhofer, 'Point-set topological spatial relations. Geographical Information Systems', 5,2 161-174 (1991).

[14] M.Nicolescu, M.J.Mataric´, 'Task Learning Through Imitation and Human-Robot Interaction, in Models and Mechanisms of Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions, Kerstin Dautenhahn and Chrystopher Nehaniv Eds., 407-424, (2006).

Imitation in animals in lack of causal understanding?

Zsófia Virányi
Konrad Lorenz Institute for Evolution & Cognition Research, Altenberg, Austria

Various experimental results have shown that when causal information is available about a problem and a demonstrator's method to solve it chimpanzees prefer emulation and try to find their own (more efficient) method when presented with the same problem. At the same time it has also been suggested that they switch to imitation when causal structure of the problem and its demonstrated solution is unclear. It is questionable, however, whether more precise copying of the demonstrated actions in lack of knowledge about their relevance can be considered as imitation, or rather reflects emulation in animals which expect others' behaviour be efficient in lack of contradictory information.

In the present literature on human and non-human imitation two phenomena are described as unclear causal structure without clear differentiation between them: 1) in case of the above mentioned lack of full information of the observed action and its constraints efficiency of the action cannot be evaluated but can be assumed; 2) full information is available about the physical constraints and effects of the observed action, but they are in contradiction with the action itself (choice of the action cannot be explained by them).

Purpose of the poster is to draw attention to the need of differentiating between these two kinds of lack of causal understanding of social learning situations in order to avoid possible false comparisons between species and to make viable theoretical interpretations.

Selective imitation in dogs

F Range+, Zs Viranyi* §, L Huber+
+ Department for Behaviour, Neurobiology and Cognition, University of Vienna, Austria
§ Department of Ethology, Eötvös University, Budapest, Hungary
* Konrad Lorenz Institute for Evolution & Cognition Research, Altenberg, Austria

The transmission of cultural knowledge requires learners to identify what relevant information to retain and selectively imitate when observing other's skills. By one year of age human infants - without relying on language or theory of mind – already show evidence of this ability. They are able to interpret others' behavior as goal-directed, and as a result predict the most efficient action to achieve a goal within the constraints of a given situation. One situation in which human infants are thought to manifest this non-mentalistic inferential process is their selective imitation of goal-directed actions. For example, if a model demonstrates a head action when a hand action would be more efficient to turn on a light, infants imitate the head action only if its use during the demonstration cannot be explained by their hands being occupied, suggesting imitation by preverbal infants to be a selective, interpretative process (Gergely, Bekkering, Kiraly, 2002). However, the less effective action is only copied if the demonstration is accompanied by communicative cues targeted at the infants. Thus, early sensitivity to ostensive-communicative cues and the efficiency of goal-directed actions seem to be crucial prerequisites for such relevance-guided selective imitation (Csibra & Gergely, 2006). While this competence has been thought to be human-specific, here we show an analogue capacity in a non-human species, the domestic dog (Canis familiaris). In our experimental set-up, subjects watched a demonstrator dog pulling a wooden rod using an 'ineffective' paw action instead of using a mouth action usually preferred by dogs as was shown in a control group. In one group, using the 'ineffective' action was justified by the constraints of the situation e.g. the mouth of the model dog was occupied with a ball, whereas in the second group no constraints were present to explain the demonstrator's choice. In the first trial after observing the trained conspecific model, dogs imitated the non-preferred action only in the group where no constraints were present that could have explained the model's paw use. Consequently, dogs did not blindly copy the ineffective method, but demonstrated relevance-guided selective imitation like the infants in a comparable task.

Robotic Locust: who is my friend?

Shigang Yue
Brain Mapping Unit, Downing Site

To make a robot interact with human effectively, one important thing is to make sure it is able to recognise friendly and aggressive behaviours against it.

In this movie, we showed that it is possible for a robot to recognise these two different things around it. We equipted a khepera II robot with a pair of locust's inspired visual neural systems to see its surroundings, and a motor system to interpret the outputs of the visual system into behaviours.

The visual systems were based on lobula giant movement detector (LGMD) and decending contralateral movement detector (DCMD) in locusts. The visual-motor control was based on a motor system which may control locusts' directional jumping behaviours.

As shown in the movie, the robotic locust can recognise movements towards it by comparing the spikes from its two 'eyes'- escaping if it was an aggressive one, or just sitting there if it was a slow and gentle one, like a friend's movements.

The robotic locust always be able to run away from the fast approaching objects, which is often predators, regardless these objects' color, shape and materials.

We hope this move brings new inspiration ...

Object Affordances: Linking Sensory Motor Maps and Imitation

L. Montesano, M. Lopes, A. Bernardino, J. Santos-VIctor

The concept of affordance was introduced by Gibson as relation between an agent and the environment based on the agent's action capabilities. In this paper we argue that this concept (or knowledge representation) plays an important role as a bridge between sensory motor maps and higher cognitive capabilities such as imitation. Affordances encode relationships between actions, objects and effects and are at the core of basic cognitive capabilities such as prediction and planning. Within the framework of a general developmental architecture for social robots, we address the problem of learning affordances through the interaction of a robot with the environment as a key step to understand the world properties and interact socially. We present a general model for affordances using Bayesian networks. Actions, object features and effects form the nodes of the network and the affordances are implicitly encoded by the dependencies between these nodes. The amount of prior knowledge and the selected variables define different learning scenarios ranging from parameter tuning, which is the most common problem in the literature, to more general instances that also cope with feature selection and multiple actions. Since learning is based on a probabilistic model, it is able to deal with uncertainty, redundancy and irrelevant information present in real world. In addition to this, the model allows to directly use the acquired affordances to solve prediction, recognition and planning tasks. Using the affordances, the robot is able to imitate a human based on the perceived effects and its knowledge about its own action capabilities. We demonstrate successful affordance learning on a humanoid robot interacting with objects and apply the acquired knowledge in simple imitation games.

Mindful Environments

One of the key paradigms for interaction envisaged for Ambient Environments does not just involve disappearing computers, but also a disappearing interface. "Natural interaction", or an intelligent system that can determine at any time what the inhabitants of the environment need and long for, constitutes the holy grail of ambient interaction. In this sense, the environment should be able to make conjectures of the mental state of users as accurately as possible - similar to the way we can read the minds of others. Most computational research to date on detecting the mental state of people have failed to consider the full range of mental states that people display in natural interactions and the full range of displays of the various mental states. They have not been able to capture how humans communicate their intentions, the intricacies of mental life and have often ignored ecological validity.

This workshop addresses the question how to go beyond the rather simplistic notions regarding natural interaction in mindful environments. In order to be able to build such systems, we need to integrate the knowledge we have about how people show what's on their mind and how people go about building theories of what goes on in the minds of others. One of the aims of this workshop is to bring together an interdisciplinary group of researchers to discuss the state of the art of the research on the study of theory of mind (in particular in human communication) and on computational modelling and system building that is directed towards the ability to recognize and represent the intentions and other aspects of the mental state of a person interacting with others and with computational systems in an (ambient) environment. Another aim of this workshop is to discuss how the computational models could inform empirical and theoretical research in human social processes, through formalization and simulation, for instance.

Some of the kinds of studies of interest include:

- Studies of behaviours and the models of behaviour that people display in interacting with each other and the environment. How can we really tell what goes on into another person's mind? What cues do people use and how can we rely on them? How can the features be detected? What is needed to interpret them?
- Studies into cognitive modelling: alternative theories have been proposed for how people come to understand beliefs, desires and intentions of others, a theory of mind. How can we model these theories? How do current computational models of theory of mind compare to these theories and how do we evaluate them? How can computational models and simulations inform knowledge about human processing and vice versa?
- Studies in system development for the intelligent environment such as robots and virtual humans. What should a cognitive model of an intelligent interactive environment look like? What should a representation of the mind look like? Which categories need to be represented (intentions, beliefs, attitudes, emotions, action tendencies)?

To help answer questions like these related to behaviours and modules, on modelling and simulation-based studies of communication and cognition, and on system building, we have received contributions of a variety of disciplines. From researchers studying natural systems, such as humans, that are equipped with mind-reading skills to system engineers involved in building computational systems; from linguistic, psychology, sociology, computational modelling (simulation, (multi-)agent systems) and signal processing.

Dirk Heylen & Stacy Marsella (Symposium Chairs)

Attribution of Communicative Capacity Among Agents in a Hetereogeneous Population

Melanie Baljko[1] and **Nell Tenhaaf**[2]

Abstract. In this paper, we describe our work developing *A-life sculpture*: art works that are interactive and that incorporate behaviours that are not explicitly explained to the human interactants. In interacting with A-life sculptures, human interactants *discover* the system's behaviours, a process that results in a co-construction of the artwork (and of the art work's meaning), in the sense that experience of the work is different for each participant, and many facets of the work are not immediately available, but appear during the time spent with it. During the interaction, the sculpture is also responding and reacting, in a meaningful way, to the interactant. Our work incorporates both artistic and scientific goals, and we have integrated these two perspectives during the development process, which has entailed broadening our design framework beyond traditional performance-based approaches.

1 BACKGROUND

Theories of human–human interaction tell us that human interactants construct defeasible assertions (or hypotheses) about the mental states of others when designing their actions (communicative or other). A notable example of this is Clark's Grounding Theory [4], which posits the existence of "common ground" (i.e., assertions that not only hold in the minds of each interactant, but that are also thought by each to also be held by the other interactant(s)); goals are advanced through the accumulation of common ground. Numerous empirical studies, especially those that involve task-oriented communicative exchanges (such as assembling origami shapes or navigation), have provided compelling evidence in support of this theory.

Principles of biomimicry tell us that one approach to system design would be to transplant Grounding Theory (or others like it) to human–computer interaction. According to this design approach, we should build computational systems that make conjectures of the mental states of users and that attempt to "read the minds" of others, we should develop agent architecture mechanisms whereby currently-held beliefs about the mental states of others can be revised, in view of observations accumulated in real-time (i.e., afford agents the ability to incorporate information about how the human interactants show "what's on their mind"), and we should develop techniques not only to recognize the evidence of the mental states of others, but also to represent those mental states (second-order representation). In doing this, it will be important to consider the full range of mental states (e.g., is the BDI model adequate? are more or fewer types of mental states required?); to capture the intricacies of mental life; and to ensure ecological validity in the mental landscape that is devised for computational agents. Such a direction would advance the goal of "natural" interaction because it espouses the principles of interaction that humans use.

These principles, which are seemingly innate, actually are complex and need to be acquired by language learners, as evidence from communication disorders tells us [10]. Indeed, much work has been done in this direction; both in the view of traditional, symbolic AI (e.g., the so-called "cognitive modeling" work; computational implementations of Grounding Theory, such as work by Traum [15]) and in the probabilistic framework (e.g., systems that make use of decision theory and Markov Decision Processes, such as [9], systems that make use of the partially-observable variants [16]). Perhaps an interesting point of discussion for the workshop participants will be whether the logical endpoint of this direction is tantamount to strong AI.

None of the previously-mentioned computational interactive systems, however, attempt to do away completely with the notion that the computational system needs to have a manifestation (via its interface). The tacit assumption is made that the human interactant needs to imbue the system, via an interface, with certain characteristics or properties in order to treat it as a co-interactant and in order to make presuppositions about the system's agency. If the system's interface were to disappear altogether, the presuppositions could be challenged, possibly in a detrimental way. At any rate, one reason the aforementioned "biomimicry" approach actually works is that humans apply social rules, even when interacting with computational media (and even when they *know* the media is computational — the so-called "Media Equation" [12]). Computational agents are anthropomorphized by human interactants. A research question that follows, then, is to what degree can the interface "disappear" (or the manifestation of the interface be subverted) and still elicit this strong response of anthropomorphization? We will revisit this question below, but first we will address a related matter.

Recently, HCI practioners have been looking to build a bridge between traditional performance-based HCI to the "new world" of HCI where other, harder-to-quantify factors like aesthetics, values, and emotional experience are more important (e.g., [6, 11, 14]). This "new world" of HCI is motivated by the observations that computational technologies have migrated into everyday lives; that the extant, performance-based frames of reference may no longer be adequate; and that new frames of references may be needed in order to understand and design human-computer interaction. It has been hypothesized that aesthetics may play a role in experience design and understanding user experience. This hypothesis resonates with our approach, and, as promising as it is, the field is grappling with foundational, conceptual issues, as well as methodological ones (how to

[1] Department of Computer Science and Engineering, Faculty of Science and Engineering, York University, email: baljko@cse.yorku.ca

[2] Department of Visual Arts, Faculty of Fine Arts, York University, email: tenhaaf@yorku.ca

design for an aesthetics of interaction? how to evaluate the aesthetics of interaction, whether qualitatively, quantitatively, or both?).

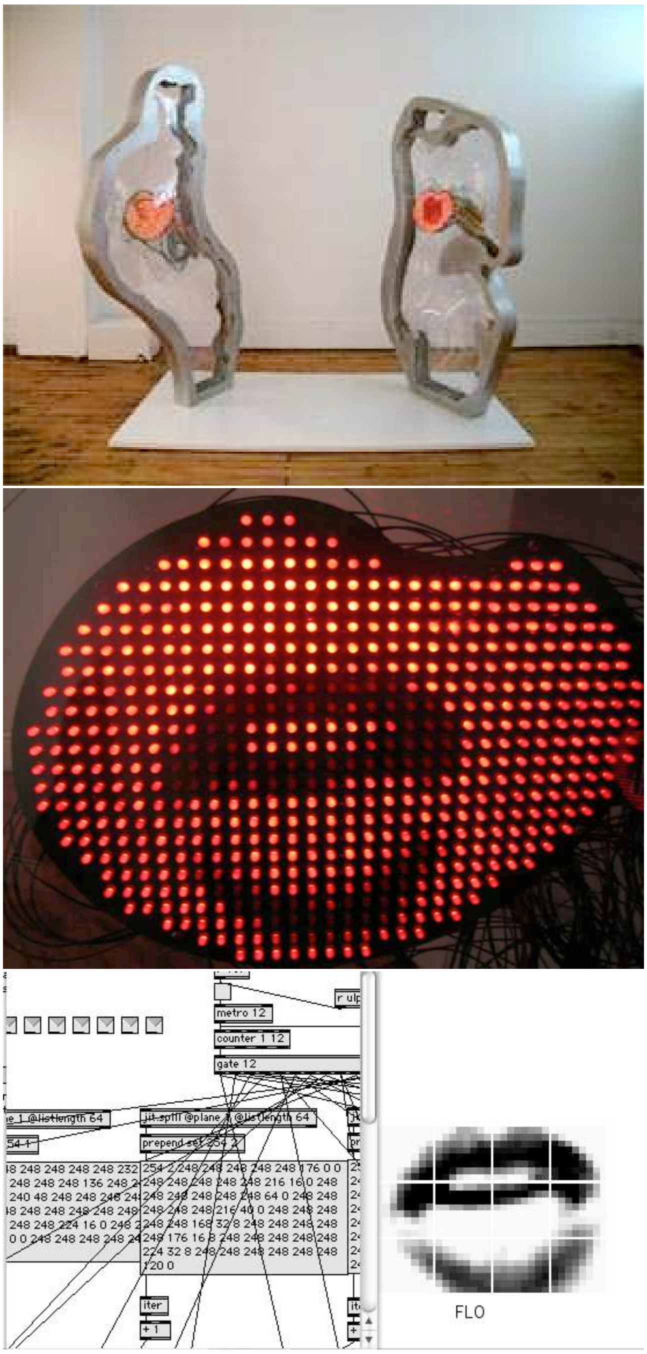
In our work, we too are investigating technqiues to combine the "old" with the "new" HCI. We are building dynamic art works that are based on multi-agent systems consisting of both artificial and human-representative agents. This heterogeneous community of agents is observable through sculptural, abstracted visual and auditory displays (A-life sculptures). Agents are embodied as composites of electronic components, such as clusters of LEDs and stereo speakers. Such embodiments provide the physical infrastructure for the mounting of sensors in a non-obvious manner (which provide the agent's *modes of sensory-perception*). We consider such embodiments — which we describe as *low-fidelity embodiments* — to be preferable to high-fidelity embodiments, such as humanoid-like, digitally-rendered characters, because they circumvent the cliches and expectations attached to humanoid characters, avatars, or (even worse) cartoons. They have another advantage: their high level of abstraction. Both a single agent and a population of such agents can be embodied by the same physical infrastructure. The human interactant distinguishes between the two cases on the basis of the behaviour of the articulators (whether the pixellated lights and audio displays cohere into perceivable sub-units). The "observer" is understood to refer to both the human interactant(s), as well as to passive viewers of the art works.

Figure 1 shows a detail from *Flo'nGlo* [Tenhaaf, 2005;2007], which is an artwork that makes use low-fidelity embodiment. The two characters Flo and Glo are much more cartoon-like than we plan for the representation of agents in our current work. In Flo'nGlo, low-fidelity embodiment is used primarily to induce the viewer to look more closely at the characters' "guts" so as to identify what he or she is seeing. As well, the pixellated boards serve to unify the video imagery of each character with the pixellated expression of an optimizing algorithm that periodically re-sets the conversational give-and-take between them (which involves butting in, conceding the turn, etc.) The two characters talk only to each other, not to the viewer; in this earlier work that precedes our collaboration, the sense of a conversation between the agents was achieved in a metaphorical way, reinforced by exclusion of the viewer.

In developing these art works, we are looking to aspects of A-life research in which the emphasis moves away from building and studying just the A-life artefact, and toward exploring the social environment in which such artefacts are deployed. This allows us to consider directly how biases and assumptions are built into every system design, and how agents in a system might tell a human interactant who is not the system builder what the premises of the system are, via an interface that is designed to guide her/him. In an artwork, such information tends to be not overt but integrated into interactions that are designed to be intuitive.

The art works we are currently building are interactive in that the manifest behaviour of the work depends on the behaviours and choices made by the human interactant(s) in a conversational exchange involving the community as a whole. The artistic process of interaction entails that the human interactant recognizes that he or she is a member of the agent population. That internal, mental state of the human interactant becomes system behaviour once the human takes action that is predicated on her/his mental state.

We refer to the sum of these interactions as system behaviour, which can be considered *emergent* from the perspective of an "observer" who is engaged in the process of engagement in the artistic process, who seeks to arrive at a collaborative solution to a conversational task.



```
254 2 248 248 248 248 248 176 0 0
248 248 248 248 248 216 16 0 248
248 248 248 248 248 64 0 248 248
248 248 248 216 40 0 248 248 248
248 248 168 32 8 248 248 248 248
248 176 16 8 248 248 248 248 248
224 32 8 248 248 248 248 248 248
120 0
```

Figure 1. Low-fidelity embodiment used in the artwork *Flo'nGlo* [Tenhaaf, 2005; 2007]. The top image shows the entire piece, the middle image shows a detail of LED clusters, and the bottom image shows the software simulation of the hardware instantiation (implemented using the Max/MSP programming environment).

1.1 Goals

This project has both artistic and scientific goals.

A primary artistic inspiration for this project stems from the desire to incorporate, in art works, complex forms of interactivity that call on deep involvement by viewers, using the modes and strategies of human conversation as our model. Achieving this goal will entail, in part, the cueing of behaviours in the human interactant. The conversation takes place not between the human interactant and the artwork, but among a population of agents that are represented to the human interactant *by* the artwork; the human participation is mediated by their *representative agent*. When a human interactant joins the conversation, their representative agent is instantiated into the population. The human interactants first need to figure out the mechanisms for recognizing themselves in the population, and then for taking action as a member of the community. Human interactants are colour-coded to differentiate them from virtual agents, and in parallel we have two kinds of sound identifiers. This is a behaviour-based system that we are building, with an emphasis on affording immediate actions that allow the overall behaviour — as well as the underlying motivational levels — of the system to become apparent. But the appearance as well as other sensory (output) modes of the system have to be "readable" by participants and observers. We have hypothesized that we can use these representational features to help guide the interactant's interpretation of system behaviour.

An immediate shared goal for the mixed society of human and virtual agents will be to perform a conversational exchange that achieves a pleasing sound progression, combining the two different types of sounds. "Pleasing" is a deliberately subjective word, as the task could be fulfilled in many different ways as long as there is a combined effort among all agents in the population (which may be as small as two). This still achieves an explicit "best behaviour" goal. Empirical studies by Sengers et al. [14] shows that the public approach an art installation in a gallery or museum (as opposed to an informational device about the art) ready for critical reflection and with open-endedness in their expectations. Indeed, it is more likely that an art viewer will engage with an interactive installation than an everyday user of technology (where such viewers are more willing to co-construct meaning from the experience). Yet many of the same HCI design challenges apply, because an art viewer first of all must be transformed into a participant, and secondly, must be guided in the co-construction of meaning in such a way that whatever they experience counts, no matter what feedback they get from the system.

The scientific inspiration of the project is to devise a set of mechanisms whereby the structure of human–human conversational turn-taking is emergent in the interactions between humans and artificial agents. A further scientific hypothesis of this work is that explicitly cueing the anthropomorphization of agents (e.g., such as which is acheived through the use of human- or animal-like characters) is not necessary in order to engage turn-taking behaviours in human interactants, and that human interactants will bestow sufficient turn-taking "agency" to the artificial entities, provided their behaviour is sufficiently nuanced, given whichever articulators are available to it. "Evaluation" of the art work in situ is tantamount to the (scientific) evaluation of this hypothesis. We are, in effect, not taking for granted the gesture of anthropomorphizing, but breaking it down into steps that are more interesting — both scientifically and artistically. In each of these domains, the term is used very loosely to attempt to account for engagement of humans with non-human agents, whereas it actually obscures processes of engagement and explains nothing. Our cueing steps explain a lot more: recognition by the interactant of their representative agent in the population is accompanied by attribution of agency to any similar agent representation; and, the subsequent co-construction of both the experience and the meaning of the artwork reveals attribution of communicative intent.

We feel that scientific evaluation is most appropriately applied to analyzing the elicited interactions themselves, that is, the richness of their structure. We will apply sets of criteria that characterize the pattern and the structure of conversational turn-taking that has taken place, and the degree to which the human interactant adapts his or her use of modes to the artificial agent. We plan to compare these attributes to those in face-to-face human-human interactions and to evaluate the differences, if any, among the interactions elicited in the various installation sites.

We also have the goal to derive an evaluation methodology that incorporates both functional and aesthetic criteria.

2 APPROACH

2.1 Embodiment

In order to test our hypotheses that turn-taking behaviours will be elicited in humans and that agency will be bestowed on virtual agents even without relying on overt anthropomorphization, we have elected to endow the artificial agents with *low-fidelity* embodiments.

Our underlying premise is that for an interactant to meaningfully experience communication and conversation with a machine or a digital system, the system should have characteristics that are unique to it. Strong emotions in humans are elicited by technology itself, by systems that present themselves as the machines that they are, rather than as simulated humans (see [1]). Note that a conversational exchange among human and artificial agents can call on biomimicry even without the virtual agents themselves mimicking humans. This is why we are moving completely away from "characters" and toward image and sound modes that are abstract and essentially machinic: the system reveals its workings through low-resolution video, mixed in with a display of the system's calculations, and through sound that it is based on pure signal and leans toward a noise aesthetic. The interactivity is not predictable and is based on inputs that are processed in real-time, which we hypothesize will add to a sense of the system's autonomy. It works like an evolving feedback loop, in which artificial agents adapt on their side of the interface as we do on ours, in a relational, interactive process.

Our system is built on the principles of Embodied Artificial Life (EAL), because "cognitive-oriented concepts" such as communication and intent are grounded in context, not in abstraction. They rely on "the dynamics inside an agent and its coupling to the environment" [5, p. 603].

Our system, in its current prototype form, relies on the interactant's movements as the basis for the system's sensory input. Tracked by an overhead camera, the interactant moves from side to side or forward and backward in space, taking her/his representative agent along. The prototype presently makes use of a screen-based simulation, but our next design iteration will make use of small clusters of LEDs. Turn-taking will thus be negotiated by body movement on the part of the interactant, a feature that we feel is significant for intuitive and emotional engagement with the art work. Attribution and reading of "mental state" is expanded well beyond cognitive processes alone, although we believe that intellectual consideration of concepts and social topics is a pleasurable aspect of aesthetics for most viewers of contemporary art.

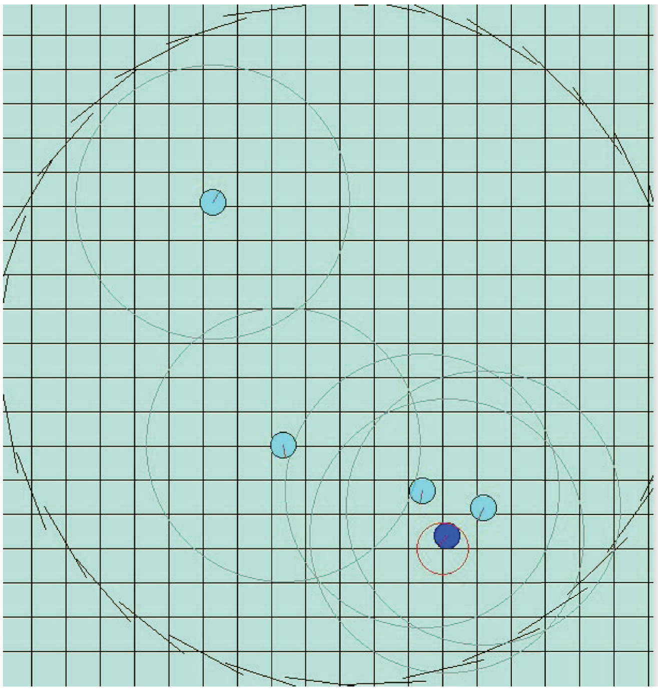


Figure 2. An illustration of the herding task. The nest is shown as a red circle, the target is shown as a blue circle and the agents in the population are shown as cyan circles.

2.2 Task Domain

We have constructed the art works such that they present to the interactants both a *navigation collaborative* task (implemented) and a *conversational* task (in progress). A herding task has been implemented. A software prototype has been developed in which the agents are represented simply as two-dimensional circles. In this task, agents in the population must direct the target entity into what we term *the nest*. The behaviour of the target entity is implemented by the simple rule to evade the agents. The behaviour of the artificial agents is governed by a set of simple rules that express attraction and repulsions to other entities in the arena (e.g., attraction to the target and the nest, repulsion from close proximity to other agents and the arena walls). A single agent cannot perform this task in isolation (except in special cases), since the target robot simply evades the agent. However, when multiple agents are placed in the arena, they collectively are able to herd successfully the target entity to the nest. The behaviour of the human-representative agent is under the control of the interactant. This mechanism involves a phase during which the representative agent entrains itself to the interactant via tracking by overhead camera, and subsequent gesture mimesis (the agent moves when the interactant does, and the sound changes to reflect the movement). A schematic of the task, shown in the software prototype, is shown in figure 2.

The metaphor of human conversation has been applied to human-computer interaction previously, but, in these previous cases, the embodiments of the agents have been what we would describe as high-fidelity: either the agents themselves have been highly anthropomorphized (e.g., Rea and other embodied conversational agents [3]) or, if the human-like embodiments have been discarded, the agents themselves have retained human-like modes of articulation, most typically speech or text glosses of speech (such as the systems of [7], and others). Other multimodal interfaces (e.g., [8]) are not truly "conversational", since the turn-taking is pre-structured (conversation is characterized by the free exchange of turns among the interactants). The focus of the scientific investigation in this project is the means and mechanisms for eliciting turn-taking in interactions with low-fidelity agents, which is a focus that has not been previously undertaken. Artistically, we want to see how readily an interactant understands the task without any instruction at the interface, and whether the experience goes beyond either fun or frustration to provoke a reflection about the "lifelike" qualities of technology.

2.3 Agent Architecture

The interactive dynamic of the proposed artwork entails bestowing artificial agents with the autonomous abilities to monitor the state of a conversational interaction, to take the conversational turn and hold it. The architectures must also cue the desired behaviours in the interactant.

The artificial or virtual agents in the system are built to take their cue from the human interactant's actions, and also to initiate and perform conversational tasks. The states and transitions of the system are dictated by one of several possible finite state machines (FSMs). We plan to solicit empirical evidence on the behaviour of human interactants with these various FSMs. The FSMs are based directly on the "systemics" of turn-taking described by Sacks et al. [13] (different versions exist for whether barge-ins are allowed or not, and for different ways of handling turn-taking "collisions"). The states of the FSMs represent the various conversational states (e.g., which interlocutor holds the turn, whether the turn is presently under negotiation). We are building an architecture for the virtual agents that affords to them knowledge of the current state and to implement a particular conversation strategy, given that state. Whenever a turn-relevant point is detected (i.e., a point at which the current turn-holder offers the turn, either by specifically targeting the next potential turn-holder or generally offering the turn to any interactant), the possible responses are to accept or decline the offer (collisions can occur if there are multiple accepts). Speakers can also attempt to barge in at any time, whether this strategy is successful or unsuccessful depends on the particular FSM that is instantiated.

How can the trait of intentionality and the ability to read cues be built into the virtual agents' architecture? The human agent has will or intentionality, along with her or his ability to read cues. Will or intention is currently not implementable (given the absence of strong AI), so virtual agents need an algorithmic logic of behaviour instead — a logic that is based on computing routines like sorting, optimizing, timing. Such a logic translates intentionality in its human sense into a parallel trait that is specific to computational entities, conferring the ability to act according to internally-set parameters in parallel with the ability to read cues from the environment and from human agents.

3 WORK TO DATE

We have developed the first version of the artwork, and have used it to successfully demonstrate that human interactants can entrain themselves to their human-representative agent [2]. We elicited entrainment in the interactant re her/his movement and gesture captured by an overhead camera. Qualitative evaluation of videotaped interactions with multiple participants demonstrate that human interactions

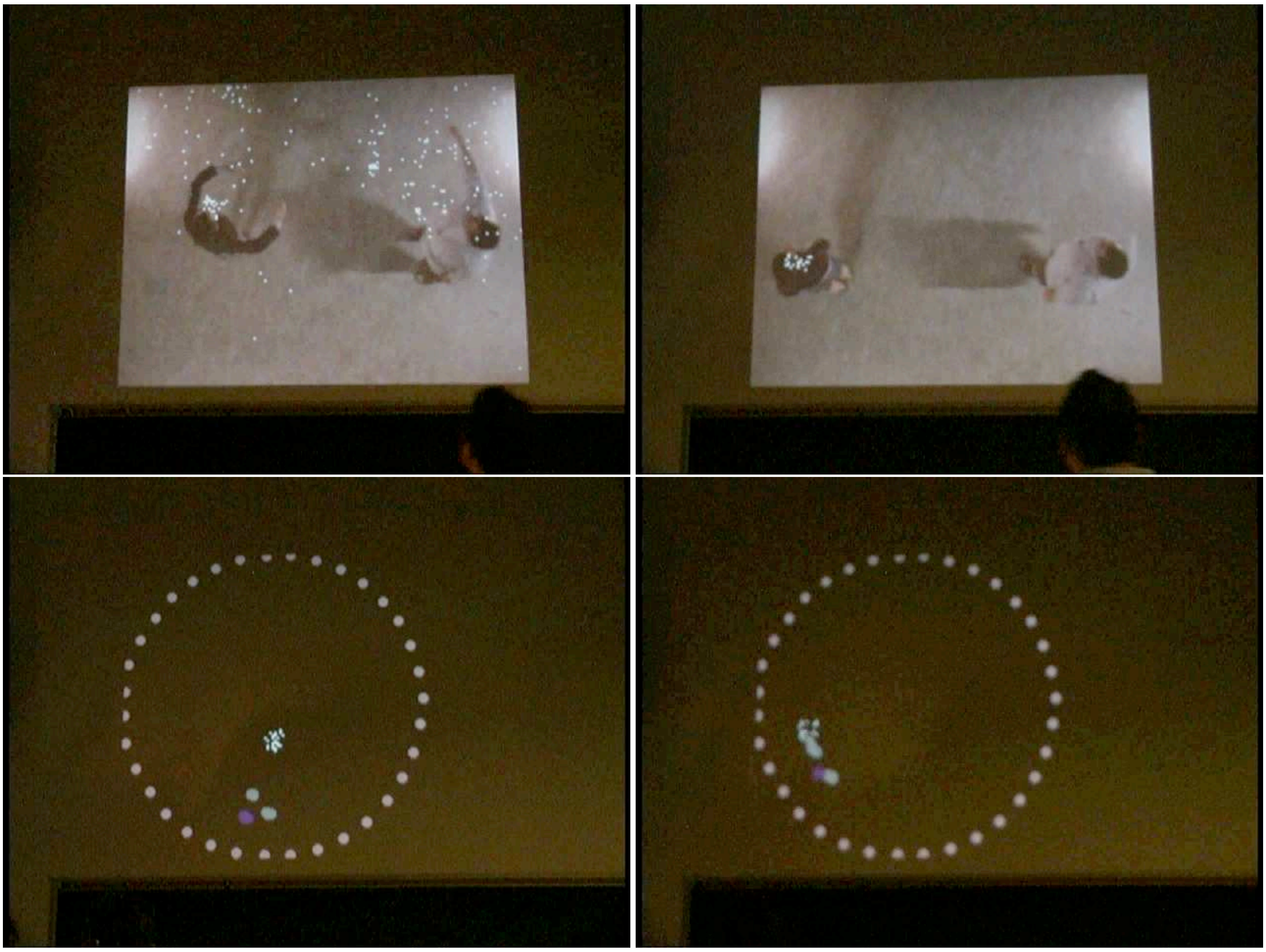


Figure 3. Illustration of process of entrainment, shown as a sequence of snapshots (shown clockwise, starting in top lefthand corner). The interactants are engaged with the prototype (in which the agents are instantiated using LCD projections and not hardware-instantiated low-fidelity enmbodiments).

are able to entrain themselves to the human-representative agent in the population. This phenomenon is illustrated in figure 3.

We have built rules into the system, but they are not made explicit — rather, they are discovered by the human participants as they interact with the A-life sculpture. This process results in a co-construction of the artwork, in the sense that experience of the work is different for each participant, and many facets of the work are not immediately available but appear during the time spent with it. Arising from this is one of the central challenges of this project: the need to devise means by which interactants can recognize and interpret emergent phenomena, without recourse to overt explanations using language. The means for interpretation should be interesting to both artistic and scientific questions.

In particular, we are investigating how low-fidelity embodiments can be used effectively. The navigation task we have implemented is a form of herding that requires collaborative movements between the human representative agent and the virtual agents. A software prototype has been developed in which the agents are represented simply as two-dimensional circles (which functioned as extremely low-fidelity embodiments; only the attributes of colour and spatial placement served as "output" modes). In the next prototype iteration, the agent population will be given slightly different low-fidelity embodiments. In this task, the agents in the population must direct the target entity into what we term *the nest*. The behaviour of the target entity is guided by the simple goal *to evade the agents*. The behaviour of the artificial agents is implemented by a set of simple rules that express attraction and repulsions to other entities in the arena (e.g., attraction to the target and the nest, repulsion from close proximity to other agents and the arena walls). From these simple rules, complex collective behaviours emerge. In the mixed population of human-representative and artificial agents, all of the agents must work together to accomplish the set task.

4 DISCUSSION

We are currently working on an artwork that will elicit entrainment in the interactant with respect to selecting a turn-taking conversation mode strategy. We anticipate that this will come about by the interactant "practising" a particular mode strategy for a certain number of turns, then switching. For example, the interactant may learn to concede or offer the turn by moving away from the artwork; barge in by moving in closer to the artwork (which might not necessarily be successful since the artificial agents have built-in counter-strategies to attempted barge-ins, so repeated tries might ensue). The desired result is the realization by the interactant that their human-representative agent can converse with others.

We plan to experiment with posing different goals for the interaction. For instance, one goal would be for the agent population to make an image of turn-taking become visible in a real-time, low-resolution video display and audible in sound sequences. A form of optimizing the conversational behaviour can be to assign an advantage to a virtual agent in choosing one mode strategy, one that is private yet leads to the common goal of social exchange. In order to accomplish this goal, the agents must discover and settle into a target pattern of turn-taking (possible target patterns include each agent takes one turn in sequence, or the turn alternates between one dominant agent and each of the other agents in turn; different variants could easily be programmed into the agent architecture). The sound differentiates between human and virtual agent actions, so it provides a cue to both distinguishing among actions and successfully combining them into a target pattern. The clarity of the target image in the low-resolution video and audio display is a function of the similarity of the population's actual turn-taking pattern to the target pattern (e.g., the more similar the turn-taking pattern, the more distinct the image). Moreover, regions of the image will be associated with each agent, so that the impact of a particular agent's lack of cooperation can be visualized. Thus, making the video image emerge, via the activity of making a particular turn-taking pattern emerge, is the collective goal of the agents in the population. The behaviour of the population co-evolves. When a new interactant enters, the pre-existing members of the population will attempt to resume a target pattern of turn taking. If the human interactant chooses — and is able — to cooperate, then the target image will re-emerge in the low-resolution video display. In order for the human interactant to participate, he or she must adapt his or her actions to those of the population.

5 CONCLUSION

We have described a project that integrates both artistic and scientific goals. Each of these types of goals carries their own perspectives on the motivation for building the artefact (e.g., interactive artwork and/or computational interactive system). By comparing and contrasting these perspectives, however, we have identified assumptions about "scientific" computational system building (e.g., a bias toward quantitative, performance-based evaluation) and identified a role for "artistic" processes (e.g., the role of aesthetics, the role of the interactant's own mental processes not only about the interaction, but about his or her own assumptions and biases about the virtual interactants themselves).

ACKNOWLEDGEMENTS

The authors would like to thank the Canada Council for the Arts and the Natural Sciences and Engineering Research Council of Canada for financial support of this research project, under the auspices of the New Media Intiative. The sound artist for the project is John Kamevaar. Thanks also to Computer Science graduate student David Jacob for his work on developing and implementing the software of the agent architectures for performing the herding task and the agent population environment.

REFERENCES

[1] Antonella De Angeli, Sheryl Brahnam, Peter Wallis, and Alan Dix, eds. *Proceedings of the CHI 2006 Workshop on Misuse and Abuse of Interactive Technologies*, Montreal, Canada, April 22 2006.

[2] Melanie Baljko and Nell Tenhaaf, 'Different experiences, different types of emergence: A-life sculpture designer, interactant, observer', in *Proceedings of AAAI Fall 2006 Symposium on Interaction and Emergent Phenomenon in Societies of Agents*, Arlington, VA, (12–15 October 2006).

[3] *Embodied Conversational Agents*, eds., Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, MIT Press, April 2000.

[4] H. H. Clark, *Using Language*, Cambridge University Press, 1996.

[5] K. Dautenhahn, 'The art of designing socially intelligent agents: Science, fiction, and the human in the loop', *Applied Artificial Intelligence*, **12**(5), 573–617, (1998).

[6] Lars Hallnas and Johan Redstrom, 'From use to presence: On the expressions and aesthetics of everyday computational things', *ACM Transactions on Computer-Human Interaction (TOCHI)*, **9**(2), 106–124, (June 2002).

[7] Peter A. Heeman and Graeme Hirst, 'Collaborating on referring expressions', *Computational Linguistics*, **21**(3), 351–382, (1995).

[8] M. Maybury, *Intelligent Multimedia Interfaces*, AAAI Press/MIT Press, Cambridge, MA, 1993.

[9] Tim Paek and Eric Horvitz, 'Conversation as action under uncertainty', in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI-2000)*, pp. 455–464, Stanford, CA, (June 2000).

[10] Michael R. Perkins, 'Is pragmatics epiphenomenal? evidence from communication disorders', *Journal of Pragmatics*, **29**(3), 291–311, (1998).

[11] Marianne Graves Petersen, Ole Sejer Iversen, and Peter Gall Krogh, 'Aesthetic interaction — a pragmatist's aesthetics of interactive systems', in *Proceedings of the 2004 conference on Designing Interactive Systems (DIS'04)*, pp. 269–274, Cambridge, MA,, (2004).

[12] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places*, Cambridge University Press, 1996.

[13] H. Sacks, E. A. Schegloff, and G. A. Jefferson, 'A simplest systematics for the organization of turn-taking in conversation', *Language*, **50**, 696–735, (1974).

[14] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph "Jofish" Kaye, 'Reflective design', in *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility CC'05*, pp. 49–58, Arhus, Denmark, (August 21–25 2005).

[15] David R. Traum, 'Speech acts for dialogue agents', in *Foundations And Theories Of Rational Agents*, eds., Michael Wooldridge and Anand Rao, 169–201, Kluwer Academic Publishers, (1999).

[16] Jason D. Williams, Pascal Poupart, and Steve Young, 'Partially observable markov decision processes with continuous observations for dialogue management', in *Proceedings of the 6th SigDial Workshop on Discourse and Dialogue*, Lisbon, Portugal, (2005).

Facial Feedback Signals for ECAs

Elisabetta Bevacqua[1] and **Dirk Heylen**[2] and **Catherine Pelachaud**[1] and **Marion Tellier**[1]

Abstract. One of the most desirable characteristics of an intelligent interactive system is its capability of interacting with users in a natural way. An example of such a system is the embodied conversational agent (ECA) that has a humanoid aspect and the capability of communicating with users through multiple modalities such as voice, gesture, facial expressions, that are typical of human-human communication. It is important to make an ECA able to fit well in each role in a conversation: the agent should behave in a realistic and human-like way both while speaking and listening. So far most of the work on ECAs have focused on the importance of the ECA's behaviour in the role of the speaker, implementing models for the generation of verbal and non-verbal signals; but currently we are mainly interested in modelling the listening behaviour. In this paper we will describe our work in progress on this matter.

1 Introduction

In conversations participants produce behaviours that are intended to convey meaning or intentions. The producer of communicative behaviours wants the intentions he has with them to be recognized by the addressees of his message. Conversation is thus a particular, socially developed instrument to enable mindreading. Communication as we understand it here requires a Theory of Mind on the side of both producers and recipients of communicative behaviours. Producers need to design their communicative actions taking into account what they believe to be the mental state of the recipients (audience design). Recipients need to be able to recognize that behaviours were produced because of an intentional action. This is the notion of non-natural meaning as discussed by Grice [14]. In [18], Levinson paraphrases Grice's definition of non-natural meaning as follows:

> [C]ommunication consists of the 'sender' intending to cause the 'receiver' to think or do something, just by getting the 'receiver' to recognize that the 'sender'is trying to cause that thought or action. So communication is a complex kind of intention that is achieved or satisfied just by being recognized. In the process of communication, the 'sender's' communicative intention becomes mutual knowledge to 'sender' (S) and 'receiver' (H), i.e. S knows that H knows that S knows that H knows (and so ad infinitum) that S has this particular intention. Attaining this state of mutual knowledge of a communicative intention is to have successfully communicated. (Levinson, p. 16)

During a conversation the listener is called to provide information on the successfulness of the communication. In order to ensure closure on the communicative actions, speakers will monitor listeners for cues of recognition to establish grounding which can be both natural cues as intentional signals produced by listeners to provide feedback on the speech. The term back-channel (stemming from [26]) is commonly used to denote the communicative behaviours that are produced by participants in a conversation as feedback on the reception of the communicative behaviours of the other participants. Both through linguistic and gestural signals, the listener can show his level of engagement in the conversation. According to the listener's responses the speaker can estimate how his/her interlocutor is reacting and can decide how to carry on the interaction: for example by interrupting the conversation if the listener is not interested or re-formulating a sentence if the listener showed signs of not understanding and so on.

In our research, we want to analyse not only how this behaviour is displayed, but also what kind of information it provides about the listener's reaction towards the speaker and his/her speech. Our aim is thus twofold: on the one hand we want to implement back-channel behaviour in a conversational agent in order to make it more realistic and human-like, and on the other hand we want to make sure that the user is able to interpret the agent's signals as 'intended' by the ECA, so that the user feels the ECA is displaying the appropriate level of understanding and participates actively in the conversation.

Through one or more channels like voice, head, face, gaze, posture and gesture, listeners provide back-channels signals of perception, attention, interest, understanding, attitude (belief, liking. . .) and acceptance towards what the speaker is saying [26, 1, 21]. A back-channel can be positive or negative and can have several meanings (understanding but not acceptance, believing but not agreeing and so on). Moreover, the listener can emit signals with different levels of control and intentionality: consciously deciding to emit a signal in order to show a reaction to the speaker's speech (and even deliberately choosing a specific one to provoke a particular effect on the speaker, for example: the listener decides to stare at the speaker to show disbelief or surprise expecting a confirmation by the speaker) or emitting cues without thinking, automatically reacting to the speaker's behaviour or speech, generating back-channels at a very low level of control [1].

In this paper we present our first experiments along these lines. We start with a characterisation of back-channels. Then, we present a perceptual test we have conducted and preliminary results. Finally we explain how we aim at introducing the evaluated signals in a computational model for a conversational virtual **listener**.

[1] University of Paris8, IUT de Montreuil, 140 rue de la Nouvelle France, 93100, Montreuil, France
email: e.bevacqua,c.pelachaud@iut.univ-paris8.fr

[2] University of Twente PoBOX 217, 7500 AE Enschede, The Netherlands, email: heylen@ewi.utwente.nl

2 Back-channels

Several research traditions have studied the behaviours that listeners display in conversations. Back-channels, or similar phenomena with a different name such as response tokens, have been studied in the conversational analysis literature, for instance, with the purpose of understanding what role the various contributions of all of the participants play in shaping the conversation. Most relevant in this respect are papers such as [23], [24], [16] but there are many others. The literature on turn-taking, both from the CA and other perspectives, also provides useful notes on the behaviours of participants that assume the primary speaker role and the auditors. In the series of papers by Duncan and co-authors[3], for instance, auditor back-channel signal are one of three classes of signals, besides speaker within-turn and speaker continuation signals, that serve to mark units of interaction during speaking turns.

A general assumption behind the concept of back-channel is that all the participants in a face-to-face conversation are both producers and recipients of communicative signals, but that there are different levels on which this occurs. Communicative signals on the primary track, to use the term by [5], are by the participants that have the floor and the secondary track, 'in the back', is constituted by the feedback on the behaviours in the primary track. As [26] points out there may be cases of iteration where speakers provide feedback on the back-channels of listeners.

Several studies of nonverbal behaviours have paid attention to the behaviours displayed by listeners. One kind of phenomenon that has received some attention is the way in which behaviours of participants are synchronized and in particular how body movements of listeners are coordinated with the verbal utterances of the speaker. [15] showed that about a quarter of the head movements by listeners are in sync with the speaker's speech. Interactional synchrony in this sense has been studied, amongst others by [17], [22], [6]. Mirroring is a particular type that has often been commented upon. Scheflen suggests that this often reflects a shared viewpoint. Also [17] hypothesized that the level to which behaviours are synchronized may signal the degree of understanding, agreement or support. These kinds of phenomena show that the behaviours of listeners arise not only from 'structural concerns' (e.g. turn-taking signals) but also from 'ritual concerns'. We take these terms from [12] who points out that it is sheer impossible to assign to behaviours a function of only one of these types of concerns (see also [3]).

Besides these synchrony behaviours, listeners display various other nonverbal behaviours as feedback. [4], looking in particular at facial expressions, classifies these behaviours in a small set of semantic categories of listener comment displays. These are, besides displays for agreement:

- Back-channel: Displays that were produced by listeners while the speaker was talking or at the end of the speaker's turn. They take the form of brow raises, mouth corners turned down, eyes closed, lips pressed. In Chovil's corpus the displays could be accompanied by typical back-channel vocalizations such as "uhuh", "mhmm", "yeah", etc.
- Personal reaction displays: A reaction in response to what the speaker had said rather than just acknowledging the content.
- Motor mimicry displays: displays that might occur in the actual situation that the speaker is talking about (e.g. wincing after hitting ones' thumb with a hammer, eyes widened and an open mouth in response to a frightening situation). These are interpreted as messages that indicated a sincere appreciation of the situation being described.

In the discussion so far, we have mentioned several functions that are served by the behaviours of listeners. They provide feedback to the speaker, acknowledging reception of the signal, possibly its understanding or some kind of comment expressing a particular attitude towards what is being expressed. From its nature as a kind of joint communicative action, conversations require that participants come to react to each other's actions to ground the actions and provide closure. Feedback is an important part of establishing grounding in the interactional achievement of having a conversation. The variety of functions that feedback serves is partly explained by the various levels on which grounding needs to take place: i.e. levels at which the participants need to have a mutual understanding of each other's intentions. [5] suggests that grounding needs to occur on at least four levels with each step a kind of joint action.

1. Joint[A executes behavior t for B to perceive; B attends perceptually to behavior t from A]
2. Joint[A presents signal s to B, B identifies signal s from A]
3. Joint[A signals to B that p, B recognizes that A means that p]
4. Joint[A proposes a joint project to B, B takes up the joint project]

As speakers make their utterances, they are usually also monitoring the interlocutors behaviours to find signs of their participatory involvedness on all of these levels.

1. A monitors B for signs of perception activity / B's behaviour provides cues of perception activity
2. A monitors B for signs that B has identified the signal / B indicates that he has identified the signal...

The utterance of speakers and the accompanying behaviours will often be designed to invoke behaviours of interlocutors to ensure this. A typical case of this behaviour is analysed by [13], consisting of hesitations and repetitions of speakers at the beginning of their utterance to evoke gaze behaviours in interlocutors.

In a similar vein, [1] distinguishes four basic communicative functions on which the speaker may require feedback:

1. Contact: is the interlocutor willing and able to continue the interaction
2. Perception: is the interlocutor willing and able to perceive the message
3. Understanding: is the interlocutor willing and able to understand the message
4. Attitude: is the interlocutor willing and able to react and respond to the message (specifically accepting or rejecting it).

The various feedback behaviours are thus not only varied in their form but also in their function. In one of the experiments that we are carrying out and report on below, we are looking at back-channel behaviours in which facial expressions, gaze, and head movements are controlled. We look at various classes of expressivity. The general classes that we consider are the following:

- Performatives such as agree, disagree, criticize, refuse, accept, approve, confirm, question
- Affectives: liking, disliking, disgust, sorry-for, surprise, fear, anger, reproach, gratitude
- Epistemics: believe, disbelieve, scepticism, certainty, doubt

[3] See [7], [8], [9], [10], [11],.

- Meta-cognitives: thinking, planning, remembering

These functions and behaviours go beyond the usual back-channel behaviours such as nodding that are mostly discussed in the computational literature. An important issue to consider is the degree to which people agree on the interpretation of the behaviours. The experiment described next is supposed to shed some light on this.

3 Recognition Test

3.1 Hypotheses

Our first hypothesis is that most back-channel signals either convey a positive or a negative connotation. Therefore, we are trying to find out the general meaning for each signal when there is one. We can assume that signals containing nods and smiles will be interpreted as positive feedback signals such as agree, accept, like, understanding and believe whereas signals containing shakes, frown and tension will rather be associated with negative meanings such as disagree, refuse, dislike, do not understand and disbelieve. Our second hypothesis is that back-channel signals are polysemic: the same signal can have different meanings and a single meaning can be expressed with different signals or a combination of signals. We are thus assuming that a single signal can be interpreted by subjects in different ways. To test these hypotheses we have conducted recognition tests on subjects who were asked to judge a set of 14 different signals displayed by a 3D agent Greta [20].

3.2 Participants

Twelve French students have been tested so far. They are students in computer science, age range 18-20.

3.3 Material

The test was done with our 3D agent Greta. The graphic interface of the test application can be seen in Figure 1. In the little window on the left Greta's videos are shown once at a time. Two buttons under the window, *play* and *next movie*, allow the user respectively to play the movie (in this way the movie is shown only when the user is paying attention) and to move on to the next movie. For a more controlled procedure, we decided that participants could not rewind the video. On the right a list of possible meanings is proposed to the participant who, after each movie and before moving on, can select one meaning according to his/her opinion about which meaning fits that particular back-channel signal best. It is possible to select several meanings for one signal and when none of the meanings seems to fit, partecipants can just click on *next movie*. In this case the absence of answers will be treated as "no answer" in the data.

In this test we decided to consider the meanings belonging to the class of expressivity *performative*:

- agree (AG)
- disagree (DA)
- accept (AC)
- refuse (RE)
- interested (IN)
- bored (BO)

From the class *epistemics* we selected the following meanings:

- believe (BE)
- disbelieve (DB)
- understand (UN)
- not understand (NU)

Finally, from the class *affective*, we considered:

- like (LI)
- dislike (DL)

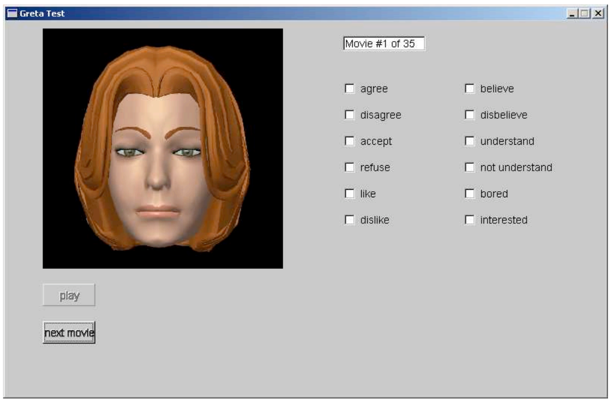


Figure 1. Graphic interface of the test application

To make the videos we have selected 14 signals chosen among the back-channel signals which were analysed and proposed by [2, 21]. Some signals are simple, containing just a single action (like a nod or a shake), while others are obtained by combining several actions (like a nod and a raise eyebrows or a head tilt and a frown). The 14 signals are:

1. a single head nod (N)
2. a head nod with a smile (NS)
3. a head nod and a raise of the eyebrows (NRE)
4. a head shake (S)
5. a head shake and a frown (SF)
6. a head shake, a frown and a tension in the lips that tighten getting thinner (SFT)
7. a frown and a tension in the lips that tighten getting thinner (FT)
8. a raise of the left eyebrow (RLE)
9. the eyes roll up in the head (ER)
10. a head tilt on the left and sad eyebrows (TSE)
11. a head tilt on the left and a frown (TF)
12. a head tilt on the right and raise eyebrows (TRE)
13. a head tilt on the right and the gaze turns on the down right (TG)
14. the eyes wide open (EWO)

3.4 Procedure

Participants were given instructions for the test through a written text. They were told that Greta would display back-channel signals as if she was talking to an imaginary speaker. They were asked to evaluate these signals by choosing among the available list of meanings. This way we made sure that participants were aware that they were evaluating back-channel signals. The signals were shown randomly at least twice and at most three times so that the participants had to watch 35 movies in all: we wanted to find out whether people gave always the same answer, or if they tended to remember the signals and associate them to more possible meanings.

Signals	Positive answers	Negative answers	No answer	Total of answers
N	26	2	0	28
NS	38	1	0	39
NRE	57	1	0	60
RLE	17	22	3	42
TSE	5	27	0	32
ER	0	21	4	25
TG	6	29	8	43
SFT	2	35	0	37
FT	3	27	0	30
S	0	38	2	40
SF	1	31	0	32
TF	8	33	1	42
TRE	18	20	2	40
EWO	11	13	13	37

Table 1. Results positive and negative for each signal.

3.5 Results and Discussion

Signals	Significant	Meaning
N	Yes p<0.0001	positive
NS	Yes p<0.0001	positive
NRE	Yes p<0.0001	positive
RLE	No p=0.5224	No distinct meaning
TSE	Yes p<0.0001	positive
ER	Yes p<0.0001	positive
TG	Yes p<0.0001	positive
SFT	Yes p<0.0001	positive
FT	Yes p<0.0001	positive
S	Yes p<0.0001	positive
SF	Yes p<0.0001	positive
TF	Yes p<0.0001	positive
TRE	No p=0.8714	No distinct meaning
EWO	No p=0.8388	No distinct meaning

Table 2. Results of the binomial tests.

One of our main concerns in this experiment was to find out whether certain signals are globally considered positive or negative. We also expected to find meaningless signals that is to say signals that do not convey positive nor negative meaning on their own and need to be matched with other signals to be meaningful. To analyse the data, we coded the answers given by the subjects as positive or negative, according to the following principles: agree, accept, like, believe, understand and interested were considered as positive answers and disagree, refuse, dislike, not understand and bored were considered as negative answers. Table 1 shows the results for each signal. For the treatment of the data we have left out the cases in which subjects have not answered ("no answer") but it will be taken into account during the analysis of the results when relevant. The differences in the total number of answers is explained by the fact that some signals have been presented twice to subjects and others three times and by the fact that subjects could give several answers for each signal. The null hypothesis is that a signal has no distinct meaning (positive or negative) so that there is no significant difference between the amount of positive and negative answers for that signal. The alternative hypothesis is that the amount of answers for one signal is so high that it proves that subjects detected a distinct meaning. To test the hypothesis, binomial tests have been performed for each signal. Signals for which the p value is less than the 0.05 level of significance, reject the null hypothesis. Table 2 shows the results.

Thus, only three signals do not reject the null hypothesis: "raise of the left eyebrow", "head tilt on the right" and "raise eyebrows and eyes wide open". This means that these three signals do not convey enough meaning when displayed alone. Looking at the distribution of the answers, we can notice that subjects' answers are almost equally shared between positive and negative items and for the last signal, "eyes wide open", the amount of "no answer" is extremely high (13 out of 37) which confirms the meaningless aspect of this particular signal. Every other signal reject the null hypothesis which proves that they either convey a positive or negative connotation. Our data shows that the positive meaning of "head nod", "head nod and smile" as well as "head nod and raise of the eyebrows" is significant. It also shows that the negative aspect of "head tilt on the left and sad eyebrows", "eyes roll up in the head", "head tilt and gaze", "head shake, frown and tension", "frown and tension", "head shake", "head shake and frown" and "head tilt on the left and frown" is significant.

Table 3 shows the statistical results en percentage, signals were played two or three times and the table contains the results of the all the repetitions. On the rows there are the signals, while on the column (from the second to the fourteenth one) there are the meanings. The first column (#Ans) contains the number of answers given for the corresponding signal.

In general we have seen that subjects tend to give more and more answers for each signal as the test goes on, probably because they become accustomed to the movies and to the aim of the test. Moreover the more complex is the signal the more answers the subjects gave. For example "head nod and smile" obtained 39 answers while "head nod" 28. "head nod and raise of the eyebrows" had 60 answers, but it is important to notice that it was displayed three times while "head nod and smile" and "head nod" just twice. However in the first two repetitions the signal “head nod and raise of the eyebrows” obtained more answers that “head nod and smile” and “head nod”.

We have the same result for the signals “head shake”, “head shake and frown” and “head shake, frown and tension”; results in the table 3 show that “head shake” obtained more answers than “head shake and frown” and “head shake, frown and tension” which are more complex signals, however “head shake” was displayed three times while “head shake and frown” and “head shake, frown and tension” just twice. During the first two repetitions “head shake and frown” and “head shake, frown and tension” still obtained more answers than “head shake”.

As expected, participants associated positive meanings to signals containing nods and smiles and in particular they related the smile to the meaning of liking (39.90%). Negative meanings were linked to shakes and frowns; for example the signal “head shake and frown” was associated above all to refuse and disagree (37.5%). The other signals were less easily associated to a constant set of meanings, as we assumed head tilts and rolling of the eyes were seen as signals of disbelief, not understanding and boredom, but they also suffered the more evident dispersion of answers and sometimes the percentage are not so relevant. For example, the signal “head tilt and gaze” (TG) was interpreted above all as a back-channel of boredom (37.20%), but all the other meanings were also selected at least once and the 18.60% of answers was “no answer”. The signal “eyes wide open” was the hardest to interpret, most answers were “no answer” (35.14%) and even if the second highest percentage classify this signal as a back-channel of disbelief, such percentage is not very relevant (16.23%). Perhaps these signals were hardest to interpret because they can convey more meanings according to the context and to the listener’s personality.

Some of the signals we took into account in this test are complex signals, composed by several single actions which have not been all tested individually. Thus, in further experiments we will analyse some actions separately, for example “smile”, “frown” and “sad eyebrows”.

4 Future Work

In the future we will submit the test to a more relevant number of subjects in order to obtain more accurate and significant results. Moreover we aim at proposing this test to subjects of different cultures in order to see if back-channel signals are perceived differently in other countries or if they are interpreted in the same way.

With this test we also want to define a set of recognizable signals to be used in the implementation of a listener model for our conversational agent. As we said in the Introduction, the listener can emit signals with different levels of control and intentionality, he can provide a back-channel consciously or unconsciously. Consequently a single listener model is not enough; two computational models are needed, respectively a cognitive model (to generate intentional back-channel signals) and a reactive model (to generate non-intentional back-channel signals). Since the instinctive listener’s back-channel is often elicited by the speaker’s behaviour, a set of rules can be defined to implement a reactive model [19]. For example, from a corpus of data, Maatman derived a list of rules useful to predict when back-channel can occur according to the speaker’s actions. Back-channel continuers (like head nods, verbal responses) appear at a pitch variation in the speaker’s voice; frowns, body movements and gaze shifts are produced when the speaker shows uncertainty; facial expressions, postural and gaze shifts are provided to reflect those made by the speaker (mimicry). Even variation in the speaker’s pitch of voice usually elicits a back-channel signal from the listener [25].

As for the cognitive model, it is complex to implement. To elaborate reasoned reactions from a listener, one must have access to not only the extrapolated speech content, but also information about the listener’s personality. For this reason, we will begin by implementing a Wizard of Oz system to provide consciously back-channel. The intentional listener behaviour is driven by a wizard while our virtual agent interacts with a user.

REFERENCES

[1] J. Allwood, J. Nivre, and E. Ahlsén, ‘On the semantics and pragmatics of linguistic feedback’, *Semantics*, **9**(1), (1993).

[2] Jens Allwood and L. Cerrato, ‘A study of gestural feedback expressions’, in *First Nordic Symposium on Multimodal Communication*, eds., P. Paggio, K. Jokinen, and A. Jönsson, pp. 7–22, Copenhagen, (2003).

[3] J.E.G.F.J. Bernieri, ‘The importance of nonverbal cues in judging rapport’, *Journal of Nonverbal behavior*, **23**(4), 253–269, (1999).

[4] Nicole Chovil, ‘Social determinants of facial displays’, *Journal of Nonverbal Behavior*, **15**(3), 141–154, (1991).

[5] Herbert Clark, *Using Language*, Cambridge University Press, Cambridge, 1996.

[6] W.S. Condon and W.D. Ogston, ‘A segmentation of behavior’, *Journal of Psychiatry*, **5**, 221–235, (1967).

[7] S.D. Duncan, ‘Some signals and rules for taking speaking turns in conversations’, *Journal of Personality and Social Psychology*, **23**, 283–92, (1972).

[8] S.D. Duncan, ‘Towards a grammar for dyadic conversations’, *Semiotica*, 29–46, (1973).

[9] S.D. Duncan, ‘On the structure of speaker-auditor interaction during speaking turns’, *Language in Society*, **2**, 161–180, (1974).

[10] S.D. Duncan, ‘Language, paralanguage, and body motion in the structure of conversations’, in *Language and Man. Anthropological Issues*, eds., W.C. McCormack and S.A. Wurm, 239–268, Mouton, The Hague, (1976).

[11] S.D. Duncan and G. Niederehe, ‘On signalling that its your turn to speak’, *Journal of Experimental Social Psychology*, **10**, 234–47, (1974).

[12] E. Goffman, *Forms of Talk*, Oxford University Press, Oxford, 1981.

[13] C. Goodwin, *Conversational Organization: Interaction between Speakers and Hearers*, Academic Press, New York, 1981.

[14] H. P. Grice, ‘Logic and conversation’, in *Syntax and Semantics: Speech Acts*, 41–58, Academic Press, New York, (1975).

[15] U. Hadar, T.J. Steiner, and Clifford F. Rose, ‘Head movement during listening turns in conversation’, *Journal of Nonverbal Behavior*, **9**(4), 214–228, (1985).

[16] John Heritage, ‘A change-of-state token and aspects of its sequential placement’, in *Structures of Social Action*, eds., J. M. Atkingson and J. Heritage, Cambridge University Press, Cambridge, (1984).

[17] Adam Kendon, ‘Movement coordination in social interaction: some examples described’, *Acta Psychologica*, **32**, 100–125, (1970).

[18] Stephen C. Levinson, ‘Putting linguistics on a proper footing: explorations in goffman’s concept of participation’, in *Erving Goffman. Exploring the Interaction Order*, eds., Paul Drew and Anthony Wootton, 161–227, Polity Press, Cambridge, (1988).

[19] R.M. Maatman, J. Gratch, and S. Marsella, ‘Natural behavior of a listening agent’, in *5th International Conference on Interactive Virtual Agents*, Kos, Greece, (2005).

[20] Catherine Pelachaud and Massimo Bilvi, ‘Computational model of believable conversational agents’, in *Communication in Multiagent Systems*, ed., Marc-Philippe Huget, volume 2650 of *Lecture Notes in Computer Science*, 300–317, Springer-Verlag, (2003).

[21] I. Poggi, ‘Backchannel: from humans to embodied agents’, in *Conversational Informatics for Supporting Social Intelligence and Interaction - Situational and Environmental Information Enforcing Involvement in Conversation workshop in AISB’05*, University of Hertfordshire, Hatfield, England, (2005).

[22] A.E. Scheflen, ‘The significance of posture in communication systems’, *Psychiatry*, **27**, 316–331, (1964).

[23] Emanuel A. Schegloff, ‘Discourse as interactional achievement: Some uses of ”uh huh” and other things that come between sentences’, in

	#Ans	AG	DA	AC	RE	LI	DL	BE	DB	UN	NU	BO	IN	NONE
N	28	21.43		39.28		7.15	3.75	3.75		17.86		3.75	3.75	
NS	39	17.95		15.38		35.90		5.13		5.13		2.56	17.95	
NRE	60	38.33		20	1.67	6.67		6.67		13.33		3.33	10	
RLE	42	14.28	2.38	4.77	4.77	2.38		7.14	23.80	4.77	9.53	11.90	7.14	7.14
TSE	32		12.5		6.25		9.38	3.12	18.75	3.12	31.25	6.25	9.38	
ER	25		4		4		12		20		12	32		16
TG	43	2.33	4.65	2.33	4.65	2.33	6.97	2.33	9.30	2.33	4.65	37.20	2.33	18.60
SFT	37	2.70	24.33		24.33		18.91		8.10		16.22	2.70	2.70	
FT	30		36.67	3.33	20	3.33	3.33		13.34		13.34	3.33	3.33	
S	40		25		40		20		7.5		2.5			5
SF	32		37.5		37.5		6.25		6.25	3.13	9.37			
TF	42	4.76	19.04	2.39			9.52	2.39	19.04	7.14	21.42	9.52	2.39	2.39
TRE	40	2	10	10		5	5	17.5	12.5		7.5	15	7.5	5
EWO	37	5.40	5.40	5.40			2.70		16.23	5.40	5.40	5.40	13.53	35.14

Table 3. Statistical results en percentage. On the rows there are the signals, while on the column there are the meanings. In Table 4 a reminder of the meaning of the abbrevations.

Signals		Meanings	
N	a single head nod	**AG**	agree
NS	a head nod with a smile	**DA**	disagree
NRE	a head nod and a raise of the eyebrows	**AC**	accept
RLE	a raise of the left eyebrows	**RE**	refuse
TSE	a head tilt on the left and sad eyebrows	**IN**	interested
ER	the eyes roll up in the head	**BO**	bored
TG	a head tilt on the right and the gaze turns on the right and down	**BE**	believe
SFT	a head shake, a frown and a tension in the lips that tighten getting thinner	**DB**	disbelieve
FT	a frown and a tension in the lips that tighten getting thinner	**UN**	understand
S	a head shake	**NU**	not understand
SF	a head shake and a frown	**LI**	like
TF	a head tilt on the left and a frown	**DL**	dislike
TRE	a head tilt on the right and a raise eyebrows		
EWO	the eyes wide open		
#Ans	number of answers given for the corresponding signal		

Table 4. Meanings of the abbrevations.

Analyzing discourse, text, and talk, ed., D. Tannen, 71–93, Georgetown University Press, Washington, DC, (1982).

[24] Emanuel A. Schegloff, 'Issues of relevance for discourse analysis: Contingency in action, interaction and co-participant context', in *Computational and Conversational Discourse. Burning issues - An interdisciplinary account*, eds., Edward H. Hovy and Donia R. Scott, 3–35, Springer, (1996).

[25] N. Ward and W. Tsukahara, 'Prosodic features which cue back-channel responses in english and japanes', *Journal of Pragmatics*, **23**, 1177–1207, (2000).

[26] V.H. Yngve, 'On getting a word in edgewise', in *Papers from the sixth regional meeting of the Chicago Linguistic Society*, pp. 567–77, Chicago: Chicago Linguistic Society, (1970).

A Two-Level BDI-Agent Model for Theory of Mind and its Use in Social Manipulation

Tibor Bosse, Zulfiqar A. Memon, Jan Treur[1]

Abstract. This paper introduces a formal BDI-based agent model for Theory of Mind. The model uses BDI-concepts to describe the reasoning process of an agent that reasons about the reasoning process of another agent, which is also based on BDI-concepts. A case study illustrates how the model can be used for social manipulation. This case study addresses the scenario of a manager that reasons about the task avoiding behaviour of his employee. For this scenario, a number of simulation experiments have been performed, and some of their results are discussed.

1 INTRODUCTION

To function efficiently in social life and within organisations, it is useful if agents can reason about the actual and potential behaviour of the agents around them. To this end, it is very helpful for these agents to have capabilities to predict in which circumstances other agents will show certain appropriate or inappropriate behaviours. If for a considered other agent, generation of actions is assumed to be based on a BDI-model, prediction of such actions will involve reasoning based on a Theory of Mind [3], [5], [16] involving beliefs, desires and intentions as a basis for the behaviour. Reasoning based on such a Theory of Mind can be exploited in two different manners. The first manner is just to predict the behaviour in advance, in order to be prepared that it will occur (*social anticipation*). For example, if an agent B has done things that are known as absolutely unacceptable for an organisation (or a relationship), then he or she may be able to predict and therefore be prepared on what will happen after a manager (or partner) agent A learns about it.

A second manner to exploit reasoning based on a Theory of Mind is to try to affect the occurrence of certain beliefs, desires and intentions at forehand, by manipulating the occurrence of circumstances that are likely to lead to them (*social manipulation*). For example, the agent B just mentioned can try to hide facts so that the manager (or partner) agent A will never learn about the issue. Such capabilities of anticipatory and manipulatory reasoning based on a Theory of Mind about the behaviour of colleague agents are considered quite important, not to say essential, to function smoothly in social life.

This type of reasoning has an information acquisition and analysis aspect, and a preparation and action aspect. To describe the latter aspect, for the agent using a Theory of Mind, a model for action preparation based on beliefs, desires and intentions can be used as well. For example, for agent B discussed above, the desire can be generated that agent A will not perform the action to fire (or break up with) him or her, and that agent A will in particular not generate the desire or intention to do so. Based on this desire, the refined desire can be generated that agent A will not learn about the issue. Based on the latter desire, an intention and action can be generated to hide facts for agent A. Notice that agent B reasons on the basis of BDI-models at two different levels, one for B itself, and one as the basis for the Theory of Mind to reason about agent A. It is this two-level architecture that is worked out in this paper in a computational model.

The modelling approach used for this computational model is based on the modelling language LEADSTO [7]. In this language, direct temporal dependencies between two state properties in successive states are modelled by *executable dynamic properties*. The LEADSTO format is defined as follows. Let α and β be state properties of the form 'conjunction of ground atoms or negations of ground atoms'. In the LEADSTO language the notation $\alpha \twoheadrightarrow_{e, f, g, h} \beta$, means:

> *If state property α holds for a certain time interval with duration g, then after some delay (between e and f) state property β will hold for a certain time interval of length h.*

Here, atomic state properties can have a qualitative, logical format, such as an expression desire(d), expressing that desire d occurs, or a quantitative, numerical format such as an expression has_value(x, v) which expresses that variable x has value v.

In Section 2, first the general BDI-model is explained. In Section 3, this BDI-model is illustrated by a case study about an employee that shows task avoiding behaviour. Next, Section 4 describes how the simple model can be extended to a BDI-model of an agent that reasons about another agent's BDI-model and uses this for social manipulation. In Section 5, this two-level BDI-model is illustrated by a case study that elaborates upon the example addressed in Section 3. This case study addresses the scenario of a manager that reasons about the task avoiding behaviour of his employee, and how to prevent that behaviour. Based on this model, some simulation experiments and their results are presented in Section 6. Section 7 discusses related work, and Section 8 concludes the paper with a discussion.

2 THE BDI-MODEL

The BDI-model bases the preparation and performing of actions on beliefs, desires and intentions (e.g., [11], [14], [18]). This model shows a long tradition in the literature, going back to Aristotle's analysis of how humans (and animals) can come to actions; cf. [1], [2]. He discusses how the occurrence of certain internal (mental) state properties within the living being entail or cause the occurrence of an action in the external world. Such internal state properties are sometimes called by him 'things in the soul', for example, sensation, reason and desire:

> 'Now there are three things in the soul which control action and truth - sensation, reason, desire.' [1], Book VI, Part 2.

[1] Vrije Universiteit Amsterdam, Department of Artificial Intelligence.
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands.
URL: http://www.few.vu.nl/~{tbosse, zamemon, treur}.
E-mail: {tbosse, zamemon, treur}@few.vu.nl.

Here, sensation indicates the sensing of the environment by the agent, which leads, (in modern terms) to internal representations, called beliefs. Reason indicates the (rational) choice of an action that is reasonable to fulfil the given desire. Based on this, Aristotle introduced the following pattern to explain action (called practical syllogism):

If A has a desire D
and A has the belief that X is a (or: the best) means to achieve D
then A will do X

The BDI-model incorporates such a pattern of reasoning to explain behaviour in a refined form. Instead of a process from desire to action in one step, as an intermediate stage first an intention is generated, and from the intention the action is generated. Thus the process is refined into a two-step process. See Figure 1 for the generic structure of the BDI-model in causal-graph-like style, as often used to visualise LEADSTO specifications. Here the box indicates the borders of the agent, the circles denote state properties, and the arrows indicate dynamic properties expressing that one state property leads to (or causes) another state property. In this model, an action is performed when the subject has the intention to do this action and it has the belief that certain circumstances in the world are fulfilled such that the opportunity to do the action is there. Beliefs are created on the basis of observations. The intention to do a specific type of action is created if there is some desire D, and there is the belief that certain circumstances in the world state are there, that make it possible that performing this action will fulfil this desire (this is the kind of rationality criterion discussed above; e.g., what is called means-end analysis is covered by this). Whether or not a given action is adequate to fulfil a given desire depends on the current world state; therefore this belief may depend on other beliefs about the world state. Instantiated relations within the general BDI-model as depicted by arrows in graphical format in Figure 1 can be specified in formal LEADSTO format as follows:

$$\text{desire(D)} \wedge \text{belief(B1)} \twoheadrightarrow \text{intention(P)}$$
$$\text{intention(P)} \wedge \text{belief(B2)} \twoheadrightarrow \text{performs(P)}$$

with appropriate desire D, action P and beliefs B1, B2. Note that the beliefs used here both depend on observations, as shown in Figure 1. Furthermore, ∧ stands for the conjunction operator (and) between the atomic state properties (in the graphical format denoted by an arc connecting two (or more) arrows). Often, dynamic properties in LEADSTO are presented in *semi-formal* format, as follows:

At any point in time
if desire D is present
and the belief B1 is present
then the intention for action P will occur

At any point in time
if the intention for action P is present
and the belief B2 is present
then the action P will be performed

As a generic template, including a reference to the agent X concerned, this can be expressed by:

For any desire D, world state property Z, and action Y such that has_reason_for(X, D, Z, Y) holds:

$$\text{desire(X, D)} \wedge \text{belief(X, Z)} \twoheadrightarrow \text{intention(X, Y)}$$

For any world state property Z and action Y such that is_opportunity_for(X, Z, Y) holds:

$$\text{intention(X, Y)} \wedge \text{belief(X, Z)} \twoheadrightarrow \text{performs(X, Y)}$$

Here has_reason_for(X, D, Z, Y) is a relation that can be used to specify which state property Z is considered a reason to choose a certain intention Y for desire D. Similarly is_opportunity_for(X, Z, Y) is a relation that can be used to specify which state property Z is considered an opportunity to actually perform an intended action Y.

Assuming that beliefs are available, what remains to be generated in this model are the desires. For desires, there is no generic way (known) in which they are to be generated in the standard model. Often, in applications, generation of desires depends on domain-specific knowledge.

3 A BDI-MODEL FOR TASK AVOIDANCE

To illustrate the BDI-model described above by a specific example, the following scenario is addressed (in the domain of an organisation); notice that here no Theory of Mind is involved.

Task Avoidance Case

A manager observes that a specific employee in the majority of cases functions quite cooperatively, but shows avoidance behaviour in other cases. In these latter cases, the employee starts trying to reject the task if he believes that his agenda already was full-booked for the short term, and he believes that capable colleagues are available with not full-booked agendas. Further observation by the manager reveals the pattern that the employee shows avoidance behaviour, in particular, in cases that a task is only asked shortly before its deadline, without the possibility to anticipate on the possibility of having the task allocated. The manager deliberates about this as follows:

> *'If I know beforehand the possibility that a last-minute task will occur, I can tell him the possibility in advance, and in addition point out that I need his unique expertise for the task, in order to avoid the behaviour that he tries to avoid the task when it actually comes up.'*

Below, this example is formalised, using the BDI-model as introduced above. First, only the behaviour of the employee is addressed (in Section 5, the deliberation process of the manager is addressed as well). To this end, the example is made more precise as follows:

The *desire* to avoid a task is created after time t by the employee if the following holds at time t:

- the employee has the belief that a task is requested that has to be finished soon
- the employee has the belief that he did not hear of the possibility that the task may come at any earlier time point

The *intention* to avoid a task is generated after time t if the following holds at time t:

- the desire to avoid the task is available
- the belief that capable colleagues are available (not full booked)

The *action* to avoid the task is generated after time t if the following holds at time t:

- the intention to avoid the task is available
- the belief that the employee's own agenda is full

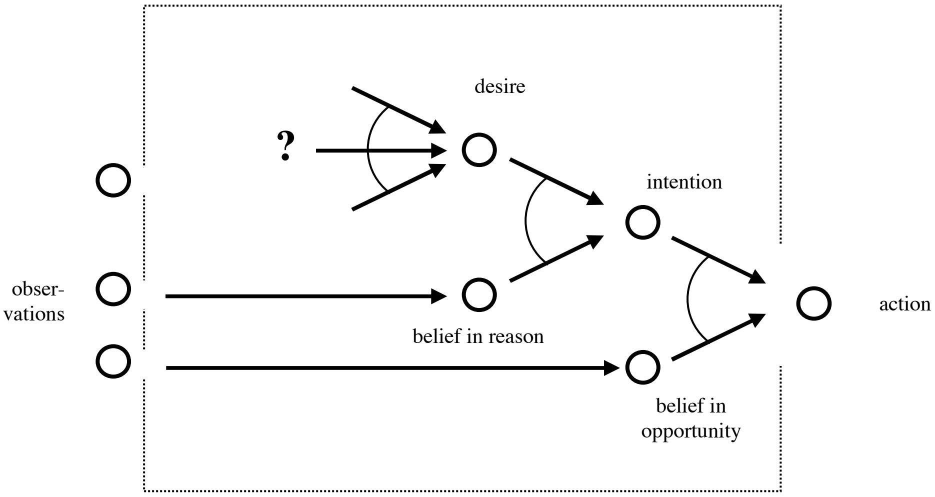


Figure 1. Structure of the general BDI-model

Using the generic template discussed at the end of Section 2, via the relations

```
has_reason_for(A, lower_workload, capable_colleagues_available,
                                                    avoid_task)
is_opportunity_for(A, own_agenda_full, avoid_task)
```

the following model for agent A is obtained:

belief(A, task_may_come) ∧ belief(last_minute_request) → desire(A, lower_workload)

desire(A, lower_workload) ∧ belief(A, capable_colleagues_available) → intention(A, avoid_task)

intention(A, avoid_task) ∧ belief(A, own_agenda_full) → performs(A, avoid_task)

4 THE TWO-LEVEL BDI-MODEL

As an instance of the *instrumentalist perspective* and opposed to explanations from a direct physical perspective (the physical stance), in [9], [10], the *intentional stance* (or folk-psychological stance) is put forward. In [10], Dennett explains the advantage of intentional stance explanations for mental phenomena over physical stance explanations:

> 'Predicting that someone will duck if you throw a brick at him is easy from the folk-psychological stance; it is and will always be intractable if you have to trace the photons from brick to eyeball, the neurotransmitters from optic nerve to motor nerve, and so forth.' [10], p.42

According to the intentional stance, an agent is assumed to decide to act and communicate based on intentional notions such as beliefs about its environment and its desires and intentions. These decisions, and the intentional notions by which they can be explained and predicted, generally depend on circumstances in the environment, and, in particular, on the information on these circumstances just acquired by interaction (i.e., by observation and communication), but also on information acquired by interaction in the past. To be able to analyse the occurrence of intentional notions in the behaviour of an observed agent, the observable behavioural patterns over time form a basis; cf. [10].

In the model presented in this paper, the instrumentalist perspective is taken as a point of departure for a Theory of Mind. More specifically, the model describes the reasoning process of an agent B that applies the intentional stance to another agent A by attributing beliefs, desires en intentions. Thus, for agent B a Theory of Mind is obtained using concepts for agent A's beliefs, desires and intentions. For example, in case a manager has an important last-minute task for his employee, but he knows that this employee often shows avoidance behaviour for last-minute tasks, he may analyse in more detail under which circumstances the employee may generate the desire and intention to avoid this task, and the related beliefs in reason and opportunity.

As a next step, the model is extended with BDI-concepts for agent B's own beliefs, desires and intentions as well. By doing this, agent B is able to not only *have* a theory about the mind of agent A, but also to *use* it within its own BDI-based reasoning processes. To this end, a number of meta-representations expressed by meta-predicates are introduced, e.g.:

belief(B, desire(A, D))

This expresses that agent B believes that agent A has desire D.

desire(B, not(intention(A, X)))

This expresses that agent B desires that agent A does not intend action X.

belief(B, depends_on(performs(A, X), intention(A, X)))

This expresses that agent B believes that, whether A will perform action X depends on whether A intends to do X. Note that the third meta-statement has a more complex structure than the other two, since it represents a statement about a *dynamic property*, rather than a statement about a *state property*. These dependencies can be read from a graph such as depicted in Figures 1 and 2 (right hand side). For example, it is assumed that agent B knows part of this graph in his Theory of Mind, expressed by beliefs such as:

belief(B, depends_on(performs(A, X), intention(A, X)))
belief(B, depends_on(performs(A, P), belief(A, B2)))
belief(B, depends_on(intention(A, P), desire(A, D)))
belief(B, depends_on(intention(A, P), belief(A, B1)))

belief(B, depends_on(desire(A, D), belief(A, B3)))
belief(B, depends_on(belief(A, X), hears(A, X)))

Desire refinement in the BDI-model for an agent B attributing motivations to an agent A is formulated (in LEADSTO format) by:

desire(B, X) ∧ belief(B, depends_on(X, Y)) ↠ desire(B, Y)
desire(B, not(X)) ∧ belief(B, depends_on(X, Y)) ↠ desire(B, not(Y))

Moreover the following schemes for intention and action generation are included in the model. For any desire D, world state property Z, and action Y such that has_reason_for(B, D, Z, Y) holds:

desire(B, D) ∧ belief(B, Z) ↠ intention(B, Y)

For any world state property Z and action Y such that is_opportunity_for(B, Z, Y) holds:

intention(B, Y) ∧ belief(B, Z) ↠ performs(B, Y)

Moreover, some dynamic properties of the world are needed:

performs(B, tell(A, C)) ↠ holds_in_world(communication(B, A, C))
holds_in_world(communication(B, A, C)) ↠ hears(A, C)

For an overview of the complete two-level BDI-model, see Fig. 2.

5 A TWO-LEVEL BDI-MODEL FOR REASONING ABOUT TASK AVOIDANCE

The above model can be used to describe how the manager agent (from the case described in Section 3) can reason and act in an anticipatory manner to avoid the employee's avoidance desire, intention and/or action to occur. The initial desire of B is that A does not perform the action to avoid the task:

desire(B, not(performs(A, avoid_task)))

Fulfilment of this desire can be obtained in the following three manners:

Avoiding A's desire to occur
This can be obtained when the employee hears in advance that possibly a last minute task may occur. This will make the second condition in A's desire generation as described in Section 3 fail.

Avoiding A's intention to occur (given that the desire occurs)
This can be obtained by refutation of the belief that plays the role of the reason to generate the intention in A's intention generation as described in Section 3, e.g., when the employee hears that colleagues do not have the required expertise.

Avoiding A's action to occur (given that the intention occurs)
This can be obtained by refutation of the belief that plays the role of opportunity in A's desire action as described in Section 3, e.g., when his agenda is not full-booked.

For convenience, the model does not make a selection but addresses all three options to prevent the avoidance action. This means that B generates desires for:

- A hears about the possibility of a last-minute task in advance

  hears(A, task_may_come)

- A hears that no colleagues that are capable of performing the task are available

  hears(A, not(capable_colleagues_available))

- A hears that his agenda is not full-booked

  hears(A, not(own_agenda_full))

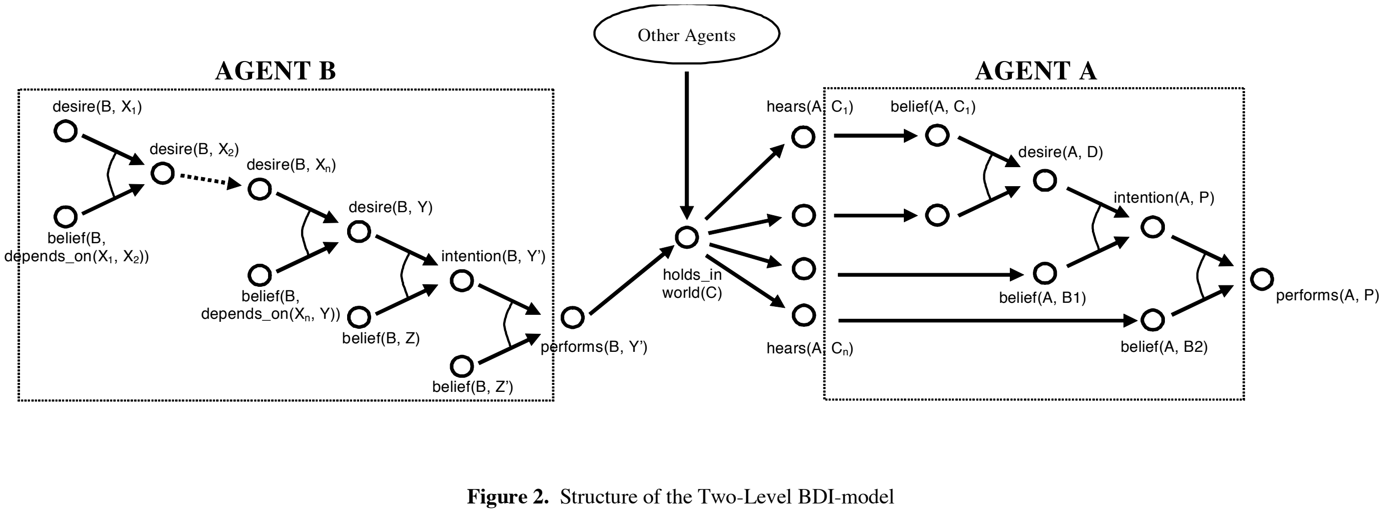


Figure 2. Structure of the Two-Level BDI-model

To fulfil these desires, intentions are to be generated by B to perform actions such as:

- B tells A about the possibility of a last-minute task in advance

  performs(B, tell(A, task_will_come))

- B tells A that no colleagues that are capable of performing the task are available

  performs(B, tell(A, not(capable_colleagues_available))

- B tells A that some of the (perhaps less interesting) tasks were taken from A's agenda and were re-allocated to a colleague

  performs(B, tell(A, not(own_agenda_full))

Reason for B to choose for these actions is

- the belief of B that telling something will lead to the person hearing it
  belief(B, adequate_communication(B, A))

Moreover, these intentions of B can lead to the corresponding actions when the following belief of B in opportunity is there:

- the belief that A is available for B to talk to
  belief(B, available_for(A, B))

In addition to the generic BDI-model shown in Section 4, the following specific relations are used to model the case study:

has_reason_for(B, hears(A, C), adequate_communication, tell(A, C))
is_opportunity_for(B, available_for(A, B), tell(A, C))

Note that the last minute request itself is an event that not necessarily comes from agent B; it can come from any agent, for example a Director agent. It is modelled as an event in LEADSTO.

6 SIMULATION EXPERIMENTS

In a number of simulation experiments, the two-level BDI-model has been applied to the case study as described in Section 5. To this end, the LEADSTO software environment [7] has been used. In Figure 3 and 4, examples of resulting simulation traces are shown. In these figures, time is on the horizontal axis; the state properties are on the vertical axis. A box on top of a line indicates that a state property is true. Note that, due to space limitations, only a selection of the relevant atoms is shown. Figure 3 is the resulting simulation trace of the situation explained in Section 3 in which *no* Theory of Mind is involved, i.e., only the behavior of employee is addressed, without social manipulation. The trace depicts that the employee initially receives some inputs (e.g., indicated by the state property

hears(employee, capable_colleagues_available)

at time point 1). As a result, the employee has made some beliefs (e.g., the state property

belief(employee, capable_colleagues_available)

at time point 2), which persists for a longer time. Next, when the employee receives a last minute request at time point 6

hears(employee, last_minute_request)

he eventually generates desire to avoid the task at time point 8

desire(employee, avoid_task)

Based on this desire and the input received earlier

hears(employee, capable_colleagues_available)

the employee generates the intention to avoid the task at time point 9:

intention(employee, avoid_task)

Based on this intention and the input received earlier

hears(employee, own_agenda_full)

at time point 1, the employee eventually performs the action of avoiding the task at time point 10.

Figure 4 is the resulting simulation trace of the case study described in Section 5, in which the manager agent can reason and act in an anticipatory manner to avoid the employee's avoidance desire, intention and/or action to occur. Figure 4 shows that the manager initially desires that the employee does not perform the action to avoid the task:

desire(manager, not(performs(employee, avoid_task)))

Based on this, he eventually generates number of more detailed desires about what the employee should hear (see, for example, the state property

desire(manager, not(hears(employee, capable_colleagues_ available)))

at time point 3). Next, the manager uses these desires to generate some intentions to fulfil these desires (e.g., the state property

intention(manager, tell(employee, not(capable_colleagues_ available)))

at time point 4). Eventually, these intentions are performed, and the employee receives some new inputs (e.g., the state property

hears(employee, not(capable_colleagues_available))

at time point 7). As a result, when the employee receives a last minute request at time point 11

hears(employee, last_minute_request)

he does not generate the action to avoid the task.

Figure 3. Simulation trace of task avoidance behaviour (without ToM)

Figure 4. Simulation trace where task avoidance behaviour is prevented by social manipulation

Note that in the scenario sketched in Figure 4, the manager takes all possible actions (within the given conceptualisation) to fulful its desires. This is a rather extreme case, since according to the employee's BDI-model, modifying only one of its input will be sufficient to make sure that (s)he does not avoid the task. Other traces can be generated in which the manager takes less actions to fulfil its desires.

7 RELATION TO OTHER WORK

The two-level BDI-Agent Model for Theory of Mind presented here is an example of recursive modeling and its use in social manipulation; see also [6], [8], [12], [17]. In the field of Agent Theory, the idea of recursive modeling has been described in [12] as follows:

> 'Recursive modeling method views a multiagent situation from the perspective of an agent that is individually trying to decide what physical and/or communicative actions it should take right now. [...] In order to solve its own decision-making situation, the agent needs an idea of what the other agents are likely to do. The fact that other agents could also be modeling others, including the original agent, leads to a recursive nesting of models.' [12], p.125

In [17], PsychSim - an implemented multi-agent based simulation tool for modeling interactions and influence among groups or individuals - has been described in the context of childhood bullying and aggression, which provides interesting insight into the role that Theory of Mind plays in human behavior. In this work, an agent's theory of mind about other agents is crucial in the following sense:

> 'For example, a bully motivated by the approval of his classmates would use his mental model of them to predict whether they would enjoy his act of aggression and laugh along with him. Similarly the bully would use his mental model of the teacher to predict whether he will be punished or not' [17], p. 247

In PsychSim, agents maintain models of each other's beliefs, goals, policies, etc., and are able to reason about it. This is a form of recursive agent modeling specifically organised to model psychological factors that play a role in influencing human communication and human social interaction in general.

The two-level BDI-Agent model for Theory of mind for social manipulation presented in this paper can be considered a recursive model with two levels of nesting. At level 1, the manager uses BDI-concepts *within* the Theory of Mind to describe the reasoning process of the employee. At level 2, the manager uses BDI-concepts for reasoning *about* the Theory of Mind for its own reasoning (meta-reasoning) about the reasoning process of the employee.

The work by [8] considers many topics, like foundation of sociality (cooperation, competition, groups, organization, etc), levels of coordination and cooperation, emergent pre-cognitive structures and constraints. Specifically it addresses influencing other agents and trying to change their behavior based on Theory of Mind of the agent:

> 'The explicit representation of the agents mind in terms of beliefs, intentions, etc., allows for reasoning about them, and – even more importantly – it allows for the explicit influencing of others, trying to change their behavior (via changing their goals/beliefs).[...] The agents should have some decision function (that implicitly or explicitly presupposes some goal/desire/preference). The influencing agent should give them some hints for this decision, in order to change their behavior' [8], p. 178

However, in that work no formalisation is presented. In contrast, the model presented here has been formally specified.

8 DISCUSSION

In order to function efficiently in social life, it is very helpful for an agent to have capabilities to predict in which circumstances the agents in its environment will show certain behaviours. To this end, such an agent will have to perform reasoning based on a Theory of Mind [3]. This paper presents a model for reasoning based on a Theory of Mind, which makes use of BDI-concepts at two different levels. First, the model uses BDI-concepts *within* the Theory of Mind (i.e., it makes use of beliefs, desires and intentions to describe the reasoning process of another agent). Second, it uses BDI-concepts for reasoning *about* the Theory of Mind (i.e., it makes use of beliefs, desires and intentions to describe an agent's meta-reasoning about the reasoning process of another agent). At this second level, meta-statements are involved, such as 'B believes that A desires d' or 'B desires that A does not intend a'. These meta-statements are about the states occurring within the other agent. In addition, meta-statements are involved about the dynamics occurring within the other agents. An example of such a (more complex) meta-statement is 'B believes that, if A performs a, then earlier he or she intended a'.

The two-level BDI-based model as presented can be exploited both for *social anticipation* (i.e., in order to be prepared for the behaviour of another agent) and for *social manipulation* (i.e., in order to affect the behaviour of another agent at forehand). The model has been formalised using the high-level modelling language LEADSTO, which describes dynamics in terms of direct temporal dependencies between state properties in successive states. Based on the formal model, a number of simulation experiments have been performed within a specific case study, addressing the scenario of a manager that reasons about the task avoiding behaviour of his employee. A main difference with the earlier work described in [15] is that in the current paper the agent model is executable and therefore can easily be used for simulation. Moreover, it not only addresses reasoning about the other agent's beliefs, desires and intentions, but also integrates this with reasoning about the agent's own beliefs, desires and intentions, and actions in order to perform social manipulation. This part was not formalised in [15].

The case study illustrates how the two-level model can be used for social manipulation. For this purpose, the crucial steps are to find out which situations would lead to undesired behaviour of another agent, and to prevent these situations from occurring (or, similarly, to establish situations that would lead to desired behaviour). In addition, the model can be used for social anticipation. In that case, the main steps are to predict the behaviour that another agent will show given the current situation, and to prepare for this. Also this second type of reasoning based on a Theory of Mind is essential to function smoothly in social life.

In the model, not only Agent B has a Theory of Mind about Agent A, but it also uses this theory in its own reasoning process in order to do social manipulation, i.e., to change the behavior of Agent A. The model was designed in such a way that the Agent A does not know beforehand that Agent B is trying to manipulate him by changing some of the beliefs. Thus, the situation was not considered that Agent A tries to not to be manipulated by Agent B. In future research it will be addressed how such elements can be included in the model. For instance, the employee may have models of other employees to infer who else is available for a task.

In [8], a basic problem of social life among cognitive agents has been addressed, that is: *how to induce the other agent to believe on us and even to do something*, or in other words: *why should he care about our goals and expectations?* This problem can be solved normally – but not necessarily – by communicating. Generally, in order to induce another agent to do or not to do something, we need power of influencing him. However, the most important basis of our power is the fact that also our actions are potentially interfering with his goals. This can be exploited to change his mind and induce him to do or not to do something. In the case study considered here, as the manager has power over his employee, he interfered in the goal of employee: by performing the action of changing his beliefs by communicating to him. As a result, the behavior of the employee has been changed, i.e., he doesn't avoid the task.

From a theoretical angle, much literature is available in foundations of approaches as the one presented here. For example in literature such as [13], [18], [19], a modal logic perspective is used to obtain formal semantics for languages that allow to express that an agent has reflective knowledge about what another agent knows. However, most of such modal logic approaches do not address the dynamics of the agents' processes in an executable manner. An exception is [4], where executable temporal logic is used as a basis; however, there the reflective aspect is not incorporated.

REFERENCES

[1] Aristotle (350 BC).!*Nicomachean Ethics* (translated by W.D. Ross)

[2] Aristotle (350 BC).!*De Motu Animalium* On the Motion of Animals! (translated by A. S. L. Farquharson)!!

[3] Baron-Cohen, S. (1995). *Mindblindness*. MIT Press.

[4] Barringer, H., M. Fisher, D. Gabbay, R. Owens, & M. Reynolds (1996). *The Imperative Future: Principles of Executable Temporal Logic*, Research Studies Press Ltd. and John Wiley & Sons.

[5] Bogdan, R.J. (1997). *Interpreting Minds*. MIT Press.

[6] Boella, G. and van der Torre, L. (2004) Groups as agents with mental attitudes. In: Jennings, N.R., Sierra, C., Sonenberg, L., and Tambe, M. (eds.), *Proceedings of the third international joint conference on Autonomous Agents and Multi Agent Systems, AAMAS'04*, pp. 964-971.

[7] Bosse, T., Jonker, C.M., Meij, L. van der, and Treur, J. (2007). A Language and Environment for Analysis of Dynamics by Simulation. *International Journal of Artificial Intelligence Tools*. To appear, 2007. Shorter, earlier version in: Eymann, T., Kluegl, F., Lamersdorf, W., Klusch, M., and Huhns, M.N. (eds.), *Proc. of the Third German Conference on Multi-Agent System Technologies, MATES'05*. LNAI, vol. 3550. Springer Verlag, 2005, pp. 165-178.

[8] Castelfranchi C. *Modelling Social Action for AI Agents*. Artificial Intelligence, vol. 103, 1998, pp. 157-182

[9] Dennett, D.C. (1987). *The Intentional Stance*. MIT Press. Cambridge Mass.

[10] Dennett, D.C. (1991). Real Patterns. *The Journal of Philosophy*, vol. 88, pp. 27-51.

[11] Georgeff, M. P., and Lansky, A. L. (1987). Reactive Reasoning and Planning. In: Forbus, K. and Shrobe, H. (eds.), *Proceedings of the Sixth National Conference on Artificial Intelligence*, AAAI'87. Menlo Park, California. American Association for Artificial Intelligence, 1987, pp. 677-682.

[12] Gmytrasiewicz P. J. and Durfee. E. H. A rigorous, operational formalization of recursive modeling. In: Lesser, V. (ed.), *Proceedings of the First International Conference on Multiagent Systems*, pp. 125-132, 1995.

[13] Halpern, J.Y., Fagin, R., Moses, Y., and Vardi, M. Y. (1995). *Reasoning About Knowledge*. MIT Press, 1995.

[14] Jonker, C.M., Treur, J., and Wijngaards, W.C.A., (2003). A Temporal Modelling Environment for Internally Grounded Beliefs, Desires and Intentions. *Cognitive Systems Research Journal*, vol. 4(3), 2003, pp. 191-210.

[15] Jonker, C.M., Treur, J., and Vries, W. de, (2002). Temporal Analysis of the Dynamics of Beliefs, Desires, and Intentions. *Cognitive Science Quarterly*, vol. 2, 2002, pp.471-494.

[16] Malle, B.F., Moses, L.J., Baldwin, D.A. (2001). *Intentions and Intentionality: Foundations of Social Cognition*. MIT Press.

[17] Marsella, S.C., Pynadath, D.V., and Read, S.J. *PsychSim: Agent-based modeling of social interaction and influence*. In: Lovett, M., Schunn, C.D., Lebiere, C., and Munro, P. (eds.), *Proceedings of the International Conference on Cognitive Modeling, ICCM 2004*, pp. 243-248 Pittsburg, Pensylvania, USA.

[18] Rao, A.S. and Georgeff, M.P. (1995) BDI-agents: from theory to practice. In: Lesser, V. (ed.), *Proceedings of the International Conference on Multiagent Systems,* pp. 312 – 319.

[19] Rao, A.S. and Georgeff, M.P. (1991). Modelling Rational Agents within a BDI-architecture. In: Allen, J., Fikes, R. and Sandewall, E. (eds.), *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, *(KR'91)*. Morgan Kaufmann, pp. 473-484.

Intention Recognition with Divergent Beliefs for Collaborative Robots

Jesse Gray and **Matt Berlin** and **Cynthia Breazeal**[1]

Abstract.

Robotic systems that aim to collaborate effectively with humans in social environments must be able to respond flexibly to the intentions of their human partners. Dynamic environments may further require robots to respond intelligently to the actions of humans with false or incomplete situational beliefs. We present an integrated architecture which incorporates simulation-theoretic mechanisms to allow a robot to infer the task-related beliefs and intentions of its interaction partners based on their observable motor behavior and visual perspective. The performance of this architecture is evaluated on a set of novel benchmark tasks requiring a robotic system to demonstrate appropriate collaborative behaviors in the presence of potentially false beliefs. Results are compared against human performance on a similar task suite.

1 SELF-AS-SIMULATOR COGNITIVE ARCHITECTURE

Our approach to endowing machines with socially-cognitive learning abilities is inspired by leading psychological theories and recent neuroscientific evidence for how human brains might infer the mental states of others and the role of imitation as a critical precursor. Specifically, *Simulation Theory* holds that certain parts of the brain have dual use; they are used to not only generate our own behavior and mental states, but also to predict and infer the same in others. To understand another person's mental process, we use our own similar brain structure to simulate the introceptive states of the other person [1, 5–7].

For instance, Gallese and Goldman [6] propose that a class of neurons discovered in monkeys, labeled mirror neurons, are a possible neurological mechanism underlying both imitative abilities and Simulation Theory-type prediction of the behavior of others and their mental states. Further, Meltzoff and Decety [10] posit that imitation is the critical link in the story that connects the function of mirror neurons to the development of mindreading. In addition, Barsalou [1] presents additional evidence from various social embodiment phenomena that when observing an action, people activate some part of their own representation of that action as well as other cognitive states that relate to that action.

Inspired by this theory, our simulation-theoretic approach and implementation enables a humanoid robot to monitor an adjacent human collaborator by simulating his or her behavior within the robot's own generative mechanisms on the motor, goal-directed action, and perceptual-belief levels. This grounds the robot's information about the human in the robot's own systems, allowing it to make inferences about the human's likely beliefs in order to better understand the intention behind the human's actions. Our architecture, which extends [2] and [8], is designed to run on the 65 degree of freedom humanoid robot Leonardo and its graphical simulator (Fig. 1).

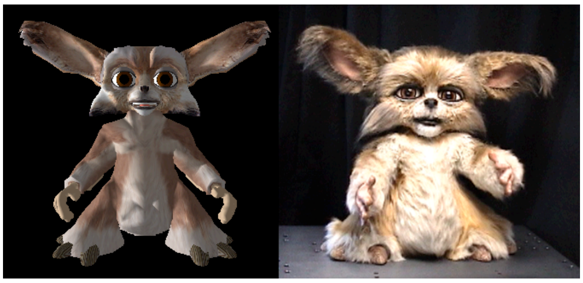


Figure 1. The Leonardo robot and graphical simulator.

Our implementation computationally models simulation-theoretic mechanisms throughout several systems within the robot's overall cognitive architecture. See Figure 2 for a system diagram. For instance, within the motor system, mirror-neuron inspired mechanisms are used to map and represent perceived body positions of another into the robot's own joint space to conduct action recognition. Leo reuses his belief-construction systems, and adopts the visual perspective of the human, to predict the beliefs the human is likely to hold to be true given what he or she can visually observe. Finally, within the goal-directed behavior system, where schemas relate preconditions and actions with desired outcomes and are organized to represent hierarchical tasks, motor information is used along with perceptual and other contextual clues (i.e., task knowledge) to infer the human's goals and how he or she might be trying to achieve them (i.e., plan recognition).

1.1 Perspective taking mechanisms

We turn now to the robot's visual perspective taking mechanisms. While others have identified that visual perspective taking coupled with spatial reasoning are critical for effective action recognition [9] and human-robot collaboration on a shared task within a physical space [11], and collaborative dialog systems have investigated the role of plan recognition in identifying and resolving misconceptions (see [4] for a review), our work is novel in its emphasis on simulation-theoretic mechanisms for inferring introceptive states (e.g., beliefs and goals) in human-robot collaboration.

When collaborating on a shared task, it is important for all parties involved to have a consistent representation of the task context. However, in complex and dynamic environments, it is possible for

[1] MIT Media Lab, U.S.A., email: {jg, mattb, cynthiab}@media.mit.edu

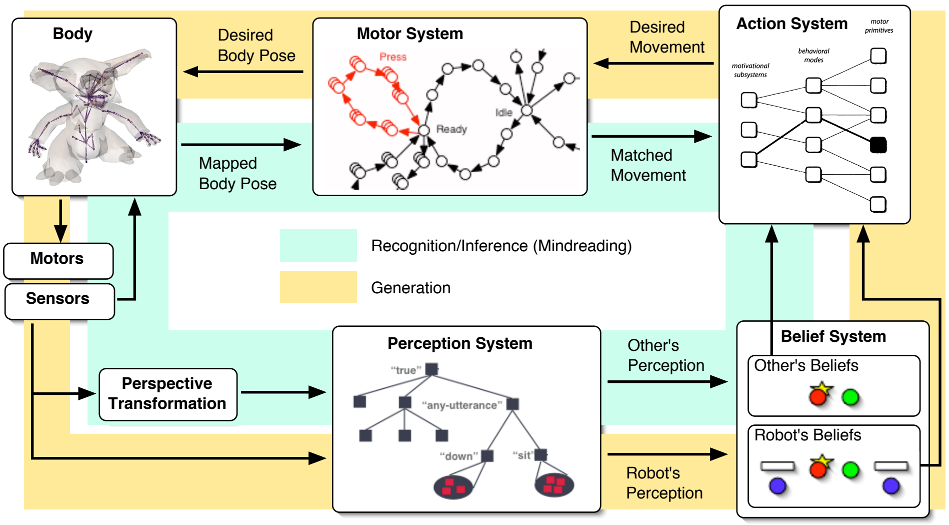


Figure 2. System architecture overview.

one collaborator's beliefs about the context surrounding the activity to diverge from those of other collaborators. For example, a visual occlusion could block one's viewpoint of a region of a shared workspace (but not that of another) and consequently lead to ambiguous behavior where one collaborator does not realize that the visual information of the scene differs between the participants.

To address this issue, the robot must establish and maintain mutual beliefs with its human partners about the shared context surrounding collaboration. The robot keeps track of its own beliefs about object state using its Belief System, described in detail in [3]. In order to model the beliefs of a human partner as separate and potentially different from its own, the robot re-uses the mechanism of its own Belief System. These beliefs that represent the robot's model of the human's beliefs are in the same format as its own, but are maintained separately so the robot can compare differences between its beliefs and the human's beliefs.

Belief maintenance consists of incorporating new sensor data into existing knowledge of the world. The robot's sensors are all in its reference frame, so objects in the world are perceived relative to the robot's position and orientation. In order to model the beliefs of the human, the robot re-uses the same mechanisms used for its own belief modeling, but first transforms the data into the reference frame of the human (see Fig. 3).

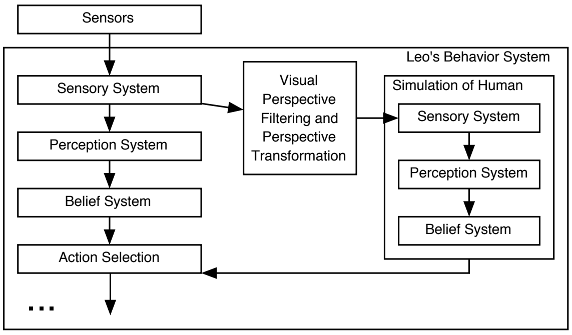


Figure 3. Architecture for modeling the human's beliefs re-uses the robot's own architecture for belief maintenance.

The robot can also filter out incoming data that it believes is not perceivable to the human, thereby preventing that new data from updating the model of the human's beliefs. If the inputs to the robot's perceptual-belief pipeline are the sensory observations $O = \{o_1, o_2, ..., o_N\}$, then the inputs to the secondary pipeline that models the human's beliefs are O', where:

$$O' = \{P(o')|o' \in O, V(o') = 1\} \tag{1}$$

where:

$$V(x) = \begin{cases} 1 & \text{if } x \text{ is visible to human} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

and:

$$\begin{aligned} P :& \{\text{robot local observations}\} \\ & \quad \rightarrow \{\text{person local observations}\} \end{aligned} \tag{3}$$

Visibility can be determined by a cone calculated from the human's position and orientation, and objects on the opposite side of known occlusions from the human can be marked invisible.

Maintaining this parallel set of beliefs is different from simply adding metadata to the robot's original beliefs because it reuses the entire architecture which has mechanisms for object permanence, history of properties, etc. This allows for a more sophisticated model of the human's beliefs. For instance, Fig. 4 shows an example where this approach keeps track of the human's incorrect beliefs about objects that have changed state while out of the human's view. This is important for establishing and maintaining mutual beliefs in time-varying situations where beliefs of individuals can diverge over time.

1.2 Goal inference

The concept of self as simulator is also applied to goal inference through dual use of the robot's action production systems to not only

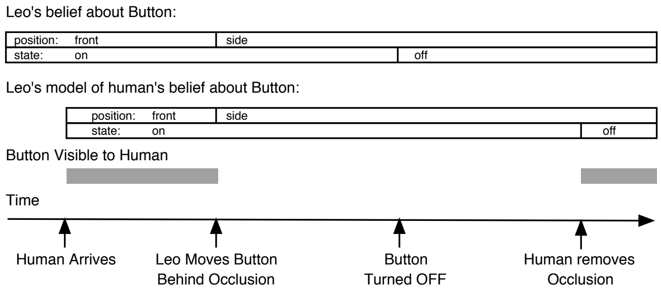


Figure 4. Timeline following the progress of the robot's beliefs for one button. The robot updates its belief about the button with any sensor data available - however, the robot only integrates new data into its model of the human's belief if the data is available when the human is able to perceive it.

direct the robot's actions but to recognize the goals of others. To accomplish this, we reuse three separate levels of abstraction: body position, movements through many body positions, and schemas made of movements, contexts, and goals.

Leonardo has the capacity to physically imitate the position of humans he is observing. This capacity is learned through an imitative interaction, where a human first imitates Leo and Leo can then save the correlation between observed human positions and his own motor coordinates. This data can be used to produce a reverse mapping, so Leo can then imitate the human or understand their pose in terms of his own motor coordinates.

Once Leo can map a single body position from the human into his own motor space, he can match time sequences of poses against his own motions. If he concludes that a human's motion is similar to one he can perform, Leo can represent the human's activity in this compact format useful for further inference.

Leo has a particular schema representation to control his actions based on his goals. This representation is also flexible enough to be used in reverse to simulate the behavior of a human partner, determining their goals from their observed actions. This technique not only allows re-use of schemas for goal inference, but it also ensures that any observed behavior is in a format immediately useful to the robot - namely goals that are expressed in its own network of schemas.

Within the deliberative system of the robot, the atomic-level representation of a goal-directed behavior is a schema that associates its necessary perceptual preconditions with a specific action (optionally performed on a particular object, or with other parameters) to achieve an expected outcome (its goal). Schemas can be organized sequentially and/or hierarchically to create larger structures to represent tasks and execute them. When chaining sequential schemas, the goal of one schema becomes the precondition of the subsequent schema. Compound tasks are specified as a hierarchy of schemas, where the expected result of multiple schemas are the inputs (i.e., listed in the preconditions) of the subsequent schema. To achieve some desired task goal, only the relevant schema need be activated and all necessary preconditions will be fulfilled. Each of the condition elements that separates the actions has the capacity to map parameters in either direction (Fig. 5). This is what allows bidirectional activation of the network. By activating a goal, appropriate pre-conditions will occur ("Open Box 1" will be preceded by "Unlock Lock 1" based on the parameter mapping of the "Box Unlocked" condition); similarly, from a precondition bound to a target, possible goal targets can be calculated. This mapping process takes place in the "context" of a particular set of beliefs about the world - that is, the mapping uses current knowledge about the world to determine how objects may relate.

When simulating the goal-directed behavior of others, the deliberative system begins by first attempting to determine which of its schemas might match the person's current contextual situation and action. Once the robot classifies an observed motion as matching one that it can perform, it searches its schemas for any that evoke that same motion. If multiple schemas involve the same motion, the set of candidate schemas is narrowed by matching the current human's perceptual context against the necessary context for that schema. Note that the perceptual context must be based on the the model the robot is maintaining of the human's beliefs about the situation, not on the robot's. For example, in Figure 5, the robot's and human's beliefs disagree on the purpose of "Key1". If the robot is to determine the goal of a human holding "Key1", it must use the human's beliefs, even if they are incorrect.

Once a schema has been selected as a suitable match, the robot infers that the human's immediate goal is the expected result of that schema. It can also anticipate less immediate goals by following the schema chain upwards while computing the targets for each action in the context of the human's beliefs. Once it has inferred the goal of the human, the robot can be helpful in a number of situations. If the human fails an action (the robot notices the action being performed in an appropriate context, but the goal is not achieved) then the robot can complete the action for the human. When the beliefs of the human and the robot differ, the robot may be able to provide even more useful assistance. Because of the differing beliefs, the schema networks for the human's understanding of the situation will differ from the robot's, allowing opportunities for the robot to suggest or

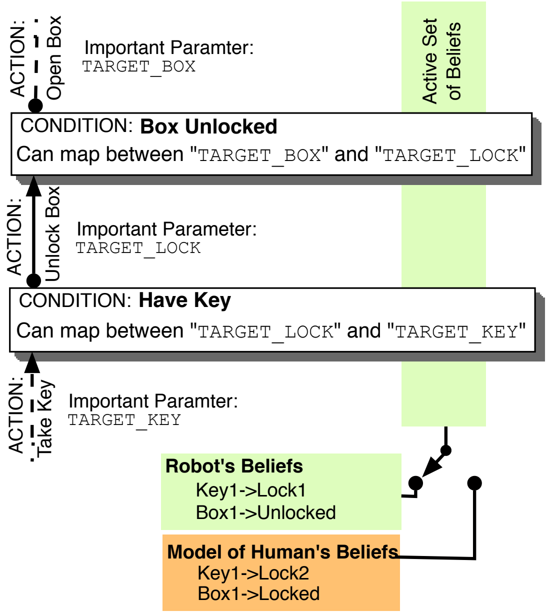


Figure 5. Section of action schema. Schemas can be traversed in either direction, upwards to infer goals and downwards to find ways to achieve those goals. Parameters for the actions are converted by the Conditions, using data in the model of the human's beliefs (for goal inference) or the robot's own beliefs (for task accomplishment)

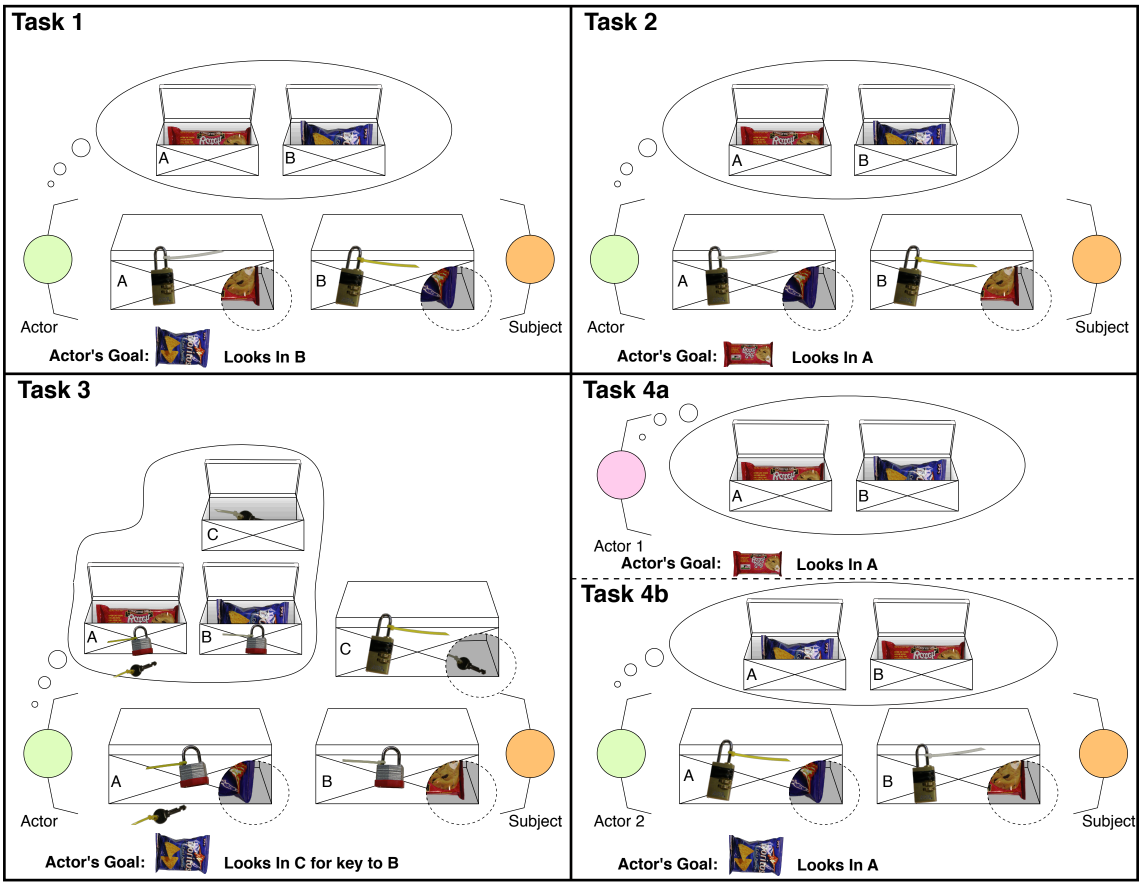


Figure 6. The four collaborative benchmark tasks: (1) simple goal inference, (2) goal inference with false beliefs, (3) goal inference with false beliefs and indirect, dislocated action, and (4) goal inference with multiple agents and false beliefs. Shown are the actual world state and the actor's "belief" state at the moment when the subject's behavior is classified.

perform other actions. The robot searches for these helpful behaviors by first moving up the schema network (in the context of the human's beliefs) to the goal of the human. From there, it can search back down (using its own belief context) for the shortest currently possible chain that can achieve the goal. If that chain starts with an action which differs from the human's current action, the robot can attempt this action or point it out to the human. If that chain depends on knowledge that the human doesn't have, the robot can also try to point out that information to the human.

For example, with the beliefs listed in Fig. 5, the robot might witness the human taking a key to Box1. It would infer that the human wants to unlock Box1 in order to open it. Searching back down from that goal in the context of its own beliefs, however, the robot would discover that it is possible to open Box1 immediately, since it is unlocked. It could then suggest the action or perform it itself.

2 BENCHMARK TASKS

In order to evaluate our cognitive architecture, we have developed a novel set of benchmark tasks that examines the use of belief reasoning and goal inference by humans and synthetic agents in a collaborative setting. In this section, we present the details of this benchmark suite and a task-by-task comparison of human and robotic performance data.

Our benchmark tasks are variants of the classic false belief task from developmental psychology [12]. In the classic task, subjects are told a story with pictorial aides that typically proceeds as follows: two children, Sally and Anne, are playing together in a room. Sally places a toy in one of two containers. Sally then leaves the room, and while she is gone, Anne moves the toy into the other container. Sally returns, and the subject is asked: where will Sally look for the toy?

Our benchmark tasks embed the false belief task within a live, collaborative setting. Participants interact face-to-face with a collaborative partner, and are prompted to assist their partner in any way they see fit. Instead of evaluating the participant with an explicit prompt (e.g. "where will your partner look for the cookies?"), our analysis focuses on the participant's implicit, non-linguistic reasoning: we observe the participant's behavior as they attempt to assist their partner. Thus we are attempting to examine the spontaneous use of goal inference and false belief reasoning in collaborative activity.

A schematic of the four benchmark tasks is shown in Figure 6. In each task, the participant (Subject) interacts with a collaborative partner (Actor) who is an experimental confederate. The participant has access to a collection of food objects that are identical to hidden target objects that their partner may be searching for. It is thus possible for the participant to assist their partner by handing them an object which matches the target of their search.

Task 1 is a control task examining simple goal inference. The Subject and Actor both watch as the experimenter hides a package of cookies in Box A and a bag of chips in Box B. The experimenter then seals both boxes. The Actor receives instructions written on a notecard to deliver a bag of chips to the experimenter. The Actor proceeds to attempt to open Box B, and the Subject's subsequent behavior is recorded. In order to successfully assist the Actor, the Subject must infer that because the Actor is attempting to open Box B, the Actor's goal is to acquire the chips contained within the box.

Task 2 examines goal inference with false beliefs. The setup proceeds as in Task 1, with Subject and Actor both observing cookies hidden in Box A and chips hidden in Box B. After the boxes are sealed, the Actor is asked to leave the room, at which point the experimenter swaps the contents of the boxes. The Actor returns, receives instructions, and attempts to open Box A. In order to successfully assist the Actor, the Subject must infer that the Actor's goal is to acquire the cookies, even though Box A currently contains the chips.

Task 3 examines goal inference with false beliefs and indirect, dislocated action. The setup proceeds as in Task 2, however, in this case, the experimenter locks both Box A and Box B with color-coded padlocks. The key to Box A is left in plain view, but the key to Box B is sealed inside of a third box, Box C. The Actor is then asked to leave the room, at which point the experimenter, using a master key, swaps the contents of Box A and Box B, leaving both boxes locked. The Actor returns, receives instructions, and attempts to open Box C. In order to successfully assist the Actor, the Subject must infer that the Actor's goal is to acquire the chips, even though the immediate target of the Actor's actions, Box C, contains neither the chips nor even the key to a box containing chips.

The final task, Task 4, examines goal inference with multiple agents and false beliefs. In this task, the Subject is introduced to two collaborative partners, Actor 1 and Actor 2. All three watch as the experimenter hides cookies in Box A and chips in Box B, and then seals both boxes. Actor 1 is then asked to leave the room, at which point the experimenter swaps the contents of Box A and Box B in view of both the Subject and Actor 2. Actor 2 is then asked to leave, and Actor 1 returns. Actor 1 receives instructions and attempts to open Box A. The Subject's subsequent behavior is recorded (Task 4a). Finally, Actor 1 leaves, and Actor 2 returns, receives instructions, and also attempts to open Box A. The Subject's behavior is recorded (Task 4b). In order to successfully assist both actors, the Subject must keep track of Actor 1's false beliefs about the object locations as well as Actor 2's correct beliefs about these locations.

2.1 Human Subjects Study

We conducted a human subjects study to gather human performance data on our collaborative benchmark tasks. Figure 7 shows some of the essential elements of our study setup. Target objects were hidden in three flight cases (A), (B), and (C). Our experimental confederate and the study participant were seated opposite each other at locations (D) and (E), respectively. The participant's stock of food objects was located on a stool, (F), adjacent to their chair and out of the reach and view of the confederate. The target objects, (H), were a bright red package of chocolate-chip cookies and a bright blue bag of corn chips. Also shown are the viewpoint from the participant's location, (I), and the viewpoint from the confederate's location, (J) - note that the stock of food objects is not visible from this location.

A detail of our box-sealing mechanism is shown in (G). In all of the cases which called for boxes to be sealed, we sealed them with color-coded combination locks similar to the one depicted. Note that two of the lock's four numeric dials have been covered up and fixed in place by electrical tape, leaving only two dials free for manipulation. This lock mechanism served an important timing function in our study, introducing a delay in the Actor's process of opening any sealed box. When attempting to open a sealed box, the Actor, who secretly knew the correct combination, would make a show of attempting to "hack" the lock by trying all 99 combinations. The Actor would start at zero and slowly increment the combination, tugging at the lock with each iteration. The correct code was always 21, although the experimenter pretended to change the codes between tasks. This meant that the Actor could successfully "hack" the lock within 30 to 45 seconds if necessary, giving the Subject sufficient time to consider the Actor's goal and contemplate potential helpful actions, while keeping the experiment running at a reasonable pace.

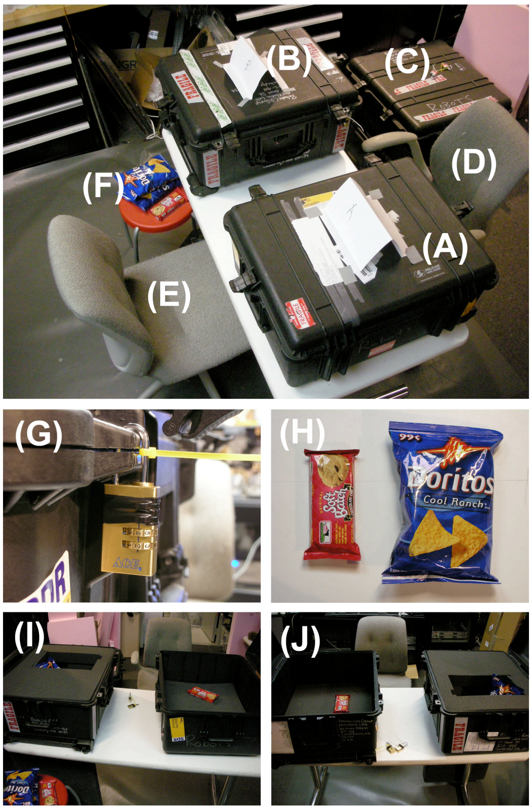


Figure 7. Setup of the human subjects study. (A,B,C) Boxes in which target objects were hidden. (D) Confederate's chair. (E) Participant's chair. (F) Objects available to participant. (G) Detail of box with combination lock. (H) Target objects. (I) Participant's viewpoint. (J) Confederate's viewpoint.

We gathered data from 20 participants, who were each presented with the four benchmark tasks in randomized order. Participants were instructed not to talk to their partner, but were told that they were otherwise free to perform any action or gesture that might help their partner achieve the goal. Participants were instructed that they might find the objects on the stool next to their chair useful, but that they could only use one of these objects per task.

The results of the study are summarized in Table 1. Participant behavior was partitioned into six categories, from most helpful to least helpful: correct object presented, guidance gesture presented, grounding gesture presented, other, no action, incorrect object presented. Behavior was classified as follows. If the participant presented the correct target object to their partner, they were tallied as "correct," and if they presented the wrong object, they were tallied

Table 1. Behavior demonstrated by study participants on benchmark tasks.

Task	Correct Object	Guidance Gesture	Grounding Gesture	Other	No Action	Incorrect Object
Task 1	16	0	0	1†	1	2
Task 2	14	1	2	0	0	3
Task 3	13⋆	5	2	0	0	0
Task 4a	14	2	1	0	3	0
Task 4b	13	0	1	1‡	1	4

⋆ One participant produced the object only after the key had been retrieved from box C.

† Participant successfully pried open the locked target box.

‡ Participant discovered the combination lock code and revealed it gesturally.

as "incorrect." Participants who did not present either object were classified according to the gestures that they displayed. "Guidance" gestures included only direct pointing or manipulation towards the correct target box, lock, or key. "Grounding" gestures included bidirectional pointing gestures indicating that the box contents had been swapped, as well as the use of the matching food objects as a "map" to indicate the correct contents of the various boxes. In the absence of such gestures, behavior was tallied as "no action." Finally, two unexpected cases were tallied as "other" as described in the table notes. It should be noted that in the case of Task 3, guidance gestures were almost as helpful as producing the correct object, since indicating the correct padlocked box or its readily-available key resulted in the rapid acquisition of the contents of the box.

These results indicate that participants were largely successful at inferring the goals of their collaborative partners and engaging in helpful behaviors even in the presence of false beliefs, multiple agents, and indirect goal cues. It should also be noted, however, that success was not uniform: many participants found some of the tasks to be quite challenging, and many reported difficultly in remembering the locations of the hidden objects and the divergent beliefs of their collaborative partners.

2.2 Robot Demonstration and Discussion

The collaborative benchmark tasks from our study can be used to examine the performance of the self-as-simulator cognitive architecture. Our architecture allows the Leonardo robot to track the beliefs of his human collaborators, infer their goals, and engage in helpful behaviors.

In our setup, instead of handing matching food objects to his collaborative partner, the robot indicates the locations of hidden objects by pointing, allowing the human to retrieve the specified objects. In all other respects, we follow the identical task protocol as was used in the human subjects study.

We use a ten-camera Vicon motion capture system to track the positions of reflective markers mounted to people and objects involved in the benchmark tasks with high spatial resolution. Customized tracking software allows the robot to uniquely identify rigid and near-rigid objects and track their position and orientation. This sensory apparatus is used to track the head and hand pose of the robot's human collaborators, as well as the position and extent of the boxes, box lids, and target food objects.

Table 2 displays the behavior generated by our architecture on the various benchmark tasks in two conditions: with matching target objects available to the robot, and with no objects available. With matching target objects available, the robot successfully reveals the correct target object to his collaborative partner on all benchmark tasks, matching the behavior observed in the majority of study participants in each case.

Table 2. Behavior demonstrated by robot using self-as-simulator cognitive architecture.

Task	Target Objects Available	Target Objects Unavailable
Task 1	reveals correct object	no action
Task 2	reveals correct object	points to target location
Task 3	reveals correct object	points to key
Task 4a	reveals correct object	points to target location
Task 4b	reveals correct object	no action

With no objects available, the robot can only provide gestural support to his collaborative partner. On Tasks 1 and 4b, the collaborator is attempting to open the correct box, and the robot generates no assistive behaviors. On Tasks 2 and 4a, the robot can use his knowledge of the human's beliefs to infer which object they are trying to acquire. Using this goal in conjunction with his own true knowledge of the world state allows the robot to direct the human to the correct box via a pointing gesture (see Figure 8 for an example of the network of action schemas related to Task 2). The robot uses the same inferential mechanism on Task 3 to generate a pointing gesture towards the key lying on the table which opens the correct padlocked box.

While the robot is not able to generate the full range of gestures and actions observed in our study participants, the self-as-simulator cognitive architecture nevertheless allows the robot to produce helpful behaviors on a number of sophisticated collaborative tasks requiring goal inference in the presence of potentially divergent beliefs.

3 CONCLUSION

Robotic systems that aim to collaborate effectively with humans in dynamic, social environments must be able to respond flexibly to the intentions of their human partners, even when their collaborators' actions are based on false or incomplete beliefs. Our integrated architecture incorporates simulation-theoretic mechanisms which allow a robot to infer the task-related beliefs and intentions of its interaction partners based on their observable motor behavior and visual perspective. This approach enables appropriate behavioral responses in complex collaborative scenarios involving divergent, false beliefs.

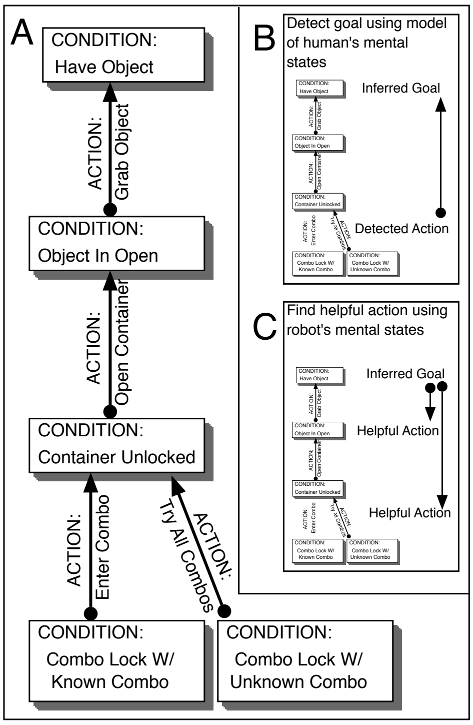

A
CONDITION: Have Object
ACTION: Grab Object
CONDITION: Object In Open
ACTION: Open Container
CONDITION: Container Unlocked
ACTION: Enter Combo
ACTION: Try All Combos
CONDITION: Combo Lock W/ Known Combo
CONDITION: Combo Lock W/ Unknown Combo
B
Detect goal using model of human's mental states
Inferred Goal
Detected Action
C
Find helpful action using robot's mental states
Inferred Goal
Helpful Action
Helpful Action

Figure 8. A) A network of action schemas related to Task 2. B) The robot detects a "Try All Combos" action on Box A. In the context of the human's beliefs, this indicates a goal of "Have Object" applied to the cookies. C) Using this goal, the robot can traverse back down the network (using its own world knowledge) to find the closest action that can lead to this goal. If the robot has access to hidden cookies, the shorter arrow indicates that the robot should draw attention to the condition unknown to the human - that cookies are ready to grab but out of sight. If the robot has no cookies, the longer arrow indicates that the robot should call attention to the action "Try all Combos" on Box B. If the human is already doing the closest possible action (as would be the case in Task 1), the robot takes no action.

ACKNOWLEDGEMENTS

The work presented in this paper is a result of ongoing efforts of the graduate and undergraduate students of the MIT Media Lab Robotic Life Group. This work is funded by the *Digital Life* and *Things That Think* consortia of the MIT Media Lab, and supported by the Office of Naval Research (YIP grant N000140510623). Jesse Gray would like to thank Samsung Electronics for their support. We would also like to thank Roman Zadov, Amy Shui, Andrew Brooks, Mikey Siegel, Guy Hoffman, Andrea L. Thomaz, Alea Teeters, and Stefanie Tellex for their assistance with our study.

REFERENCES

[1] L. W. Barsalou, P. M. Niedenthal, A. Barbey, and J. Ruppert, 'Social embodiment', *The Psychology of Learning and Motivation*, **43**, (2003).

[2] Bruce Blumberg, Marc Downie, Yuri Ivanov, Matt Berlin, Michael Patrick Johnson, and Bill Tomlinson, 'Integrated learning for interactive synthetic characters', *ACM Transactions on Graphics*, **21**(3: Proceedings of ACM SIGGRAPH 2002), (2002).

[3] C. Breazeal, M. Berlin, A. Brooks, J. Gray, and A. L. Thomaz, 'Using perspective taking to learn from ambiguous demonstrations', *Journal of Robotics and Autonomous Systems Special Issue on Robot Programming by Demonstration*, **54**(5), (2006).

[4] Sandra Carberry, 'Techniques for plan recognition', *User Modeling and User-Adapted Interaction*, **11**(1-2), 31–48, (2001).

[5] M. Davies and T. Stone, 'Introduction', in *Folk Psychology: The Theory of Mind Debate*, eds., Martin Davies and Tony Stone, Blackwell, Cambridge, (1995).

[6] V. Gallese and A. Goldman, 'Mirror neurons and the simulation theory of mind-reading', *Trends in Cognitive Sciences*, **2**(12), 493–501, (1998).

[7] R. Gordon, 'Folk psychology as simulation', *Mind and Language*, **1**, 158–171, (1986).

[8] Gray, J., Breazeal, C., Berlin, M., Brooks, A., and Lieberman, J., 'Action parsing and goal inference using self as simulator', in *14th IEEE International Workshop on Robot and Human Interactive Communication (ROMAN)*, Nashville, Tennessee, (2005). IEEE.

[9] Matthew Johnson and Yiannis Demiris, 'Perceptual perspective taking and action recognition', *International Journal of Advanced Robotic Systems*, **2**(4), 301–308, (December 2005).

[10] A. N. Meltzoff and J. Decety, 'What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience', *Philosophical Transactions of the Royal Society: Biological Sciences*, **358**, 491–500, (2003).

[11] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz, 'Enabling effective human-robot interaction using perspective-taking in robots', *IEEE Transactions on Systems, Man, and Cybernetics*, **35**(4), 460–470, (2005).

[12] H. Wimmer and J. Perner, 'Beliefs about beliefs: Representation and constraining function on wrong beliefs in young children's understanding of deception.', *Cognition*, **13**, 103–128, (1983).

Towards A Computational Model of the Self-Attribution of Agency

Koen V. Hindriks[1] and **Pascal Wiggers**[1] and **Catholijn M. Jonker**[1] and **Willem F.G. Haselager**[2]

Abstract. In this paper, a first step towards a computational model of the self-attribution of agency is presented, based on Wegner's theory of apparent mental causation. A model to compute a *feeling of doing* based on first-order Bayesian network theory is introduced that incorporates the main contributing factors to the formation of such a feeling. The main contribution of this paper is the presentation of a formal and precise model that can be used to further test Wegner's theory against quantitative experimental data.

1 INTRODUCTION

The difference between falling and jumping from a cliff is a significant one. Traditionally, this difference is characterized in terms of the contrast between something happening to us and doing something. This contrast, in turn, is cashed out by indicating that the person involved had mental states (desires, motives, reasons, intentions, etc.) that produced the action of jumping, and that such factors were absent or ineffective in the case of falling. Within philosophy, major debates have taken place about a proper identification of the relevant mental states and an accurate portrayal of the relation between these mental states and the ensuing behavior (e.g. [2, 22, 6, 4, 5, 16, 11] to name but a few). In this paper, however, we will focus on a psychological question: how does one decide that oneself is the originator of one's behavior? Where does the feeling of agency come from? Regarding this question we start with the assumption that an agent generates explanatory hypotheses about events in the environment, a.o. regarding physical events, the behavior of others and of him/herself. In line with this assumption, in [19] Wegner has singled out three factors involved in the self-attribution of agency; the principles of priority, consistency and exclusivity. Although his account is detailed, both historically and psychologically, Wegner does not provide a formal model of his theory, nor a computational mechanism. In this paper, we will provide a review of the basic aspects of Wegner's theory, and sketch the outlines of a computational model implementing it, with a particular focus on the priority principle.

The paper is organized as follows: Section 2 provides an outline of Wegner's theory and introduces the main contributing factors in the formation of an experience of will. In section 3, it is argued that first-order Bayesian network theory is the appropriate modeling tool for modeling the theory of apparent mental causation and a model of this theory is presented. In section 4, the model is instantiated with the parameters of the *I Spy* experiment as performed by Wegner and the results are evaluated. Finally, section 5 concludes and gives some directions for future research.

2 APPARENT MENTAL CAUSATION

Part of a theory of mind is the link between an agent's state and its actions. That is, agents describe, explain and predict actions in terms of underlying mental states that cause the behavior. In particular, human agents perceive their intentions as causes of their behavior. Moreover, intentions to do something that occur prior to the corresponding act are interpreted as reasons for doing the action. This understanding is not fully present yet in very young children. But by the age of 4 or 5, children also are able to distinguish intentions from desires or preferences and from the outcomes of intentional actions [3, 23].

But even to adults it is not always clear-cut whether or not an action was caused by ones own prior intentions. For example, when one finds someone else on the line after making a phone call to a friend using voice dialing, various explanations may come to mind. The name may have been pronounced incorrectly making it hard to recognize it for the phone, the phone's speech recognition unit may have mixed up the name somehow, or, alternatively, one may have more or less unconsciously mentioned the name of someone else only recognizing this fact when the person is on the line. The perception of agency thus may vary depending on the perception of one's own mind and the surrounding environment.

In the self-attribution of agency, intentions play a crucial role, but the conscious experience of a feeling that an action was performed by the agent itself still may vary quite extensively. We want to gain a better understanding of the perception of agency, in particular of the attribution of agency to oneself. We believe that the attribution of agency plays an important role in the interaction and the progression of interaction between agents, whether they are human or computer-based agents. As the example of the previous paragraph illustrates, in order to understand human interaction with a computer-based agent it is also important to understand the factors that play a role in human self-attribution of agency. Such factors will enhance our understanding of the level of control that people feel when they find themselves in particular environments. One of our objectives is to build a computational model to address this question which may also be useful in the assessment by a computer-based agent of the level of control of one of its human counterparts in an interaction.

As our starting point for building such a model, we use Wegner's theory of apparent mental causation [20]. Wegner argues that there is more to intentional action than forming an intention to act and performing the act itself. A causal relation between intention and action may not always be present in a specific case, despite the fact that it is perceived as such. This may result in an illusion of control. Vice versa, in other cases, humans that perform an act do not perceive themselves as the author of those acts, resulting in more or less automatic behavior (automatisms). As Wegner shows, the causal link

[1] Man-Machine Interaction Group, Delft University of Technology, The Netherlands, email: {k.v.hindriks, c.m.jonker, p.wiggers}@tudelft.nl

[2] Nijmegen Institute for Cognition and Information, Radboud University Nijmegen, The Netherlands, email: pimh@nici.ru.nl

between intention and action cannot be taken for granted.

Wegner interprets the self-attribution of agency as an experience that is generated by an interpretive process that is fundamentally separate from the mechanistic process of real mental causation [19]. He calls this experience the *feeling of doing* or the *experience of will*.[3] The fact that Wegner's theory explains the feeling of doing as the result of an interpretive process is especially interesting for our purposes. It means that this theory introduces the main factors that play a role in interpreting action as caused by the agent itself retrospectively. It thus provides a good starting point for constructing a computational model that is able to correctly attribute agency to a human agent when it is provided with the right inputs.

Wegner identifies three main factors that contribute to the experience of conscious will, or a feeling of doing: (i) An intention to act should have been formed just before the action was performed. That is, the intention must appear within an appropriately small window of time before the action is actually performed. Wegner calls this the *priority principle*. (ii) The intention to act should be consistent with the action performed. This is called the *consistency principle*. (iii) The intention should exclusively explain the action. There should not be any other prevailing explanations available that would explain the action and discount any intention, if present, as a cause of the action. This is called the *exclusivity principle*.

A crucial factor in assessing the contribution of the priority principle to the feeling of doing is the timing of the occurrence of the intention. In [21] it is experimentally established that the experience of will typically is greatest when the intention is formed about 1 second before the action is performed. As Wegner argues, the priority principle does not necessarily need to be satisfied in order to have a feeling of doing. *People may sometimes claim their acts were willful even if they could only have known what they were doing after the fact* [19]. Presumably, however, an agent that makes up an intention after the fact to explain an event will (falsely) *believe* that it occured prior to that event.

The contribution of the consistency principle to the experience of will *depends [...] on a cognitive process whereby the thoughts occurring prior to the act are compared to the act as subsequently perceived. When people do what they think they were going to do, there exists consistency between thought and act, and the experience of will is enhanced* [19]. The comparison of thought and action is based on a semantic relation that exists between the content of the thought and the action as perceived. The thought may, for example, name the act, or contain a reference to its execution or outcome. The mechanism that determines the contribution of the consistency principle to a feeling of doing thus relies on a measure of how strongly the thought and action are semantically related. Presumably, the contribution of the consistency principle is dependent on the priority principle. Only thoughts consistent with the act that occurred prior to the perceived act, within a short window of time, contribute to a feeling of doing.

The contribution of the exclusivity principle to the experience of will consists in the weighting of various possible causes that are available as explanations for an action. The principle predicts that when the own thoughts of agents do not appear to be the exclusive cause of their action, they experience less conscious will; and, when other plausible causes are less salient, in turn, they experience more conscious will [19]. People discount the causal influence of one potential cause if there are others available [1]. Wegner distinguishes between two types of competing causes: (i) internal ones such as: emotions, habits, reflexes, traits, and (ii) external ones such as external agents (people, groups), imagined agents (spirits, etc.), and the agent's environment. In the cognitive process which evaluates self-agency these alternative causes may discount an intention as the cause of action. Presumably, an agent has background knowledge about possible alternative causes that can explain a particular event in order for such discounting to happen. Wegner illustrates this principle by habitual and compulsive behavior like eating a large bag of potato chips. In case we know we do this because of compulsive habits, any intentions to eat the chips are discounted as causes by knowledge of such habits.

3 COMPUTATIONAL MODEL

One of our aims is to provide a computational model in order to validate and explicate Wegner's theory of apparent mental causation. This theory defines the starting point for the computational model. But the theory does not describe the functioning of the affective-cognitive mechanisms that lead to a feeling of doing at the level of detail which is required for achieving this goal. We thus have to make some modeling choices in order to specify *how* a feeling of doing is created. In this section a computational model is introduced that provides a tool for simulating the feeling of doing. In the next section the model is instantiated with an experiment performed by Wegner as a means to validate that the model also fits some of the empirical evidence that Wegner presents to support his theory.

It is clear that any model of the theory of apparent mental causation must be able to account for the varying degrees or levels in the experience of a feeling of doing, the variation in timing of intention and action, the match that exists between those, and the competition that may exist between various alternative causes. Neither one of these factors nor the feeling of doing itself can be represented as a two-valued, binary state, since humans can experience more or less control over particular events. As observed in [19], even *our conscious intentions are vague, inchoate, unstudied, or just plain absent. We just don't think consciously in advance about everything we do, although we try to maintain appearances that this is the case*.

Given the considerations above, it seems natural to use a probabilistic approach to model the degrees of priority, and consistency and to weigh the various competing alterative explanations. Moreover, the cognitive process itself that results in an experience of will is an interpretive or inferential process. Given the various inputs relating to time and perceived action, a cause that explains the action is inferred which may or may not induce a feeling of doing. A natural choice to model such dependencies is to use Bayesian networks. Bayesian networks [17] have been used extensively to model causal inference based on probabilistic assessments of various sorts of evidence (see for examples of this in research on a *theory of mind* e.g. [8, 18]). Bayesian networks also allow us to use symbolic representations of the thoughts formed and the actions performed by an agent, which need to be compared in order to compute a feeling of doing in the theory of apparent mental causation.

However, Bayesian networks have their limitations. Essentially, Bayesian networks define a joint probability distribution over a predefined set of propositions. To stay within the topic of this paper, one could easily construct a Bayesian network for a particular set of intentions, actions and alternative causes for the actions. As an example, Figure 1 shows a simple causal network modeling that the closing of a door can be caused either by a strong wind or because of the intention of an agent to close the door. The principles of priority and consistency can be encoded in the strengths of the depen-

[3] *Feeling of doing* and *experience of will* are used interchangeably in this paper. Wegner sometimes also uses the phrase *experience of control* as synonym for the former phrases.

dencies in the graph, i.e. in the conditional probability table associated with the node labeled as *The door closes*. Such models can be constructed for every particular situation, but obviously this would not provide a generic account. In the example, if we would like to additionally consider the possibility that another person could have closed the door, a new network would have to be introduced and a new conditional probability table would have to be defined. Instead, what is needed is a more general, higher-level theory that can be used to reason over any event and its potential causes. The model moreover should explicitly model the general principles of priority, consistency and exclusiveness introduced above as well as the interactions between them, rather than hide these contributing factors in a single probability distribution.

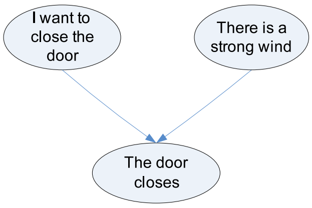


Figure 1. A Simple Causal Network

As many have noted the shortcomings of Bayesian networks, there has been a surge in research on generalizations of Bayesian networks in recent years [13, 7, 9, 14]. These formalisms differ in notation and in representational power, but central to all of these approaches is the ability to represent probability distributions over relations or predicates rather than over atomic propositions. In this paper, Multi-Entity Bayesian Network (MEBN) Theory is used [14]. MEBN is *a knowledge representation formalism that combines the expressive power of first-order logic with a sound and logically consistent treatment of uncertainty*.

An MEBN Theory consists of several MEBN fragments that together define a joint probability distribution over a set of first order logic predicates. Figure 2 shows two MEBN fragments, each depicted as a rounded rectangle, that model the priority principle. A fragment contains a number of nodes that represent random variables. In accordance with the mathematical definition, random variables are seen as functions (predicates) of (ordinary) variables.

The gray nodes in the top section of a fragment are called *context nodes*; they function as a *filter* that constrains the values that the variables in the fragment can take. In contrast to the nodes in the bottom section of a fragment, context nodes do not have an associated probability distribution but are simply evaluated as true or false. Another perspective on these nodes is that they define what the network is about. The context nodes labeled with the $IsA(t, v)$ predicate define the type t of each of the variables v used. In our model, we distinguish intentions, events, opportunities, and time intervals in which the former may occur. Intentions are *mental states* which are to be distinguished from events, which are temporally extended and may change the state of the world. Opportunities are states which enable the performance of an action. In the model, the probabilities associated with each of the nodes should be interpreted as the likelihood that the agent attaches to the occurrence of a particular state, event or other property (e.g. causal relationship) given the available evidence.

Dark nodes in the bottom section of a fragment are called *input nodes* and are references to nodes that are defined in one of the other fragments. In Figure 2, the node in the right fragment labeled $Exists(a, t_a)$ is an input node. To ensure that the model defines a proper probability distribution, a node can be defined in a single fragment only, in which it is said to be *resident*. The node labeled $Exists(a, t_a)$ is resident in the left fragment in Figure 2.

As usual, the links between nodes represent dependencies. Every resident node has a conditional probability table attached that gives a probability for every state of the node given the states of its parent nodes. Prior distributions are attached to resident nodes without parents. Essentially, every fragment defines a parameterized Bayesian network that can be instantiated for all combinations of its variables that satisfy the constraints imposed by its context nodes.

In order to be able to compute a feeling of doing, the prior probability distributions are assumed to be given in this paper. The computational model presented does not explain how explanatory hypotheses about perceived events are generated, nor does it include an account of the perception of these events. Even though the model assumes this information somehow has already been made available, it is setup in such a way that it already anticipates an account for computing at least part of this information. In particular, the mechanism approach of [1] to explain causal attribution has played a guiding role in defining the model. The basic idea of this approach is that *causal attribution involves searching for underlying mechanism information (i.e. the processes underlying the relationship between the cause and the effect)*, given evidence made available through perception and introspection. Assuming that each mechanism defines a particular covariation (or joint probability distribution) of the contributing factors with the resulting outcome, the introduction of separate probability distributions for each particular event that is to be explained can be avoided. As a result, the number of priority and causality fragments needed is a function linear in the number of mechanisms instead of the number of events.

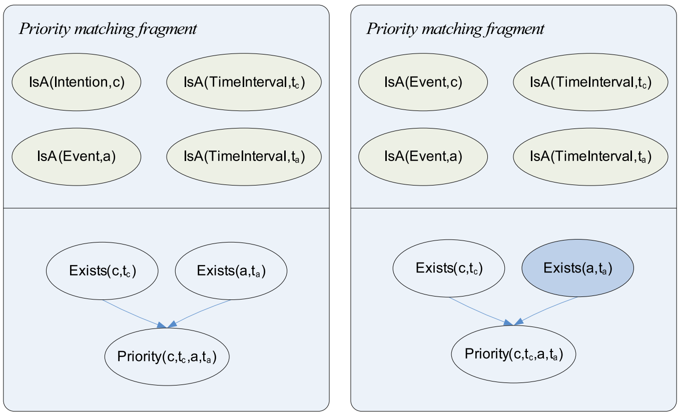


Figure 2. Priority Fragments

3.1 Priority Fragments

The priority principle is implemented by the Priority fragments in Figure 2. Though these fragments are structurally similar, two fragments are introduced in line with the idea that different causal mechanisms may associate different time frames with a cause and its effect. For reasons of space and simplicity, Figure 2 only depicts two fragments, one associated with intentional mechanisms leading to action and a second one for other causal events. The exact time differences depend on the mechanism involved. For example, when moving the steering wheel of a car one expects the car to respond immediately, but a ship will react to steering with some delay.

The *Exists* random variables model that an agent may be uncertain whether a particular state or event has actually taken place at a particular time (also called the *existence condition* in [12]). If there is no uncertainty these nodes will have value true with probability one. The probability associated with the *Priority* random variable is nonzero if the potential cause occurs more or less in the right time frame before the event that is explained by it and the associated probability that the relevant events actually occurred is non-zero. In line with [21], the probability associated with the intentional mechanism increases as the time difference decreases to about one second. As one typically needs some time to perform an action, the probability starts to decrease again for time intervals less than one second. Each of the fragments may be instantiated multiple times, illustrated in Section 4, depending on the number of generated explanatory hypotheses.

3.2 Causality Fragments

Figure 3 depicts two fragments corresponding respectively with the intentional mechanism (left) and another type of mechanism (right) that may explain an event. In this case, the fragments are structurally different in two ways. First, even though both fragments require that cause c and effect a are consistent with the mechanism associated with the fragment, the consistency nodes are different. The type of consistency associated with the intentional fragment, called *intentional consistency*, is fundamentally different in nature from that associated with other mechanisms as it is based on the degree of *semantic* relatedness of the content of intention c and the event a (represented as a probability associated with the node). This reflects the fact that one of Wegner's principles, the consistency principle, is particular to intentional explanations. Second, an additional context node representing an opportunity o to act on the intention is included in the fragment corresponding with the intentional mechanism. An intention by itself does not result in action if no opportunity to act is perceived. In line with common sense and philosophical theory [5], the intentional mechanism leads to action given an intention and the right opportunity as input. The model entails that the presence of multiple opportunities increases the probability that a relevant intention is the cause of an event. Additional detail is required to model this relation precisely, but for reasons of space we refer to [10] for a formal model.

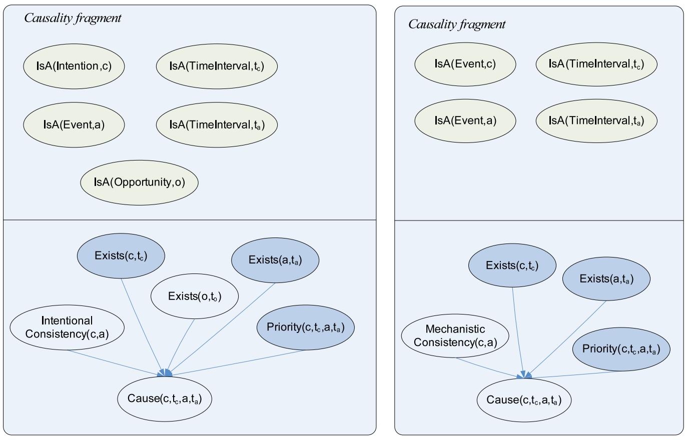


Figure 3. Causality Fragments

The node labeled $Cause(c, t_c, a, t_a)$ in the intentional fragment models the *feeling of doing*. The associated probability of this node represents the probability that the intention c of an agent has caused event a. In other words, it represents the level of self-attribution of agency for that agent. The probability associated with the node depends on the priority and consistency as well as on the presence (i.e. existence) of both c and a. Obviously, if either c or a is not present, $Cause(c, t_c, a, t_a)$ will be false with probability 1. Additionally, in the intentional fragment an opportunity o must exist.

3.3 Exclusivity fragment

In order to model the exclusivity principle, an exclusivity fragment is introduced as depicted in Figure 4. In general, if there are multiple plausible causes for an event, exclusivity will be low. Technically, this is modeled as an exclusive-or relation between the competing causes. The value of the random variable $Exclusivity$ is set to true to enforce exclusivity. As a result, given two causes of which only one is very likely, the posterior probability of the unlikely cause is reduced. This effect is known as the *discounting effect*, also called *explaining away* [17], and has been studied extensively (e.g. [1]).

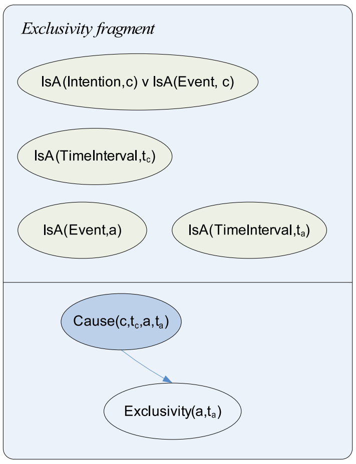
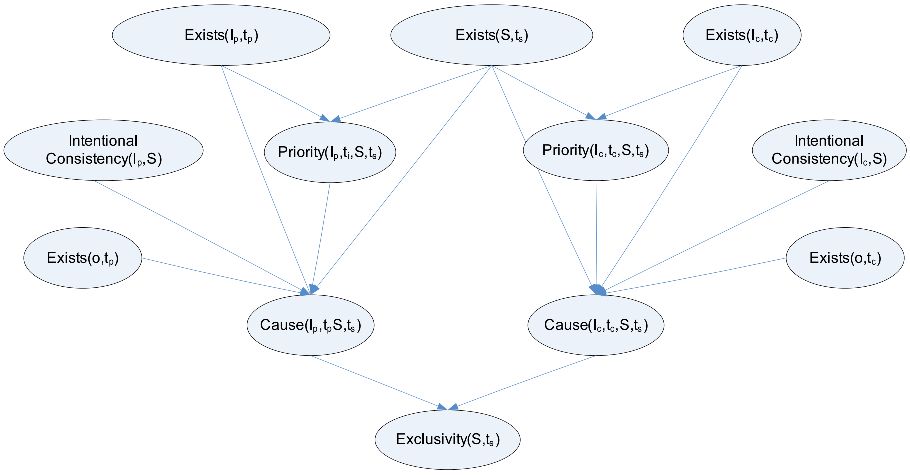


Figure 4. Exclusivity Fragment

Given an event to be explained and a number of generated explanatory hypotheses (including all contributing factors associated with a particular mechanism), each of the fragments discussed is instantiated accordingly, taking into account the context conditions. To obtain a single, connected Bayesian network, all of the resulting fragments are connected by merging the reference nodes with their resident counterparts. Using this network, the *feeling of doing* can be computed by performing probabilistic inference and querying the $Cause(c, t_c, a, t_a)$ variable in the intentional fragment given the values of the other nodes in the network. By querying other $Cause$ variables we can find by means of comparison which of the potential causes is the most plausible one. As a result, only when the node representing the feeling of doing has a high associated probability an agent would explain the occurrence of an event as caused by itself.

4 SIMULATION OF THE *I SPY* EXPERIMENT

In this section, an instantiation of the model that corresponds with an experiment performed by Wegner is presented. In [21] the results of the *I Spy* experiment are presented that tested whether participants report an experience of agency for something that is most likely the result of someone else's action. In the experiment two participants are seated on opposite sides of a table. On the table a square board that is attached to a computer mouse is located and both participants



$\text{Exists}(I_p,t_p)$
$\text{Exists}(S,t_s)$
$\text{Exists}(I_c,t_c)$
Intentional Consistency(I_p,S)
$\text{Priority}(I_p,t_i,S,t_s)$
$\text{Priority}(I_c,t_c,S,t_s)$
Intentional Consistency(I_c,S)
$\text{Exists}(o,t_p)$
$\text{Exists}(o,t_c)$
$\text{Cause}(I_p,t_pS,t_s)$
$\text{Cause}(I_c,t_c,S,t_s)$
$\text{Exclusivity}(S,t_s)$

Figure 5. An instance of the model for the *I Spy* experiment

are asked to put their fingertips on the board and to move the mouse by means of the board in slow sweeping circles. By doing so, a cursor is moved over a computer screen showing a photo from the book *I Spy* [15], hence the name of the experiment, picturing about 50 small objects. The subjects had to move the mouse for about 30 seconds after which they would hear a 10 second clip of music through headphones and within this period they had to stop moving the mouse and then rate on a continuous scale whether they allowed the stop to happen or intended the stop to happen. In addition to the music, subjects would occasionally hear words over the headphones. Participants were told that they would hear different bits of music and different words. One of the participants however did not hear music at all, but was a confederate who received instructions from the experimenter to stop on a particular picture or to let the other participant determine the stop. The forced stops were timed to occur at specific intervals from when the participant heard a corresponding word that was intended to prime a thought about items on the screen. By varying timing, priority was manipulated. For unforced stops the words heard by the participant corresponded about half of the time to an object on the screen.

It turned out that in initial experiments in which the confederate did not force stops the mean distance between stops and the pictures that were primed by words was not significantly different from the mean distance in trials in which the prime word did not refer to an object on the screen. These initial experiments were performed to confirm that participants would not stop the cursor on an object simply because of hearing the word. In consecutive experiments, however, where the stops were forced by the confederator, participants tended to perceive the stops as more or less intended, dependent on the time interval between the hearing of the prime word and the actual stop. In particular, if the word occurred between 5 and 1 seconds before the stop, a significant increase in self-attribution was observed.

4.1 Instantiating the Model

Based on the description of the *I Spy* experiment and the results presented in [21], an instantiation of the computational model has been derived.

Given the description of the experiment, a stop on or near a particular picture is only explainable by the movements made by the participant and the confederate. The stop does not need to be interpreted as an *intentional* action, however. This will depend on the likelihood assigned by the participant that a relevant intention is thought to be present by either the participant itself or the confederate. In the *I Spy* experiment it is reasonable to assume that the explanatory hypotheses generated by the participant consist only of intentions to stop the cursor on a particular picture. Given the additional fact that the confederate forces a stop on a picture that corresponds with the prime word, it is, moreover, reasonable to assume that to explain this event only an intention to stop on the picture described by the prime word and an opportunity to do so are generated. If the prime word is, for example, *swan*, the participant thus is assumed to only generate the hypotheses that the participant intends to stop on the swan picture and the confederate intends to stop on the swan picture. These intentions are supposed to be generated in conjunction with the opportunity to do so by means of moving the mouse. Finally, appropriate time intervals need to be associated with the intentions as well as the events. In the *I Spy* experiment, what matters is the actual time difference between these, so any choice of interval with the right time difference can be used.

Figure 5 shows the Bayesian network that is obtained by merging several instantiations of the model fragments as explained above and by instantiating the variables with these values. Intentions are respectively labeled I_p and I_s and the opportunity is labeled o. The event of stopping on the swan picture is denoted by S. The priority and causality fragments associated with the intentional mechanism are instantiated twice, once for the relevant participant's intention and once for the confederate's intention. As a result, two possible causes are identified which is reflected by the two $Cause$ random variables in the network. Each of the resident nodes are merged with input nodes to obtain a single connected network.

4.2 Estimating Probability Distributions

Given that the structure of the network adequately models the participant's causal inferences, the remaining challenge is to associate the appropriate (conditional) probability distributions with the nodes in the network.

In the experiment it is tested whether primed words influence the attribution of agency, or a feeling of doing. In the model this is reflected by the fact that the participant believes at least with some probability that s/he formed an intention to stop on the picture. It is not quite clear how probable the participant will think s/he had the relevant intention based on the description in [21]. It is well-known that priming may have various measurable effects but as reported in [21] the behavior of the participant is not significantly influenced. It may be that the participant constructs an intention after the fact and that this intention reconstruction is influenced by the priming. In any case, it seems that the probability should not be set too high. To incorporate a possible effect of priming it should be slightly higher than uncertainty (a probability of 50%). Similar reasoning would indicate that the participant's belief that the confederate had the relevant intention to stop on the swan picture would be less than 50%, simply because there is no reason at all to suggest that the confederate would have such an intention. Maybe the fact that during the instructions the participant is informed that the confederate hears other words may also be of influence on the relative certainty associated with the belief that the confederate does not have the relevant intention.

The prior probability associated with the opportunity to stop on a particular picture, we estimate, will be quite low. The description in [21] does not make this completely clear. The setup suggests that mouse movement will be less precise in comparison with the steering of a mouse in more normal conditions. In line with this, the probability associated with the opportunity node is set to about 30%, to reflect that it will be quite hard to steer the mouse to a target.

We assume that the participant has virtually no uncertainty about the event to be explained, i.e. the stop on the swan picture, which seems reasonable given the setup of the experiment which makes it easy to observe where the cursor is located on the screen.

Finally, the prior probability of the intentional consistency nodes has to be established. Since the prime word that the participant hears refers to the object on which the cursor stopped on the screen (although the precision is not indicated in [21]), we have set this probability quite high for both participant and confederate to about 80% (both participant and confederate's intention have the same content, which semantically represents the stop event).

The remaining nodes for which we need to define conditional probability distributions are the nodes labeled with $Priority$, $Cause$ and $Exclusivity$ random variables. These conditional probability distributions are not given through perception or other information about a particular event that is to be explained. These probability distributions are not situation-dependent in contrast with the prior probabilities discussed above. They define the logic of the corresponding fragments.

The quantitative data presented in [21] about the influence of the time interval between the primed word and the (forced) stop on the reported perceived intention can be used to assign a probability distribution to the priority node. As mentioned above, the priority fragment associates the probability that cause and effect are related to each other in the right time frame depending on the mechanism. This should be highest according to the findings presented in [21] for time differences of 5 or 1 second, and very low for time intervals of 30 second and -1 second (i.e. the prime word is provided after the stop).

The conditional probability distribution associated with the $Cause$ random variable is defined as follows: It yields a high probability when all of its inputs are true; in case one of the $Exists$ nodes is believed to be very likely to be false, the $Cause$ node has a very low associated probability; the probabilities associated with the $Priority$ and $IntentionalConsistency$ input nodes give rise to a more gradual effect on the probability associated with the $Cause$ node.

Finally, the $Exclusivity$ variable is defined as an exclusive-or with some noise to indicate that exclusivity is the preferred state, but such that the possibility of two causes that explain an event is not completely excluded.

4.3 Evaluating the Results

The resulting model including the associated probability distributions gives the same results as those reported in [21]: If the a priori probability associated with the $Priority$ variables is higher (corresponding to the time interval between 5 to 1 seconds), then a significantly higher feeling of doing is produced than otherwise. The second column of Table 1 shows the posterior probability of the $Cause(I_p, t_p, S, t_s)$ node that models the feeling of doing for several a priori probabilities of the $Priority$ variable. For a probability of 0.85 for priority the probability of $Cause$ corresponds to the feeling of doing for a time difference of about 1 second as described in [21]. Similarily, the values obtained with a probability for priority of 0.8 and 0.35 correspond to the feeling of doing reported in [21] for respectively 5 seconds and 30 seconds time diffence between the prime word and the stop of the cursor.

In [21], also the variance in feeling of doing observed in the experiment is reported. One would expect that a person's personality influences his feeling of doing. Various people, for example, might be more or less sensitive to priming or might have a strong or weak tendency to claim agency in a setup such as in the *I Spy* experiment. We tested the model with different values of priority with a moderated a priori probability for the existence of intention of 0.45 and with a high a priori probability of 0.65 for the existence of an intention. The corresponding posterior probabilities of the cause node are shown in Table 1. These probabilities adequately correspond with the variance reported by Wegner, which gives some additional support for the proposed computational model.

	$P(Exists(I_p, t_p))$		
$P(Priority)$	0.55	0.45	0.65
0.3	0.41	0.36	0.45
0.35	0.44	0.39	0.48
0.5	0.51	0.46	0.56
0.8	0.62	0.56	0.66
0.85	0.63	0.58	0.67

Table 1. Posterior probability of $Cause(I_p, t_p, S, t_s)$ for different a priori probabilities of $Priority(I_p, t_p, S, t_s)$ and $Exists(I_p, t_p)$.

5 CONCLUSION AND FUTURE WORK

In this paper, a first step towards a computational model of the self-attribution of agency is presented, based on Wegner's theory of apparent mental causation [19]. A model to compute a *feeling of doing* based on first-order Bayesian network theory is introduced that incorporates the main contributing factors (according to Wegner's theory) to the formation of such a feeling. The main contribution of this paper is the presentation of a formal and precise model that provides detailed predictions with respect to the self-attribution of agency and that can be used to further test such predictions against other quantitative experimental data. An additional benefit of the model is that given empirical, quantitative data the parameters of the network can be learned, using an algorithm as described in [14].

A number of choices had to be made in order to obtain a computational model of Wegner's theory of apparent mental causation. Not all of these choices are explicitly supported by Wegner's theory. In particular, it has been hard to obtain quantitative values to define the probability distributions in our model. The report of the *I Spy* experiment in [21] does detailed information, but did not provide sufficient information to construct the probability distributions we need. Certain values had to be guessed in order to obtain outcomes corresponding with the results in [21]. The only validation of these guesses we could perform was to verify whether variation of some of the input values of our model could be said to reasonably correspond with the reported variations in the experiment in [21]. It is clear that more work needs to be done to validate the model. In future work, we want to design and conduct actual experiments to validate and/or refine the model of self-attribution.

To conclude, we want to remark that there are interesting relations here with other work. As is argued in [18], Bayesian networks are not sufficient as cognitive models of how humans infer causes. These networks are very efficient for computing causes, but are themselves instantiations from more general, higher-level theories. In a sense, this is also the case in our model since both the consistency fragment as well as the causality fragment in our first-order Bayesian theory of apparent mental causation need to be instantiated by other domain-specific theories in order to derive the right semantic relations between thoughts and actions, and to identify potential other causes of events. Additional work has to be done to fill in these gaps in the model, starting from e.g. ideas presented in [1, 18].

REFERENCES

[1] W.-K. Ahn and J. Bailenson, 'Causal attribution as a search for underlying mechanisms', *Cognitive Psychology*, **31**, 82–123, (1996).

[2] G.E.M. Anscombe, *Intention*, Harvard University Press, 1958/2000.

[3] J.W. Astington, *The Child's Discovery of the Mind*, Harvard University Press, 1993.

[4] D. Davidson, *Essays on actions and events*, Oxford University Press, 1980.

[5] F. Dretske, *Explaining behavior*, MIT Press, 1988.

[6] H. Frankfurt, 'The problem of action', *American Philosophical Quarterly*, **15**, 157–162, (1978).

[7] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer, 'Learning probabilistic relational models', in *IJCAI*, pp. 1300–1309, (1999).

[8] A. Gopnik and L. Schulz, 'Mechanisms of theory formation in young children', *Trends in Cognitive Science*, **8**, 371–377, (2004).

[9] D. Heckerman, C. Meek, and D. Koller, 'Probabilistic models for relational data', Technical report, Microsoft Research, (2004).

[10] C.M. Jonker, J. Treur, and W.C.A. Wijngaards, 'Temporal modelling of intentional dynamics', in *ICCS*, pp. 344–349, (2001).

[11] A. Juarrero, *Dynamics in action*, MIT Press, 2002.

[12] J. Kim, *Supervenience and Mind*, Cambridge University Press, 1993.

[13] D. Koller and A. Pfeffer, 'Probabilistic frame-based systems', in *AAAI/IAAI*, pp. 580–587, (1998).

[14] Kathryn B. Laskey, 'MEBN: A Logic for Open-World Probabilistic Reasoning', Technical Report C4I-06-01, George Mason University Department of Systems Engineering and Operations Research, (2006).

[15] J. Marzollo and W. Wick, *I Spy*, New York: Scholastic, 1992.

[16] *The philosophy of action*, ed., A. Mele, Oxford University Press, 1997.

[17] J. Pearl, *Probabilistic Reasoning in Intelligent Systems - Networks of Plausible Inference*, Morgan Kaufmann Publishers, Inc., 1988.

[18] J.B. Tenenbaum, T.L. Griffiths, and S. Niyogi, 'Intuitive Theories as Grammars for Causal Inference', in *Causal learning: Psychology, philosophy, and computation*, eds., A. Gopnik and L. Schulz, Oxford University Press. In press.

[19] Daniel M. Wegner, *The Illusion of Conscious Will*, MIT Press, 2002.

[20] Daniel M. Wegner, 'The mind's best trick: How we experience conscious will', *Trends in Cognitive Science*, **7**, 65–69, (2003).

[21] Daniel M. Wegner and T. Wheatley, 'Apparent mental causation: Sources of the experience of will', *American Psychologist*, **54**, (1999).

[22] *The philosophy of action*, ed., A.R. White, Oxford University Press, 1968.

[23] H. Wimmer and J. Perner, 'Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception', *Cognition*, **13**, 103–128, (1983).

Modelling attentionally- and emotionally-sensitive social agents

Christopher Peters [1]

Abstract.
The detection of attentive and emotional behaviours of others, and the ability to infer elements of mental state and intent from such behaviour, is fundamental to the functioning of social entities. In this paper, we consider the work done so far and the next steps in our approach for modelling a real-time, socially sensitive, embodied virtual agent. Our initial aim is to create a broad but lightweight computational framework suitable for real-time agents, spanning perception through to action, so that agents can have the ability to make limited useful inferences about elements of the mental state and intentions of other agents. These inferences can be based on the synthetic visual perception of certain attentional and emotional behaviours of other agents, such as gaze direction and facial expression. A key element is that these inferences inform the agents decision making processes so that its actions are based not only on its own goals, but also on its theories of the goals of the other. We present a concrete interaction initiation scenario and its evaluation in order to demonstrate work done so far and consider some of the numerous interesting benefits and possibilities for extensions to the system.

1 INTRODUCTION

While the achievements and advancements in creating artificial social entities have been exemplary, in comparison with real humans, contemporary automated humanoid agents are still often perceived as asocial and unnatural human-machine ‘hybrids’. In some ways, their appearance and behaviour may have parallels with our own, yet in other ways, it varies dramatically. Above all, although these hybrids can often make very intricate calculations and decisions, they seem insensitive to very fundamental human modes of signalling and awareness - one wonders when interacting with them whether it is the case that they just can’t detect our desires, feelings and intentions, or if they simply don’t care about them.

In this paper, we present our ongoing work on a broad, lightweight framework that we hope will give real-time embodied agents the ability to be more sensitive with respect to basic human-like behaviours. To us, the notion of social sensitivity encompasses a gamut of complexity, ranging from simply making an eye-movement to acknowledge the presence of another, for example while passing on the street, to fully empathic displays where one feels and reacts with sadness to the others anguish. The key idea behind social sensitivity is that one is given the impression that the agent is aware of and considers ones existence (by its very nature, a social consideration) in its planning: It is not asocial, and it should certainly not ignore you, even if the extent of its acknowledgement is merely confined to that of a subtle gesture of ‘social inattention’ [12].

We attempt to create socially sensitive, mindful agents by endowing them with limited faculties relating to *theory of mind* [22]. Our modelling is broad in the sense that it starts from perception and takes consideration of factors through to behaviour. As such, it does not have particular depth in any one specific area - for example, there is not yet complex gaze tracking data or facial expression recognition, but these are rather taken from the animation of the other agent in the virtual environment using synthetic senses. The approach taken here is that more elaborate parts can be added or interchanged as desired in a modular manner, but will contribute to the same core set of high-level inferences (in our case, particularly that of *interest*) and theories that are used as part of the decision making process. As such, modelling has been taking place in an iterative process. Our key objective in early iterations has been to investigate if and how basic ‘mind-reading’[3] skills from visual signals, such as eye and head direction, can be interpreted into a more complex mental state of *interest in interaction* and how this can in turn be used to produce more elaborate behaviours between autonomous agents. Generally speaking, the first part of this paper is concerned with this topic. The second half is concerned with an important extension that we are modelling in the next iteration of elaboration, which is the inclusion of a mechanism for allowing mindfulness towards the emotions of the other as well as basic empathic reactions towards them. As such, Section 2 reviews important related areas of research that consider computational theories of mind for inferring aspects of mental state of another from their behaviour. An overview of our modelling approach is presented in Section 3. This considers some of the theory that informs the design of our model. What we consider to be the attention-related aspects of our model are covered in Section 4, including a description of a prototype scenario and evaluation study. Section 5 details the key new addition we propose for the model, which allows it to show sensitivity towards emotion-related aspects of behaviour, including an empathic response. Section 6 concludes by discussing important considerations for the work presented.

2 BACKGROUND

As mentioned, our work uses research based on theory of mind for driving agent behaviour and in this way it is related to research being pursued in a number of domains, spanning topics such as recognition and interpretation of signals from visual input to methods for conducting in-depth theory of mind reasoning and game playing. In terms of the former, a vast amount of work (see [18] for survey) has been conducted on the recognition and interpretation of human features and displays from visual data e.g. faces and facial expressions. Much of this work, however, only goes as far as establishing a relatively low- or mid-level categorisation, for example the cod-

[1] LINC Laboratory, University of Paris 8, email: c.peters@iut.univ-paris8.fr

ing of facial actions or basic emotion category [9]. These seem to stop somewhat short of providing a higher-level account for use in inferring possible mental states or intentions of the user. Some exceptions go beyond categorisation of basic emotions to 'mind-read' more diverse and complex mental states, *interested* and *thinking* for example, based on head gestures and/or facial expressions [10], and these are particularly relevant to our studies here.

Social roboticists have been constructing and demonstrating impressive practical models of theory of mind that encompass perceptual, reasoning and behavioural output aspects. This work should probably stand out as a source of inspiration to illuminate the path for the agents community in constructing broad practical theory of mind models. Scassellati [24], for example, constructed a humanoid robot as a test bed for the evaluation of models of human social development. The robot has been endowed with a theory of mind based on a merger of two theoretical models, by Leslie [15] and Baron-Cohen [2]. Early stages of the system use the movement of environmental stimuli to distinguish between animate/volitional and inanimate/nonvolitional objects. Animate stimuli are tagged as *intentional* and then further processed by successive layers of theory of mind. Such work follows a theme whereby theory of mind mechanisms help the robot to be more mindful of users, for example in relation to developmental models of shared and joint attention [14][8].

In the agents community, most approaches to theory of mind are focused purely on reasoning aspects [16][7][6]. For example, Pynadath and Marsella [23] present PsychSim, a multiagent-based simulation tool for modelling interactions and influences between groups and individuals. Each agent has beliefs about its environment and recursive models of other agents, allowing it to communicate beliefs about other agents' beliefs, goals and intentions and be motivated to use communication to influence other agents' beliefs about agents.

In contrast, the approach described here is focused more on providing a framework incorporating perceptually-oriented rather than reasoning-oriented theory of mind capabilities for agents. We view these two strands as being complimentary to each other, as many of the reasoning-based approaches do not consider perceptual and behaviour output aspects and seem to lack principled modular approaches, something of great necessity for programming practical computational models for agents.

3 OVERVIEW

Our approach to modelling socially sensitive agents has been divided into two main stages of development. Both stages relate to different aspects of the theoretical literature that we base our model on [2][4].

The first stage of our research has been on an influential model that Baron-Cohen refers to as a *mind-reading* system [2] and relates to how non-verbal behaviour can be used to infer behavioural intentions. In particular, it emphasises the role of the eyes and the evolutionary importance of gaze detection, not only in humans but also in many primates. It consists of a series of specialised modules (see Figure 1), including:

- Eye-direction Detector (EDD) The EDD is a social cognition module exclusively based on vision. It functions by detecting the presence of eyes or eye-like stimuli in the environment and computing the direction of gaze (e.g. directed or averted).
- Intentionality detector (ID) The ID module attributes the possibility of an object having goals and desires based on self propulsion, i.e. notions of animacy and intention. One should not, for example, attribute volitional behaviour to a brick, even if it is moving in the environment.

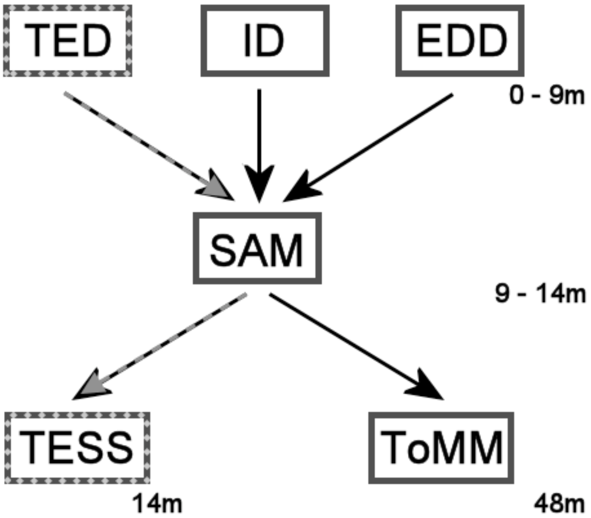


Figure 1. The high-level schematic for the full system proposed by Baron-Cohen [2][4]. The DAD of Perrett and Emery [19] replaces the EDD. The more recent additions to the model, the TED and the TESS (dashed lines) enable the system to detect emotions and engage in appropriate empathic responses towards them, respectively.

- Shared Attention Mechanism (SAM) The SAM is an important conduit in the model. It connects diadic information from the EDD and ID together in order to form triadic relationships. In this way, it can be known that one is engaged with another in paying attention to a third region or object of interest. Although such a module could also be capable of generating shared attention behaviours, this is a broad topic that we will not address specifically here.
- Theory of Mind Mechanism (ToMM) This module stores the attribution of mental states to the other agent and is based on the results of interactions between the other modules. It contains working theories that may not necessarily be correct, but are nonetheless vital for forming an internal representation of the possible motives behind the actions of other living entities.

Perrett and Emery [19] evaluated this system from a neurophysiological perspective and also proposed some further modules:

- Direction of attention detector (DAD) This is a more general form of the EDD above, that combines information from separate detectors that analyse not only gaze, but also body and direction of locomotion.
- Mutual attention mechanism (MAM) This is a special case of shared attention where the relationship is dyadic, involving mutual gaze and eye contact. In this situation, the goal of the participants' attention is each other.

We have already implemented, evaluated and demonstrated a prototype of such a system which uses simple interpretations from gaze, body and locomotion direction to allow an agent to engage in a conversation initiation scenario. The implementation is described in more detail in Section 4.

One vital aspect missing from the previous model is consideration of emotional aspects (although we note that the label *interest*, which one would presume is based also to a large degree on the previous factors, is sometimes referred to as an affective state). The second stage of our work, and one that we have embarked upon recently, relates to the detection of emotional expressions in order, not only to provide more information for disambiguating the possible mental state of the other, but also to facilitate the generation of an appropriate emotional response. Baron-Cohen refers to this as the 'empathizing' system [4], and has proposed two extra modules that augment his previous system of mind-reading (see dashed boxes, Figure 1):

- The emotion detector (TED) This can detect the basic emotions [9] of the other and has dyadic represents that can contain affective states. For example, 'Mother - is *unhappy*'.
- The empathising system (TESS) This allows for an appropriate affective state to be triggered in an observer by the emotional state identified in another. For example, 'I am *horrified* - that you are *in pain*'. It is also pointed out that this module helps ensure a drive for organisms to help each other.

Our preliminary ideas for steps towards implementing a computation equivalent, and its possible uses, are described in Section 5.

Although we use Baron-Cohens work as a high-level guide, in terms of computational implementation, it is very broad and somewhat sparse: there are plenty of details and blanks to be filled in and space for modifications. Nonetheless, Baron-Cohens model is useful for modelling social agents since it provides a modular framework based on suspected real-life dissociations, it incorporates perceptual processes, and its validity can be (and is being [19][5], see [8] for an alternate view) tested. Indeed, implementing such computational models and observing behavioural output is a further way in which validation may take place.

4 AN ATTENTIONALLY-SENSITIVE PROTOTYPE

We propose the general notion that an entity is *attentionally-sensitive* when it is particularly attuned and adapt at paying attention to the attention-related behaviours of other entities in order inform it about the possible presence of opportunities or dangers in its environment. These potential opportunities or dangers may arise directly from the entity being observed (e.g. the entity staring the perceiver in the eye) or may be located elsewhere in the environment, but being signalled indirectly by the scrutiny of the observed entity (e.g. we may follow the gaze of another to establish the source of its interest). In this Section, we describe our achievements so far in using the previously described theory (Section 3) to construct attentionally-sensitive autonomous agents.

Our model is composed of computational equivalents of the modules mentioned earlier, such as the intentionality detector, *ID* and the direction of attention detector, *DAD*, as well as a synthetic sensory component for simulating the vision of the agent within the virtual environment and some memory components for storing the results of processing operations along the way. The overall purpose of the system is to extract animate entities (in this case, other agents) from the visual perception of the agent, and then process the direction of the entities' subparts to calculate where their attention is directed. Over time, these calculations are accumulated into higher-level metrics and simple theories, such as if our agent believes that the other has seen it, or believes that the other believes it has been seen. These theories then inform the decision making mechanism of the agent so it can base decisions not only on its own goals, but also account for what it thinks the goals of the other might be. As such, these theories are constrained very much to our scenario of conversation initiation. Yet despite the comparative simplicity of the 'theories' we have implemented so far, they have given us the potential to produce far more elaborate social reasoning between agents. We will now describe in more detail the functioning of the model, the scenario and an evaluation study.

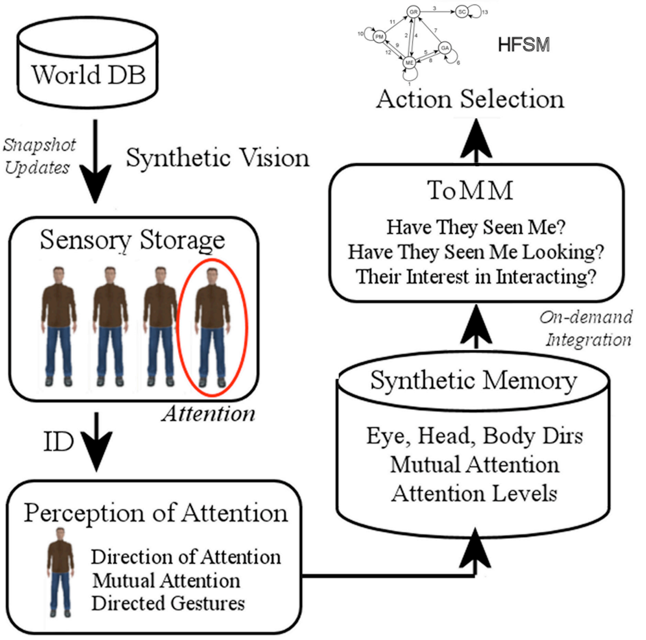


Figure 2. An overview of the main stages in the model. In the diagram, the flow of processing proceeds according to the arrows, from top to bottom, left to right. Representations become more explicit as processing progresses: a large amount of information must be processed at the level of the synthetic vision system, while only a few high-level values are stored at the level of the theory of mind module (ToMM).

4.1 Flow of Processing

The flow of processing through the various components for a single agent is summarised as follows (see Figure 2): The vision system of the agent takes frequent snapshots of the virtual environment in order to establish basic visibility information about what the agent can see. Visible items are stored as false-color percepts in a short-term sensory memory, or *STSS*. At each visual update, these percepts are processed by the ID module (see Section 3), which filters agents and their subparts into a *person percepts list*. Attended-to entries in the person percepts list, as determined by a visual attention module, are resolved and elaborated before being processed by the DAD, which measures the orientation of subparts with respect to the self. Information from the DAD is used by the MAM to establish if there is eye-contact or if agents are paying attention to each other. This information, along with the output of the DAD are time-stamped and stored as a record in the short-term memory, or *STM*. Records in the STM are integrated, on-demand, over time to provide updates to the simplified theories stored in the ToMM.

4.2 Attention, Interest and Intention

The prototype that we present here has been specially geared towards the detection of attentive motions from other entities, for example if they gaze at the observer or move towards it. This sensitivity is achieved by the DAD, which integrates the orientation of other agents' subparts (eyes, head and body) into a single attention metric, called an *attention level*, for a single instant of time. Those subparts oriented towards the agent receive a higher weighting, and the eyes receive a higher weighting than the other subparts. A further metric is then derived from the temporal integration of the attention levels. It is referred to as the *interest level*. Thus, agents utilising this model are sensitive to the attention behaviours of other agents, such as if and when they have been looked at, the attention that another agent may be paying to them or the amount of interest that other agents have in them in a specific temporal window. Furthermore, agents also store their theory as to whether they think the other agent has seen the observer looking at them.

4.3 Scenario

We have created a street conversation scenario to test this model and see if more complex social behaviours could be obtained. Implementation utilised the Torque game engine (http://www.garagegames.com) and runs in real-time on a desktop PC. Since Torque supports Linux and Macintosh platforms, the work presented here should also be capable of running on these, although thus far it has only been tested on desktop and laptop Windows systems. Two agents are placed in a street environment, each with varying goals and relationships. One or both agents may have the goal to engage in conversation, but each does not know the goal of the other. The likelihood that one agent will attempt to start a conversation with another is based not only on its own goal, but on its theory as to the others goal. This is based on the attention and interest of the other. Attention and interest may often be used as a subtle, covert cue for starting conversation: we may use attention to initially signal our intention to communicate or search for the others intention to communicate before we commit. This may save us the socially embarrassing situation of opening communicating with a recipient who does not reciprocate [13]. The more overt our cues are, the riskier it may be that those around us will notice if the other does not reply - it is a wise policy socially to establish the likelihood that the other will respond before we become overt with our requests.

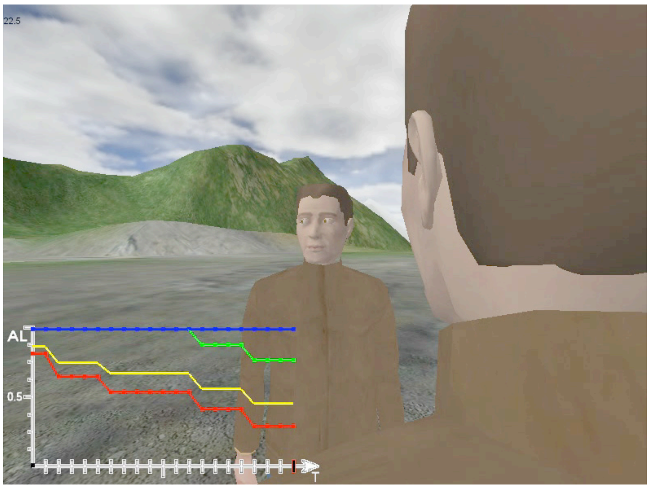
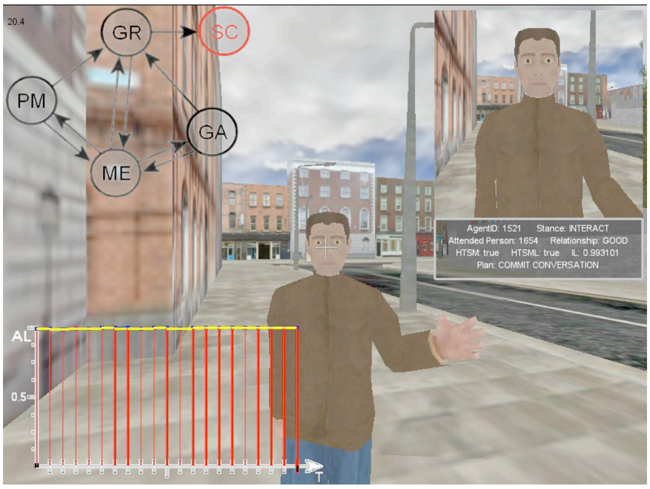


Figure 3. Depicted here is a graph (bottom left) of the eye (red), head (green), body (blue) and corresponding attention level (yellow) that the observing agent nearer the camera perceives from the observed agent facing the camera based on the orientation of its body-parts with respect to the observer. This is the main function of the DAD module.



AgentID: 1521 Stance: INTERACT
Attended Person: 1654 Relationship: GOOD
HTSM: true HTSML: true IL: 0.993101
Plan: COMMIT CONVERSATION

Figure 4. An agent decides to start conversation with the observer. The FSM is pictured on the top-left, with the relevant information about agent goals on the top right. The graph showing the agents perception of the attentive behaviours of the other is shown on the bottom-right: the red lines here signify that the observer perceives that mutual gaze has occurred.

In creating this scenario, we equate the general notion of interest with an interest in starting a conversation, i.e. the intention to start a conversation. We point out that a more general scenario would, of course, need more cues to disambiguate between other possible causes of this interest. Facial expression for example could be another very useful cue, and something we discuss more in Section 5.

4.3.1 Behavioural output

We use a hierarchical finite state machine to implement conversation initialisation. There are five high level states in the HFSM: *Monitor Environment* (ME), in which the agent attends to the environment looking for identifying other agents, *Grab Attention* (GA) in which the agent attempts to elicit the attention of another agent, *Passive Monitoring*(PM) which represents a discrete monitoring of the other agent without trying to attract their attention, and *Gauge Reaction* (GR) where an agent is actively sending signals and interprets received signals to decide whether it should commit to conversation. The final state is *Starting Conversation* (SC) which is the terminating state and handles the situation where both participants have successfully engaged in conversation. More details are available in [21].

Importantly, state transitions in the HFSM take place determined not only by the interaction goals of an agent, but also according to their perception and theories of the state and intentions of the other agent, based on its attentive behaviours. For example, an agent that wishes to interact with another will look at the other, but if it thinks the other has seen it but does not think the other is interested, it will not make more overt attempts to start conversation.

4.4 Evaluation

When constructing computational models that involve the synthesis of human perception of social signals for creating theories of the intentions of others in a human-like manner, it is vital to take into account how humans establish theories. Theories, by their very nature, are not based on totally reliable information, but rather based on the perceivers interpretation of events. In relation to work discussed here, which focuses on eye-gaze and direction of attention, user evaluation studies are of critical importance if low-level gaze direction and duration information is to be integrated into a less volatile, higher level representation. We conducted a number of evaluation studies in order to establish, along with other details, how different orientations of eye, head and body direction of an agent in a virtual environment may be interpreted by human users in terms of the amount of attention that they are paying and interest they may have in the user.

4.4.1 *Experimental Purpose and Design*

This experiment had a number purposes. The most fundamental of these was to verify that human viewers could in fact infer attention and interest from an artificial humanoid character when viewed in a virtual environment. There were also more precise goals for each study. First of all, a study involving static imagery to test the notion that the amount of attention that users perceive from an agent in a virtual environment is related to the manner in which three main body segments (the eyes, the head and the body) are oriented with respect to the viewer. This is inspired by research from ethology [11] suggesting that a similar situation may exist in nature, whereby the eyes, head, body and locomotion direction all influence the perception of the amount of attention that one has in the self. Our initial hypothesis for this study was that the attention perceived by the user would increase as the orientation of agent subparts were directed more towards to user, with the eyes as the biggest indictor of attention, followed by the head and finally the body. Secondly, from studies with a dynamically behaving agent, we aimed to test human perception of the intention of an agent to start interaction based only on their direction of gaze and locomotion. Our hypothesis was that increased directed gaze, locomotion and gesture behaviour would result in an increased perception of an agent that was interested in the viewer and open to or seeking interaction.

4.4.2 *Population and Apparatus*

A total of 21 participants engaged in a two-stage evaluation process. All participants were French computer science undergraduate students, between the ages of 19 and 25 and all therefore had a technical background. Two were female.

The demonstration and collection process was fully automated, although a regulator was on hand to observe correct following of experimental protocol and to answer questions. Participants were initially presented with an in-engine screen providing instructions in French, their native language. From this screen, the participants were then guided through the entire evaluation process via menus, which also prompted for results. For the dynamic study in the virtual environment, since the visibility of the agent's behaviours might be difficult for the viewer when the agent was far away due to the nature of the display equipment, we placed a magnified view of the agent in the top right corner of the display.

4.4.3 *Static Evaluation Study*

The first evaluation, which we will refer to as the static evaluation study, or *SES*, consisted of participants being shown a sequence of 25 static images featuring a humanoid agent standing in an upright posture. In each trial, body segments of the humanoid agent were oriented in varying positions with respect to the viewer: in some images, only certain body segments were visible. After viewing the image, participants were asked to select, using a slider, the amount of attention that they felt the agent was paying to them ("Attention which is being paid to me"), from a range of *NONE* to *A LOT*. The default position of the slider was a center point of the scale. Participants were then required to click a button in order to proceed to the next trial image. The ordering of the trials in the sequence was randomised for each participant.

Table 1. Results for the Static Evaluation Study (SES) in descending order according to the mean amount of attention that participants reported as perceived from the static images of the humanoid agent. The agent was segmented into three main parts: eyes, head and body. Direction was encoded as facing forwards towards the viewer (F), midway (M) and to the side (S).

	Eyes	Head	Body	Av. Rating	Std. Dev.
(a)	F	M	S	0.758	0.186
(b)	F	M	F	0.744	0.169
(c)	F	F	S	0.727	0.293
(d)	F	M	M	0.710	0.163
(e)	F	F	M	0.589	0.233
(f)	F	F	F	0.507	0.322
(g)	M	M	M	0.466	0.302
(h)	M	M	F	0.372	0.322
(i)	S	F	F	0.277	0.269
(j)	S	S	F	0.221	0.240
(k)	S	S	S	0.192	0.278

The averaged results of the static evaluation study for all 21 participants are summarised in descending order according to the amount of attention as perceived from static images of the humanoid agent (see Table 1). Values have a range of 0.0 to 1.0. Participants were asked to adjust a slider that indicated the amount of attention they thought the agent was paying to them in each instance: the slider ranged from *NONE* to *A LOT*. *NONE* has been mapped onto 0 and *A LOT* onto the value of 1.

In general, from Table 1, it can be seen that participants' perception of attention from the humanoid agent was highly correlated with the agent's eye direction: that is, when the agent was looking forward into the camera (giving the impression of looking at the user), participants rated it highly in terms of paying attention. Head direction correlated less strongly than eye direction, whereas the correlation between perceived attention and body direction was very low. These results adhere to the hierarchy suggested by Emery [11] and used in the current work (i.e. eyes > head > body).

The most surprising result of this study was the relatively low rating of 0.507 for Case (f) (Table 1), the situation where all of the body segments of the agent faced the user (Eyes: front, Head: front, Body: front). One would have expected this to be one of the highest ranked situations. A closer inspection of the results suggested two distinct groups of participants: the first marked the attention level highly, as expected, while the second marked it with a low average value. The lower rating group in Case (f) may have perceived less of a signal from the agent, e.g. interpreted it as as a blank stare, due to lower contrast between segment directions and thus, possibly, a less obvious signifier of attention.

4.4.4 *Dynamic Evaluation Study*

In the second evaluation, referred to here as the dynamic evaluation study, or *DES*, participants were shown a sequence of animations featuring a humanoid agent moving around in a virtual environment and making a range of behaviours.

In each trial, the agent started in a set position and walked along a predefined path. The first sub-segment of path (1a) was directed so that the agent would move closer to the viewer without walking di-

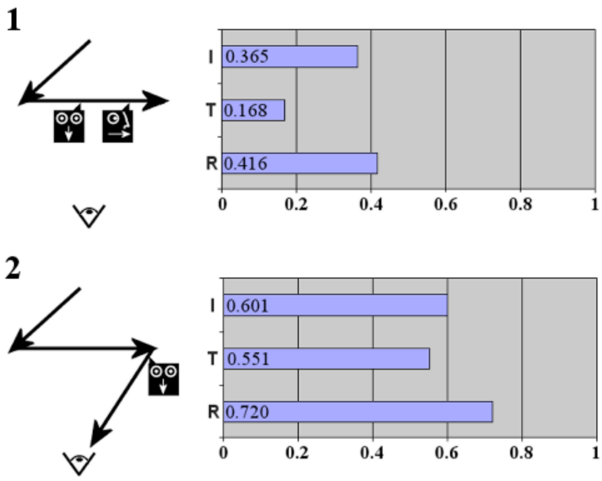


Figure 5. DES results for sample cases *1* and *2*. In each case, a top-down view of the path tagged with behaviour icons is illustrated on the left of the corresponding chart. Icons correspond to look at viewer and look away from camera respectively and are marked on the path at the time at which they occurred.

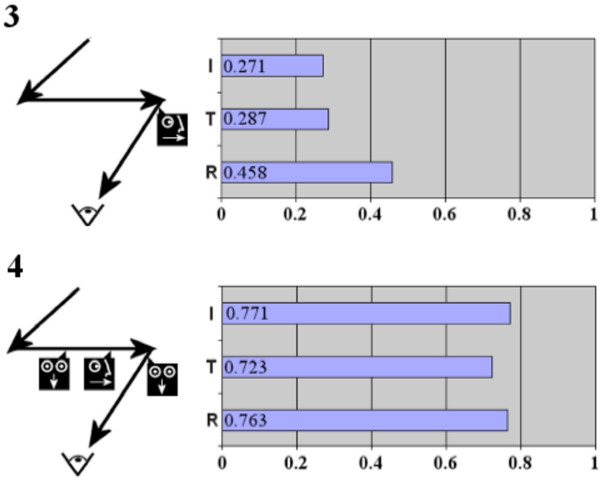


Figure 6. DES results for sample cases *3* and *4*. Each chart depicts the averaged *interest* (I), *want to talk* (T) and *would respond* (R) values as reported by all participants, out of a maximum value of 1.0.

rectly towards them. The second sub-segment (1b) was positioned so that the agent would walk perpendicular to the viewer. The second segment of path (2) was positioned so that the agent would have a locomotion direction directly towards the viewer. During trials, the agent would either walk along the first segment of path (1), or else both segments of path (1 and 2). A variety of gaze, locomotion and gesture behaviours were made during the trials: thus, trials differed in gaze (gaze at / not gaze at), gesture (wave / no wave) and locomotion direction (oblique / perpendicular / towards) with respect to the viewer.

The behaviours of the agent ranged from ignoring the user, to looking at them, walking towards them and waving. After each trial, the participants had to adjust a number of sliders indicating how they interpreted the actions of the agent in terms of e.g. the amount of interest it had shown towards them and whether they thought the agent had seen them. Participants then clicked a button to proceed to the next trial. Prior to the DES, participants were shown a test animation demonstrating the capabilities of the agent and what they were about to witness. We distinguished between two cases of an agent that was perceived to 'want to talk' to the participant (Figures 5 and 6: *T*) and one that was thought 'would respond' (Figures 5 and 6: *R*) to a talk request if made by the participant. The first case is suggestive of an agent that was perceived to be proactive in seeking interaction, while the second was a measure of the openness of the agent to interaction, but in a more passive sense.

A reference case was used that consisted of the agent walking along the path (similar to Case 1, Figure 5), but without looking towards the user at all. Comparing the reference case with Case 1, where the only difference is a brief glance at the user, users recorded an increased perceived level of interest from the agent (Interest: 0.18 → 0.36). In Case 2, where the agent did not change locomotion direction after glancing at the camera, participants had the impression that the agent was somewhat interested in them (Interest: 0.36), would be responsive to conversation (Would Respond: 0.41), but did not report that they thought the agent was actively trying to start an interaction (Want Talk: 0.16).

Cases 2 and 3 provide a similar situation in terms of interest (increase from 0.27 to 0.60) and would respond to an interaction request (0.45 → 0.71). Unlike the previous situation, there is also an increase in the perception that the agent wants to talk (0.28 → 0.55). These results seem to suggest that although the perceived openness towards an interaction is closely linked to perceived interest from an agent through gaze, the coupling between the perception of interest and the perception that the agent wants to actively start an interaction may require more overt cues signalling this intention on the part of the agent: in this case, a change of locomotion direction towards the user. This is again supported in a comparison of Cases 1 and 4: the agent glances at the user in both cases, but in Case 4, it then changes locomotion direction towards them. This locomotion change results in a large increase in user reports (Interest: 0.36 → 0.77, Want Talk: 0.16 → 0.72, Would Respond: 0.41 → 0.76), particularly the impression that the agent actively wants to talk. Case 4 received the highest ratings overall. Ethologically, this would also seem to be one of the most natural behaviours preceding the start of an interaction: from the user's perspective, the agent first walks perpendicular to them in the environment and may not appear to be aware of them. It glances over and becomes aware of them, makes a decision to engage in interaction with them, and so changes direction and walks towards them in order to start a conversation.

At a more fundamental level, it is evident from our studies that virtual humanoids have the ability to give users the impression that they are paying attention to them through their gaze, body orientation, gesture and locomotion behaviours. In addition, participants tended to report higher perceptions of interest, interaction seeking and openness overall from the agent when these behaviours were directed towards the user: In the SES, the case rated lowest in terms of the amount of attention perceived to be paid by the agent is that where all body segments are oriented away from the camera. Similarly, in the DES, the case rated lowest in all categories related to interest and interaction was the reference case: It was in this case that the agent did not look at, walk towards or gesture towards the user at all. More details about these evaluations can be found in [20].

5 TOWARDS AN EMOTIONALLY-SENSITIVE SYSTEM

Although the model presented in Section 4 has proven useful for obtaining enhanced social agents for situations involving attentive behaviours, it does not account for the important issue of emotion.

When we talk about *emotionally-sensitive* agents, we encompass a range of abilities that allow an agent to be able, not only to sense the emotions of another, but also react to them in a way that is appropriate for the type of agent being modelled. Such a reaction should, at the very least, acknowledge an emotional display of the other, even if it is just a matter of a quick glance. Such a movement could be thought of as a gesture saying 'you are of some significance to my goals' or 'I acknowledge your existence' or perhaps even 'I care enough to pay attention to your emotion'. This is the most basic level of emotional sensitivity, even though according to human norms, humans who only ever displayed such a response to the emotions of others might be considered to be very insensitive people. Yet, even this level of emotional-sensitivity is often neglected in current systems. More sensitive forms can then involve an internal alteration of ones own emotional state and an appropriate reaction.

Baron-Cohen has revised his mind-reading system to incorporate emotion and affective states [4]. As mentioned in Section 3, the revised empathising system now contains a component, linked again to the perceptual system, called TED, or The Emotion Detector. Its purpose is to deal with the recognition of affective states in others and it can built dyadic representations of the type *Agent-affective state-proposition* e.g. 'Mother-is unhappy'. TED thus allows the recognition of basic emotions. In addition, a component called TESS has also been added with the role of triggering an affective state based on the perception of an affective state of another through the TED. Therefore, this model is one of an observer propelled towards action, that is, to respond to the affective state of the other in an empathic manner.

We now discuss our thoughts on how these modules can be implemented computationally, and especially why they could prove very useful for augmenting elaborate social behaviours between autonomous agents.

5.1 A Computational TED

Of the two suggested additions, TED is perhaps the one with more substantial research already in place. For example, when working with real data, numerous works have considered how to obtain basic emotion information from the facial expressions of others, see [18] for survey. A more interesting prospect is how the data in the TED can be combined with other data to create a better view of more complex mental states of others. One interesting approach is to use a bayesian belief network and is that taken in [10] for inferring the mental state of others, such as *thinking*.

In our current case, since we are dealing with the recognition of emotions from other agents in the virtual environment, it allows us to simplify our processes considerably and maintain our attention on the holistic operation of the model. Although we will start by focusing on obtaining basic emotion [9] information in the TED, an interesting future prospect would be to consider how an appraisal-based system [25] could merge into this framework in order to produce variation in a dimensional model of the agents emotional state. Such a system would no doubt have to include information about the attention-related behaviours of the observed as well.

Although a computational TED is not enough alone to provide an emotionally-sensitive agent, it is a necessary requirement if the agent is to react appropriately. As mentioned earlier, one fundamental reaction can be the allocation of overt attention to the display of emotion by the other. As we will investigate next, TED can also be coupled with other detectors (such as the EDD), to provide a better idea of the intent of the other and also of what is happening in the environment.

5.1.1 Linking gaze, emotion and intent

An interesting question is how the emotional input from the TED can be linked into and used to form useful mental theories in the ToMM. Given the construction and purpose of our prototype model (Section 4), one very interesting research route involves the consideration of gaze as a disambiguating factor for establishing the source of what might have caused the emotional expression in the other. For example, TED could represent not only 'Mother - is unhappy', but also, 'Mother - is unhappy - *with me*'. This could be achieved either by feeding TED with information from EDD, or by integrating information at a slightly later stage, which would probably be our preferred option.

One particularly important application that we see for this is for the detection of threat. It is known that the amygdala, a region of the brain associated with emotional processing, is sensitive to gaze direction - brain imaging studies have found that a fear or anger facial expression may be interpreted differently by the viewer depending on gaze direction [1]. It is thought that, in this case, gaze acts to disambiguate the source of potential threat, so a directed-gaze angry face indicates the displayer as being the source of threat, while an averted fearful face may indicate a source of threat elsewhere in the environment. This has led us to distinguish a special category of behaviour that we refer to as being *directed*. A directed gesture is one that is made *at* somebody or something, for example waving at somebody to say hello to them. Our view is that emotional expressions (facial expressions for example) can be directed too, and in our model we will be pursuing the study of attention direction as a signaller of where or to whom the emotional expression is aimed at.

5.2 A Computational TESS

As mentioned, the purpose of the TESS is to provide an appropriate emotional reaction to the emotion of another, for example 'I am *horrified* - that you are *in pain*'. Baron-Cohen [4] defines *appropriate* as meaning that one cannot have a reaction such as 'I am *happy* - that you are in *pain*' when the TESS is functioning normally.

Our approach is likely to differ somewhat, as we would like our model to also be applicable to adult humanoid agents. This involves drawing more of a distinction in what is meant by the term *emotional reaction*: that is, to differentiate between how the TED changes the inner emotional state of the agent and how the agent expresses its emotional state to others. These may not always be the same. For example, to extend Baron-Cohen's example of a psychopath [4], a clever psychopath may indeed have the emotional reaction of feeling happy that another is in pain, but may choose to try to mask this with the emotional expression that they are sad instead, purely for social or beneficial purposes. While we certainly have no wish to try to create psychopathic agents, it would nonetheless be desirable to have a separation between emotional state and expression for other reasons, so that an agent can take other factors into account when it expresses itself. For example, to limit its display of happiness about a favourable event for it if it thinks the user may not be so happy about it. Furthermore, it would be desirable to create an expression module (see [17] for example) that accounts for both bottom-up (basic expressive reactions to the emotional stimulus) and top-down influences (attempted masking of ones true emotions for social or other purposes). In this way, TESS still has a role to play, but is only one of or is constructed of a number of other interconnected components.

A first computational prototype of this module could serve simply to copy and express the low-level or high-level characteristics of perceived emotions in order to provide an agent that imitates entities that it interacts with.

6 CONCLUSION

We have presented a prototype model and demonstrated and evaluated it in a conversation initiation scenario. The addition of simple theories about the intentions of the other to converse has provided for more complex social interaction to take place - that is, where an agent can take account of what it thinks the other knows when getting into conversation. One particular instance that we highlight is that the agent takes account of the goals of the other rather than just its own goals when attempting to open interaction, for the purposes of avoiding the social embarrassment of interaction with somebody who does not reciprocate. We believe work on this model will help illuminate complicated but interrelated notions for us, such as shared attention behaviours, engagement and empathy.

In relation to our ongoing work, the empathising modules offer a much-needed enhancement to the previous model. As well as a number of other possibilities that we have described, it allows for further disambiguation of the motives of the other as part of ones theory of their intent. Accompanied by a friendly facial expression, we may disambiguate the motives of a stranger paying close attention to us into the theory that they may just want to chat, rather than attack us. Such reasoning is still relatively basic in comparison to the human case, but nonetheless brings us somewhat closer to our goal of socially-capable agents.

Of course, this leads us again to mention that a vital aspect that should not be ignored is the context of the situation and how to reason about it. Our work is just a small part of this effort and is meant as complimentary to more in-depth reasoning approaches, which in turn can help to inform our model.

ACKNOWLEDGEMENTS

Thanks to Catherine Pelachaud for her invaluable advice throughout this work. This work has been funded by the Network of Excellence Humaine (Human-Machine Interaction Network on Emotion) IST-2002-2.3.1.6 / Contract no. 507422 (http://emotion-research.net/).

REFERENCES

[1] R. Adams, H. Gordon, A. Baird, N. Ambady, and R. Kleck, 'Effects of gaze on amygdala sensitivity to anger and fear faces', *Science*, **300**(5625), (2003).

[2] S. Baron-Cohen, 'How to build a baby that can read minds: cognitive mechanisms in mind reading', *Cahiers de Psychologie Cognitive*, **13**, 513–552, (1994).

[3] S. Baron-Cohen, *Mindblindness: An essay on autism and theory of mind*, The MIT Press, Cambridge, MA, 1997.

[4] S. Baron-Cohen, 'The empathizing system : a revision of the 1994 model of the mindreading system', in *Origins of the social mind: Evolutionary psychology and child development*, eds., B. Ellis and D. Bjorklund, The Guilford Press, (2005).

[5] S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, and I. Plumb, 'The "reading the mind in the eye" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism', *Journal of Child Psychology and Psychiatry*, **42**(2), 241 – 251, (2001).

[6] V. Carofiglio and F. de Rosis, 'In favour of cognitive models of emotions', in *Proceedings of the Joint Symposium on Virtual Social Agents (AISB)*, eds., L. Hall, D. Heylen, S. Marcella, C. Pelachaud, P. Wallis, and S. Woods, pp. 171–176, Hatfield, UK, (April 2005).

[7] C. Castelfranchi, 'Tom and bic: Intentional behavioral communication as based on theory of mind', in *Proceedings of the Joint Symposium on Virtual Social Agents (AISB)*, eds., L. Hall, D. Heylen, S. Marcella, C. Pelachaud, P. Wallis, and S. Woods, Hatfield, UK, (April 2005).

[8] G.O. Deak, I. Fasel, and J. Movellan, 'The emergence of shared attention: Using robots to test developmental theories', in *Proceedings of the First International Workshop on Epigenetic Robotics*, pp. 95–104, Lund University Cognitive Studies, (2001).

[9] P. Ekman, W.V. Friesen, and P. Ellsworth, 'What emotion categories or dimensions can observers judge from facial behavior?', in *Emotion in the human face*, ed., P. Ekman, 39 – 55, Cambridge University Press, New York, (1982).

[10] R. el Kaliouby and P. Robinson, 'Generalization of a vision-based computational model of mind-reading', in *the First International Conference on Affective Computing and Intelligent Interaction (ACII)*, eds., J. Tao, T. Tan, and R.W. Picard, pp. 582–589, Beijing, China, (October 2005). Springer-Verlag.

[11] N.J. Emery, 'The eyes have it: the neuroethology, function and evolution of social gaze', *Neuroscience and Biobehavioural Reviews*, **24**(6), 581–604, (2000).

[12] E. Goffman, *Behaviour in public places: notes on the social order of gatherings*, The Free Press, New York, 1963.

[13] A. Kendon, *Conducting interaction: patterns of behaviour in focused encounters*, Cambridge University Press, New York, 1990.

[14] H. Kozima, 'Infanoid: A babybot that explores the social environment', in *Socially Intelligent Agents: Creating Relationships with Computers and Robots*, eds., K. Dautenhahn, A. H. Bond, L. Canamero, and B. Edmonds, 157 – 164, Kluwer Academic Publishers, Amsterdam, (2002).

[15] A.M. Leslie, 'Tomm, toby, and agency: Core architecture and domain specificity in cognition and culture', in *Mapping the Mind: Domain Specificity*, eds., L.A. Hirschfeld and S.A. Gelman, 119–148, Cambridge University Press, (1994).

[16] L. Mol, R. Verbrugge, and P. Hendriks, 'Learning to reason about other peoples minds', in *Proceedings of the Joint Symposium on Virtual Social Agents (AISB)*, eds., L. Hall, D. Heylen, S. Marcella, C. Pelachaud, P. Wallis, and S. Woods, pp. 191–198, Hatfield, UK, (April 2005).

[17] M. Ochs, R. Niewiadomski, C. Pelachaud, and D. Sadek, 'Intelligent expressions of emotions', in *Proceedings of 1st International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 707–714, China, (2005).

[18] M. Pantic and L.J.M. Rothkrantz, 'Automatic analysis of facial expressions: The state of the art', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(12), 1424–1445, (2000).

[19] D.I. Perrett and N.J. Emery, 'Understanding the intentions of others from visual signals: neurophysiological evidence', *Current Psychology of Cognition*, **13**, 683–694, (1994).

[20] C. Peters, 'Evaluating perception of interaction initiation in virtual environments using humanoid agents', in *Proceedings of the 17th European Conference on Artificial Intelligence*, pp. 46–50, Riva Del Garda, Italy, (August 2006).

[21] C. Peters, 'A perceptually-based theory of mind model for agent interaction initiation', *International Journal of Humanoid Robotics (IJHR), Special Issue: Achieving Human-Like Qualities in Interactive Virtual and Physical Humanoids*, **3**(3), 321 – 340, (2006).

[22] D. Premack and G. Woodruff, 'Does the chimpanzee have a "theory of mind"?', *Behavioural and Brain Sciences*, **4**, 515 – 526, (1978).

[23] D. Pynadath and S. Marsella, 'Psychsim: Modeling theory of mind with decision-theoretic agents', in *In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1181–1186, (2005).

[24] B. Scassellati, 'Investigating models of social development using a humanoid robot', in *Biorobotics*, eds., Barbara Webb and Thomas Consi, M.I.T. Press, (2000).

[25] K.R. Scherer, 'Appraisal considered as a process of multi-level sequential checking', in *Appraisal processes in emotion: Theory, Methods, Research*, eds., K. R. Scherer, A. Schorr, and T. Johnstone, 92–120, Oxford University Press, New York and Oxford, (2001).

Anticipatory coordination through action observation and behavior adaptation

Michele Piunti[1], Cristiano Castelfranchi and Rino Falcone[2]

Abstract. To establish coordination at a cognitive level, we need to come through to some of the strict assumptions of the traditional deliberative agents. We also need to cope with some of the real world boundaries, where, for instance, knowledge and perception are affected by uncertainty and message exchange as most direct form of subjective coordination may not be reliable everywhere. Intentional, subjective aspects of coordination may concern prediction mechanisms (i.e. future state representations, mind reading), but also true cognitive expectations that agents should exploit to reconsider their intentions, and their use in reading other agents goals (for better achieving their own). On the basis of the cognitive theories of social actions and behavioral implicit communication, we here propose an observation based approach providing agents with explicit anticipatory coordination capabilities in order to exploit signs coming from other agents and, contextually, adapt behavior in anticipatory terms. Pro-activeness, adaptiveness, opportunism come out from the means-end reasoning of individual actors: agents embedding such adaptive skills are leaning to make coordination as an emergent property of their interactions.

1 INTRODUCTION

The research behind this work wishes to provide artificial agents engaged in real world applications with anticipatory coordination abilities. We here refer to subjective approaches to coordination[3], meaning those aspects of the activity of an agent specifically devoted to deal with a dynamic environment and its social interferences. In these contexts, agents continuously cope with opportunities to exploit and threats/obstacles to avoid. To coordinate herself with a give event or act, an agent has to perceive or foresee it thanks to some perceptual hints, index or sign. In real world applications most direct form of subjective coordination through message exchange is not universally serviceable. Sometimes agents may not desire to exchange information (i.e. hostile agents), otherwise also cooperative agents may be reluctant to send explicit messages due to heterogeneous models and technologies, environment and resource constraints. Direct messaging further introduce limitations and costs, namely weight for additional equipments and transmitters, bound of communication range, unreliability of services, need for standardized protocols.

On the contrary, we argue that coordination between agents is not necessarily based on explicit communication. An action performed by any one agent potentially updates the perception (and the epistemic states) of other agents, thus observing and interpreting the world where agents are pursuing their goals is an intrinsic opportunity for coordination activities. Beliefs about other's mental states are also a result of the process of interpretation of other's behavior, that can be considered as the observable *sign* for his internal state[7]. We guess one of the main functions of observation in agent living in a common world populated by other agents is coordination, while one of the main form of coordination is observation-based. Indeed, just behavior without any modification or any additional signal or mark can be exploited as a premonitory sign, thus recognition capabilities make possible for an observer to predict future actions of an observed agents. By so doing, recognizer agents should exploit these capabilities to conceive an explicit form of expectation.

In order to enhance coordination for social tasks, several coordination techniques have been developed, including those based on social conventions and norms [26], decision and game theoretical strategies [13, 14], stigmergy infrastructures [1]. Several techniques for goal and plan recognition have been proposed and applied to different application domains. The idea to exploit observation to acquire coordination hints is not new in literature [17, 12]. Less effort has been given to goal directed behavior adaptation on the basis of the anticipated outcomes of the interactions. The inferential knowledge carried out by intended plan recognition mechanisms makes possible to ascribe mental states (goals) to others: in so doing, agents are enabled to anticipate actions performed by others, thus to reconsider their intentions and/or exploit those actions as an enlarged, exogenous repertoire of actions at disposal [10]. To establish coordination at an intentional level, the interferences between agent activities have to be endogenously valued as *positive* (A2's actions realize A1's goals or create opportunities) or *negative* (A2's actions create obstacles to A1 or thwart A1's goals). To do this, a recognizer agent has to subjectively classify the expected external outcomes: valuing these expectations as positive (or negative) is made according to their contribute (or determent) to the recognizer ongoing purposes and mental states (e.g. Goal, Beliefs).

In this work we propose a design model for cognitive agents endowed with the ability to *predict the outcomes* of other agents actions, to *build a model* of future events and to *react in advance* according to these expected events. We do approach the problem by enabling agents with *mind-reading* abilities: in so doing we introduce an enriched perception module used for observing and recognizing signs, actions, and practical behavior. We further provide agents with a mean for intending other agents in form of mental states. As we show in the next sections, the subjective utilities for the expected

[1] Institute of Cognitive Sciences and Technologies - I.S.T.C. - C.n.r. and DEIS, Università di Bologna - Alma Mater Studiorum, Italy, email: michele.piunti@istc.cnr.it

[2] Institute of Cognitive Sciences and Technologies - I.S.T.C. - C.n.r., Italy, email: {cristiano.castelfranchi, r.falcone}@istc.cnr.it

[3] In the context of MAS, it rely on the viewpoint of the individual agent that can perceive and understand the actions of its peers. [24] defined subjective and objective coordination respectively as an endogenous, psychological capability for coordinating agent vs. exogenous, infrastructural system to coordinate agents.

outcomes produce a fully represented expectation that can be used to rationally change behavior, and to cause avoidance or exploitation of alternative courses of action.

The rest of this paper is organized as follows: in section 2 we define the bounds of traditional cognitive architecture in dynamic, social scenarios, showing how anticipatory competencies may overcome a set of restrictive assumptions; in section 3 we describe the architecture for anticipatory agents able to exploit observation and plan recognition as building block for anticipatory behavioral coordination; in section 4 we describe a case study through an experiment; in section 5 we conclude with final discussions.

2 FROM GOAL-DIRECTED, DELIBERATIVE AGENTS TO ANTICIPATORY, INTERACTIVE AGENTS

We refer to true cognitive, *goal-governed* and deliberative systems, able to manage subjective, internal representations for beliefs and goals[4]. Traditionally deliberative agents operate to reach a desired state of affairs from the current state by chaining their environment through a given set of actions and plan operators[5]. The wide adopted BDI-based architectures [23] focalize systems in deliberation among set of goals and means-end analysis[6] between alternative courses of actions, but let intention making and execution of plans in a functional, even purely reactive form. In this sense, deliberative agents process their information reacting in a procedural way: they choose in repertoire the plan to execute according to filtering of conditions (belief formulae, utility functions, priorities etc.), whilst the available plan library is handcrafted at design time.

Early implementations of BDI-like systems operate according to the following restrictive assumptions [22]:

1. **Static world assumption**: the world is not changing during the reasoning process.
2. **Infinite resource assumption**: even if the world is changing the agent has sufficient resources to appraise all the relevant changes and consequently revise the belief base. The agent can also plan faster than the rate at which the world is changing leaving the plans still relevant.
3. **Complete knowledge assumption**: the agent has the capacity to perceive the complete state of itself and its environment and the information describing the environment results consistent and without noise at each point in time.
4. **Determinism assumption**: each planned action will completely realize the expected outcome.
5. **Single agent assumption**: actions performed by other entities do not influence agent activities. There are no other agents to aid or thwart agent plans.

In addition, coordination competencies are generally based on the reaction upon a direct perception of some events or act (*reactive coordination*) and often treated along with the general problem of intention reconsideration [18, 25].

2.1 Breaking assumptions through anticipation

Multidisciplinary convergencies indicate agents with anticipatory capabilities be more effective to overcome a larger set of real world requirements. We define *anticipation* as the ability to coordinate the behavior with the future: more formally, anticipation enable agent to *react in advance* (at an instant t) to an event (or to a world state) that will be realized at $t + t'$. Practical anticipatory behavior should be exploited on the basis of the knowledge of the current situation *but also* on some form of expectation about future states and events. Given this, the behavior does not only depend on past and present, but also on some knowledge about the future: [3] introduced anticipatory agents entertaining *expectations* as mental representations of the future. Expectations enable agents to be anticipatory just by working on them, for virtually exploring alternatives, opportunities, events, results. Expectations are not simply predictions neither belief on the future: they are given as axiological anticipatory mental representations, also endowed with *valence* against some concern, drive, goal of the agent. As in [20], we point out that, in a cognitive system, expectations play several important roles: i) precede and control the execution of actions. ii) bias sensory processing (attention, active perception) and resource allocation. iii) are used to bias goal selection and intention reconsideration.

Furthermore, in Multi Agent Systems agents play in a shared environment and have to operate in a world eliciting interferences, where the action of an Agent $A2$ could affect the goal of another Agent $A1$. We guess that modeling mental states of individual agents allows interaction with the counterpart in the minds of other agents. Our challenge in this work is to enable the expectations about A2 actions to be used by A1 as a sign, an help in deciding to react in advance, anticipating and exploiting events and outcomes performed by the other. By so doing, we design agents able to interact following an *anticipatory coordination*, based on the anticipation of interferences, opportunities and dangers.

2.2 Interaction for goal directed agents

Interaction between agents may result at a certain grade of cooperativeness, competitiveness or in some grade in between: it may result positive or negative for agents that are helped or damaged, favored or threatened by the (effects of) the actions of the others[7]. In the cooperative case, agents are more inclined to behave pursuing joint goals: on the one side they intend to exploit actions performed by others for their purposes, on the other hand they want to help each other to achieve common goals. On the contrary, in the competitive interactions, agents intend to thwart the others: on the one side they show avoidance of undesired outcomes, on the other side they perform hostile behaviors in order to prevent adversarial threatful purposes.

[7] noticed a deeper form of interaction in attempting to influence the behavior of the others by changing their mental states. In observable environments actions acquire a communicative function by preserving their practical end through their long term effects and modification in world states. By considering each action with its necessary world contexts in terms of preconditions and outcomes, A1 may induce A2 to abort her behavior by giving misleading signs or removing the necessary conditions, or may persuade A2 to do something by intentionally signalling opportunities or creating the

[4] This classification makes sense against the category of merely *goal-oriented*, functional systems, without any internal anticipatory representation for the goal of the action, where the teleonomic character of the behavior is in its adaptive function (e.g. managed by some learning algorithms). This class of systems does define no native support for dealing with the future through representations of future states.

[5] That state also indicates the 'goal state', more precisely the representation of the goal indicating the satisfaction of a subjective desire in a future state.

[6] Deliberation is the process by which agent select the goal to be pursued; means-end is responsible to compose plans (the means) in order to achieve the previously adopted goal (the end).

[7] Notice that these notions can meaningful be applied only to systems endowed with some form of goal, where the effects of the action of an agent are relevant and impact on the goals of another.

necessary pre-conditions for A2's actions. To this end, A2 can intentionally change A1's mind through implicit communication via stigmergic traces, long term physical outcomes, environment modifications. Hence, A1 may not coordinate only by reading A2's mind (i.e. perceiving her behavior during its performance) but can exploit other post-hoc traces and outcomes of it in observable changes of the environment [27].

From the viewpoint of A1 interfering with A2, there are two strategies:

1. To adapt hers own behavior to A2's behavior, in order to exploit positive interferences (or to avoid negative ones);
2. To attempt to change A2's behavior by inducing A2 to do what she needs or to abort activities damaging A1.

Tab.1 distinguishes four different alternatives for anticipatory coordination. The first rows shows the cases of behavior adaptation: in cooperative (positive) coordination, A1 changes her (practical, purposive) plan in order to profit by a favourable circumstance; in competitive (negative) coordination, A1 is aimed at avoiding a threat. The second row shows the cases of direct influence by changing mental states of the other: A1 may induce A2 to abandon her threatening goal in order to avoid some risky effect or may persuade A2 to pursue some action in order to obtain its profitable outcomes.

	Competitive (Negative Interference)	Cooperative (Positive Interference)
Adapting behavior	avoid adversary activities	exploit teamwork activities
Changing other's mind	misleading signs, stigmergic traces	collaborative signs, stigmergic traces

Table 1. Anticipatory coordination holds to different effects according to the type of interaction between the involved agents.

2.3 Behavior as 'sign' for Anticipatory Coordination

Behavioral Implicit Communication theory [4, 5] introduces practical behavior as an important form of contextual communication between agents, without explicit messaging, neither direct speech acts. In strong BIC, agents (sources) behave intentionally with the additional motivation to make others (addressees) understand their purposes, i.e. to capture some meaning from implicit messages and, consequently, change their minds.

As for the adaptive strategy, we here refer to a weakest awareness between agents: on the one side they know to be monitored by others but do not ascribe an additional motivation in doing actions also for being recognized; on the other side, they have the goal/plan of interpreting observed behaviors, to coordinate with them and anticipate events. We present a computational model for coordinating with other predicted behavior, thus ignoring, for the moment, the possibility to induce changes in others behavior. The first layer of our design model requires the observer to *perceive (or infer) interferences*. This can be made through general plan recognition techniques:

- As in most plan recognition assumptions, agents refer to an internal knowledge and continuously match perceptual hints with it, in order to recognize other agent actions and plans.
- Through plan recognition mechanisms, agents attain *signification* (namely the semiotic ability to "ascribe sense" to the observed behaviors) and infer expectations on actions and world changes performed *by others*.

The second layer requires to *adapt behavior in anticipatory terms*, by avoiding threats or exploiting opportunities. Agent changes her own plan (sub-goal) and produces a new plan which is based on her beliefs (predictions) about the goal of another. To do this, she uses a further model for evaluating expectations and reconsider intentions:

- Agents *evaluate* positive and negative circumstances. Evaluating enable agents to read the world (i.e. actions performed by others) in terms of *positive and negative expected outcomes*.
- Agents reconsider their intentions and mental states on the basis of the new (valued) expectations.

By so doing, expectations become true representations of the future, upon which agents may concern, deliberate, reason and reconsider their plans, thus coordinating their behavior with the not *yet* existent.

Adapting behavior by working on future states elicits two main kind of appraisal. In the positive case, the agent anticipates an unexpected *help*: she can remove from the planned workflow the action that will be executed by others (agent A1 exploits A2's action, intentionally delegates and relies on it [11]). In the negative case, agent A1 anticipates an unexpected *determent*: to economize resources, she has to reconsider the ongoing intention, aborting the current action and adopting an alternative one (if present). Notice that the use of plan recognition methods introduces uncertainty in the reasoning process (coming from incomplete knowledge and errors in observation evidences, risk evaluation, learning processes etc.).

3 AGENTS AND PLAN RECOGNITION MODELS

In the following sections we present the architecture, including the plan recognition module, and we describe the reasoning process for the anticipatory coordination.

3.1 Design

As for the agents kernel we adopted the Jadex engine [2], a multithreaded BDI framework leading to loosely coupled Beliefs, Goal and Plans representation, including their mutual relations. Jadex deliberation is driven by the evaluation of logic formulae (put in form of Belief formulae) and arcs of inhibition between goals to dynamically resolve their priority. The sensor component directly gets data from the environment simulator: when an entity is sensed, its symbolic description is provided by a preceptor filter and then is used for belief revision (Fig.2). For simplicity, we assume that visual information retrieved from the environment simulator and symbolic information handled by sensor are given at the same level of representation.

A Mental States component is used to manage working memory, to allocate configuration of epistemic resources and to express attitudes and bias towards the actual state of affairs. We define Mental States through a set of related behavioral and mental changes increasing agents opportunism and proactiveness towards the environment changes. By using a functional approach, we have further defined some important roles that these affective states play for anticipation (for more details see [20]).

3.2 Plan Recognition

For the plan recognition mechanism, we assume a shared symbolic representation of purposive actions through hierarchical plans and we introduce a background process in perception filtering module (Fig. 1). Plan representation, used to match perceptions, assumes the

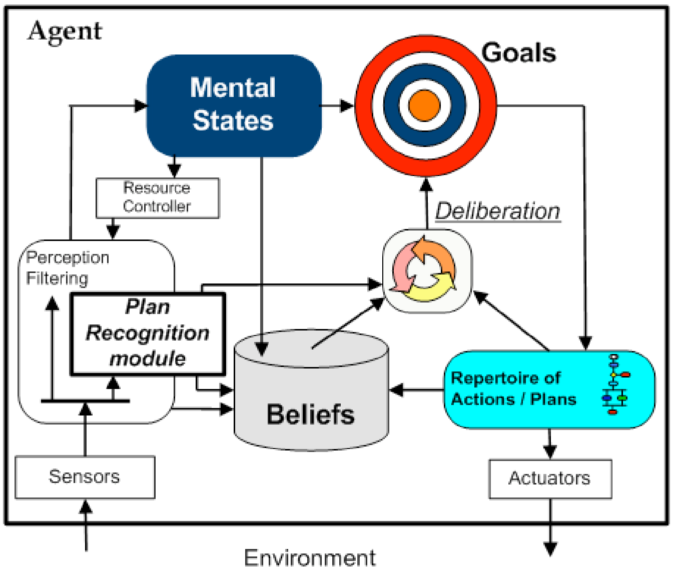


Figure 1. Agent architecture includes a BDI core.

role of the subjective belief about the preferences and the practical behavior usually performed by others.

It has been argued that plan recognition problem can be treated as the general problem of abduction. In this sense, an observer makes hypothesis following a diagnostic approach (observe action ”A” to deduce a goal ”G”). Thus, for a given set of actions, observer matches perception with the internal representation of plans to discover its best explanation. Perception filtering (Fig. 1) uses a Probabilistic Horn Abduction meta-interpreter [21] implemented within a tuProlog engine [9]. Once the prior probabilities are given, PHA calculates the list of the ’best explanations’ ordered by their crescent likelihood. The output of the recognizer is in domain of probability: the execution model manages the confirmation (failure) of the observed actions and provide the confirmation (failure) of the hypothesis. This directly results in reinforcing (inhibiting) the probability of that hypothesis (for more details on these techniques, see [8, 15, 12]).

3.3 Anticipatory process

Representation of plans is given in a single root, directed, acyclic connected graph, where roots indicate top level goals and leafs denote self-contained actions (plan steps). At any given time, the observed agent is assumed to be executing a plan decomposition path, from root to leafs, through plan tree branches (Fig. 2). We introduce:

- A repertoire of actions $\mathcal{A} = \{a_1, a_2, ..., a_A\}$ that constitute agents practical behavior as shared knowledge between observer agents. We assume plan representation used for recognition fully consistent to the plan library of operators used for practical behavior.
- A set of world states $\mathcal{S} = \{s_1, s_2, ..., s_S\}$ can be used to evaluate local background and world contexts. Notice that first order logic formulae upon $\mathcal{S}$ also constitute the preconditions for the execution of actions and for the activation of goal and plans.
- A set of outcomes $\mathcal{O}_a = \{o_{a_1}, o_{a_2}, ..., o_{a_A}\}$ indicating the world state as it is assumed to be *after* the execution an action. Notice that $\mathcal{O}_a \subset \mathcal{S}$, where each outcome indicates the expected effects of the related action in repertoire.
- A prediction function $\pi : \mathcal{A}^n \times \mathcal{S} \times \mathcal{T} \rightarrow \mathcal{A} \times \mathcal{O}_a \times \mathcal{P}$ that expresses observer’s hypotheses that, given at an instant t, n observed evidences for actions in $\mathcal{A}^n$, with the world context in $\mathcal{S}$, a certain action will be performed by an observed agent at $t + t'$, with the respective outcome in $\mathcal{O}_a$ and a probability in the distribution $\mathcal{P}$.

For an observer agent, plan recognition process provide the prediction of the next action performed by an observed agent. To this end, it refers to two sources of information: we do assume for each performed action in $\mathcal{A}$ an associate tuple of conditions on its observable features; observer agents further relate these features to some clarifying contextual world states in $\mathcal{S}$ (as noticed in [15], the use of world states significatively helps to disambiguate situations and reduce the overall complexity of the process). Given prior probabilities on plan branching, as they are reported in plan representation as meta-belief, π introduces a grade of (un)certainty in observer’s prediction. Hence, when considering what goal the observed agent might be pursuing, PHA meta-interpreter provides the best (most likely) explanation in terms of recognized goals, also evaluating the world state (in $\mathcal{S}$). When allowed by world constraints and observability, agent’s perception filtering *observes actions* performed by others and relate it to the *world context*, translating the data stream from sensors in symbols simultaneously referring to the prior knowledge of plans. By matching perception with the internal representation, a PHA-based mechanism provide *concurrent hypothesis*: observation process persists until the set of evidences in $\mathcal{A}^n$ become sufficient to disambiguate the corresponding goal: once the best explanation overcomes a fixed threshold, observer agent shapes an *expectation*, by balancing the observed predicted goal with own purposes. By so doing, observer *appraises and gives a subjective value* to the expectation: in positive terms, if the expectation is due to positive interferences (i.e. helps the pursuing of her goal); in negative terms, if the expectation is due to negative interferences (i.e. agent anticipates threats, obstacles).

In the second phase, observer agent adapts the behavior by reconsidering her intentions. We assume:

- A repertoire of *counteractions* $\mathcal{CA} = \{ca_1, ca_2, ..., ca_C\}$ that can be related to the observer goals and carried out to react to the prediction given by π.
- A set of outcomes $\mathcal{O}_{ca} = \{o_{ca_1}, o_{ca_2}, ..., o_{ca_C}\}$ indicating the expected effects for each counteraction.
- An *outcome function* $\varphi : \mathcal{CA} \times \mathcal{S} \times \mathcal{T} \rightarrow \mathcal{O}_{ca} \times \mathcal{P}$ that returns the probability for realizing the outcome of the counteraction ca_i (performed instead of a_j) when the actual world’s state is in $\mathcal{S}$. It models the uncertainty and the confidence of the observer in deciding which counteraction to take respect to the determinism of its outcome.
- An *utility function* $\upsilon : \mathcal{O}_{ca} \rightarrow \mathcal{U}$ giving the *utility value* of a certain outcome as an heuristic composition of subjective importance and desirability of the outcome, thus it is strictly related to the ongoing goal of the observer. For the observer agent, utility measures the desirability of any given outcome. Its value can be related to different domains (i.e. game-theoretic, normative) and coupled with different measures as perception of risk, urgency etc.

More formally, let h_j be a tuple $\langle a_j, s_j, t_j \rangle$: given the above definitions, $\pi(h_j)$ is the probability (provided by the plan recognition module) of a certain hypothesis j, $\varphi(ca_i, s_j, t_j)$ the confidence on the expected outcome for the counteraction ca_i (the probability that the counteraction will have its intended outcome), and $\upsilon(ca_i)$ the expected utility (given in decision theoretic account, in case of success of the respective counteraction). Agents select the reconsidered action to take by comparing, for each counteraction ca_i, the following expression:

$$\pi(h_j) \times \varphi(ca_i, s_j, t_j) \times \upsilon(ca_i) \qquad (1)$$

The above expression anticipate the effects and the subjective utility of a counteraction to take, given the anticipated effect of an action performed by the other. By so doing we do introduce subjective expectations in terms of agent's native epistemic states (Beliefs) and motivational states (Goals) (Fig. 1).

In other terms, agents adapt their plans by managing an *expected degree of adequacy* for counteractions in repertoire. Its value is a composition of an *epistemic state* (an uncertain, graded belief) and a motivational state (a graded utility and a subjective importance for the counteraction ca_i realizing a certain goal). By selecting the proper counteraction, agents may take advantage of the anticipated events, enhancing opportunism and pro-activeness, or decide to avoid them, by abandoning their activities and saving resources for alternative pursuable goals.

4 EXPERIMENT

In order to test different architectural solutions so that different strategies can be significantly compared, we engaged agents in a foraging task $T\langle LOI_n, V_n, R_L, \mathcal{A}, \mathcal{S}, r, S_r, s, \mathcal{D}\rangle$ in a 2D environment. The scenario presents a set of n Location Of Interest (LOIs) and requires a group of agents $\mathcal{A}$, each with adaptive sensor range r, sensor rate S_r, speed s, to find n types of valuables V_n, pick up them (one at a time) and bring back to the repository location R_L. Sentry agents $\mathcal{S}$ have the goal to guard LOIs and hinder agent foraging. Each valuable type V_X is coupled with a respective location LOI_X.

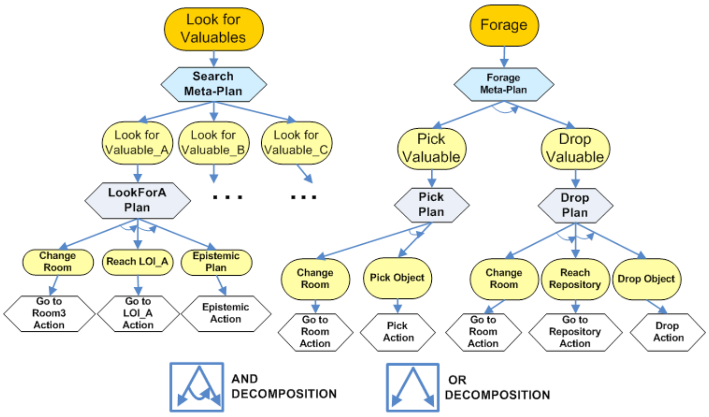


Figure 2. Plan representation used to match observations: it shows practical behavior and purposive actions for agents engaged in the foraging task.

Valuables are dynamically generated by the environment simulator close to the respective LOI, according to a probability distribution $\mathcal{D}$. Environment also present a layout of walls and doors creating room, corridors and pathways. Agents do not have an a priori knowledge of the distributions and use a library of paths and plans to move between locations.

Fig. 2 shows representation of hierarchical plans for foraging agents purposive behavior. Top level nodes (*Look for Valuables* and *Forage*) are expanded into sequences of lower level nodes, each of which is further expanded into yet lower level nodes. Thus, single plans are not just a sequence of basic actions, but may also dispatch sub-goals. The leaves of the plan structure form a non-hierarchical plan of practical actions that agents execute and observe themselves. Notice that, according to world constraints (i.e. wall, obstacles, sentries), goal/plan hierarchies may result with interleaved sequences of leaves and generate interleaved sequences of actions.

In our experiments we use three kind of valuables and three associated LOI ($n = 3$). Sensor component directly gets data from the environment simulator: when an entity is sensed, its symbolic description is provided by the simulator and the preceptor module filters it for belief revision and further reasoning processes. For simplicity, we assume that both perception data and symbolic information handled by the filter are given at the same level of representation. Foraging agent's plan knowledge (used by recognizer) is built upon internal Prolog representation. Intention reconsideration and re-planning processes are triggered by the activation of an hypothesis, namely when $\pi(h_j)$ overcomes the corresponding threshold: the process of valuing is managed at meta-level reasoning, with a meta-plan, by which the observer evaluates *on-line* the various available options.

Tab. 2 shows, from the point of view of the single agent A1 and for each sequence of observed actions and world contexts, the set of options in repertoire. Each option is a counteraction and encapsulates the respective confidence of success (due to indeterminism) and the subjective expected utility (in case of success).

In the second row of the table we show the case when agent A1 receives the evidence that agent A2 is transporting a valuable Obj_x, while the context is that A1 is looking for the same Obj_x. In this case, the time further devoted by both agents in looking for the same valuable would be wasted but A1 provide an explicit counteraction to save resources and optimize global behavior. An internal signal (from A1 perception filtering) indicates $\pi(h_j)$ is overcoming the fixed threshold: it triggers the meta-level reasoning process where confidences φ and utilities υ of counteractions in repertoire are evaluated. From a cooperative perspective, A2 not only has the goal to transport the valuable, but also the goal to make A1 aware of something: although she is not sending an explicit message, he has the goal of changing A1's mental states, updating her beliefs in order to modify behavior. In this case, the first counteraction to drop the ongoing search is taken because of its optimal expectation, hence A1 will spend her resources to look for a different kind of valuable, namely Obj_y near LOI_y (see Fig. 3).

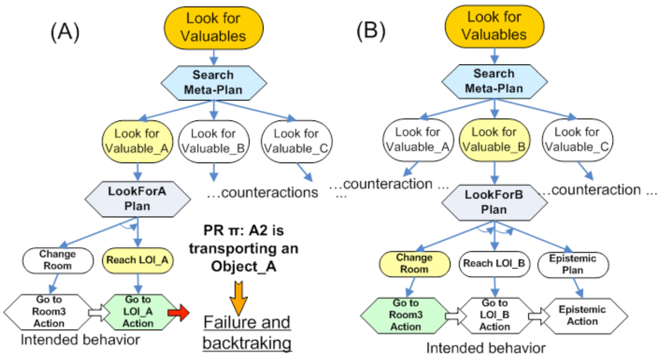


Figure 3. Adaptive Behavior: when A1 recognize A2, she receives an internal event (from Plan Recognition module), and breaks the current plan (A). The selection of the alternative course of actions (B) is driven by the evaluation of expectations for each counteractions in repertoire.

Along experiments, default values for $\varphi(ca_i)$ and $\upsilon(ca_i)$ are given in fuzzy terms, from ZERO [0.0] indicating absence of confidence

Observed action	Context World state	Options (ca_i) Repertoire of counteractions	Confidence of counteractions (φ)	Utility of the outcome (υ)
A2 approaching Obj	A2 closer to Obj	Abort Speed up Persist	MAXIMUM [1.0] LOW (0.3) MAXIMUM [1.0]	LOW [0.3] HIGH (0.8) ZERO [0.0]
A2 transporting Obj_x	A1 look for Obj_x	(Drop search Obj_x and search Obj_y) Persist	MAXIMUM [1.0] MAXIMUM [1.0]	MAXIMUM [1.0] ZERO [0.0]
A1 and A2 approaching the same Obj	Sentry close to A1	Distract the Sentry Abort Persist	LOW [0.3] MAXIMUM [1.0] ZERO [0.0]	MAXIMUM [1.0] LOW [0.3] ZERO [0.0]

Table 2. Intention Reconsideration through on-line evaluation of hypothesis and counteraction selection (agent A1 observes and anticipates agent A2). Confidences and utilities are given in fuzzy terms between ZERO [0.0] and MAXIMUM [1.0].

and utilities to MAXIMUM [1.0] indicating full utility and confidence value. Notice that agents evaluate also the hypothesis to remain committed and persist without adopting new intentions.

Belief thresholds strongly affect agent performances with space, time and activities trade offs. As in [17], the use of a belief-net may introduce learning mechanisms to adjust thresholds, confidences and utilities during the task. Thus, agents that become aware to act in particular environments (i.e. more or less risky) can adopt different strategies simply by tuning their values: by changing utility function may result in different agent personalities (e.g. individualist-autonomous, cooperative-collaborative), by changing confidence function agents become more or less self-confident etc.

5 DISCUSSION AND FUTURE WORKS

In this work we introduced anticipatory agents able to reconsider intentions on the basis of the expectations shaped on other agent recognized goals. The model enables to recognize other agent behavior as a BIC message and further provide abilities for signification and evaluation of related expectations. It further introduces noticeable properties for cognitive interaction:

- Coming through some lacks of the traditional deliberative architectures.
- Exploiting plan recognition mechanisms to really enhance pro-activeness and opportunism.
- Implementing a simplified approach to BIC which allows a wide spectrum of coordination issues to be modeled without relying on speech acts.

By introducing the ability for *intention reconsideration on the basis of expectations*, the model directly elicits anticipation and adaptivity to indeterminism, also allowing a strong subjective social interaction. Agents embeds adaptive capabilities to make anticipatory coordination an *emergent* property of the interactions: sociality is let emerge from the action and intelligence of individual agents.

Our experiments show that forms of silent, anticipatory coordination result in low cost, low complexity, highly effective mechanisms for coordination of agents with finite resources. The symbolic plan recognition engine, based on PHA, is very efficient and can serve concurrent hypotheses hence is able to predict agents pursuing multiple goals, namely interleaved plans. From a behavioral perspective, enhancements are in terms of pro-activeness, situated, real-time adaptivity to complex tasks. From the reasoning perspective, the model helps to disambiguate uncertainty, also providing strong adaptive means-end processing.

We guess this kind of architecture may contribute to the design of self-organizing/emergent societies, where virtual agents interact according to cognitive paradigms like *trust, reliance, delegation* [11, 6].

5.1 From simulation to real applications

We have made a series of assumption to simplify the domain. Moving from simulations to real applications, a series of key issues remains open. Firstly, the design of abstract actions to be recognized implicitly places the problem on the definition of the heuristics for the (reverse) process of recognizing their features: we define our representation as a series of abstract plan in first order logic terms, but to define the granularity of a real action may not be so obvious. We further have supposed complete plan representation handcrafted by the designers: in real-world scenarios this may result an intractable problem, due to complexity of tasks, heterogeneity of agents and multiplicity of their interactions. In addition, incorporating unknown goals and plans in plan representation is tractable only where the domain complexity is low [16].

Secondly, observing ongoing actions in real world application results a more complex task than we supposed: many actions have complex multi featured observable features, rather than few atomic features. Agents should embed components to resolve information processing form sensors to the internal symbolic representation. Furthermore, some of the features to observe may be intermittently lost due to noise or sensory failures. We assumed each action logically revealed without taking into account the information about its duration.

Finally, the computational costs of overwatching (matching observation against all possible actions performed by multiple agents and world contexts), may introduce overhead and serious problems in agent with finite resources.

In simplified domains (e.g. web applications), similar mechanisms for signification and plan recognition can be embedded with a different perspective, by utilizing hybrid approaches and smart infrastructures for a more *objective* coordination [24]. According to this paradigm, objective coordination is induced in MAS by means of ad-hoc abstractions of *coordination artifacts* [19] that may mediate interactions and provide coordination services. Coordination artifacts could be engineered to support previous knowledge of the plans used in the application domain and to dislocate (and automate) observation activities. As showed by many studies (i.e. [27]), this kind of infrastructure alleviates complex burdens for the involved agents: they can refer to the provided services in an uncoupled way and then decide autonomously by evaluating utilities to ascribe to the intentional options.

ACKNOWLEDGEMENTS

The research of this paper has been funded by the EU Project MindRACES within the Cognitive Systems area: Contract Number: FP6-511931. (www.mindRaces.org). We would like to thank aliCE[8] group at University of Bologna and Jadex[9] group at University of Hamburg for their components and useful support.

REFERENCES

[1] R. Beckers, O.E. Holland, and Deneubourg J.L., 'From local actions to global tasks: Stigmergy in collective robotics', *Artificial Life IV*, 181–189, (1994).

[2] L. Braubach, A. Pokahr, and W. Lamersdorf, *Software Agent-Based Applications, Platforms and Development Kits*, chapter Jadex: A BDI Agent System Combining Middleware and Reasoning, Birkhuser Book, 2005.

[3] C. Castelfranchi, 'Mind as an anticipatory device: For a theory of expectations', *In BVAI 2005*, 258–276, (2005).

[4] C. Castelfranchi, *From Conversation to Interaction via Behavioral Communication*, chapter Theories and Practice in Interaction Design, 157–179, Erlbaum, 2006.

[5] C. Castelfranchi, 'Silent agents: From observation to tacit communication.', in *IBERAMIA-SBIA*, pp. 98–107, (2006).

[6] C. Castelfranchi and R. Falcone, 'Principles of trust for mas: cognitive anatomy, social importance, and quantification', *Proc. of the Int. Conf. on Multi-Agent Systems (ICMAS'98), Paris*, 72–79, (1998).

[7] C. Catelfranchi, 'Modelling social action for ai agents', *Artificial Intelligence*, **103**, 157–182, (1998).

[8] E. Charniak and R. Goldman, 'A bayesian model of plan recognition', *Artificial Intelligence*, **64(1)**, 53–79, (1993).

[9] E. Denti, A. Omicini, and A. Ricci, 'tuprolog: A light-weight prolog for internet applications and infrastructures', in *Practical Aspects of Declarative Languages, 3rd International Symposium (PADL 2001)*, (2001).

[10] R. Falcone and C. Castelfranchi, 'Chaplin: A chart based plan recognizer', in *Proc. of the 13th International Conf. of Avignon.*, (1993).

[11] R. Falcone and C. Castelfranchi, 'Towards a theory of delegation for agent-based systems', *Robotics and Autonomous Systems*, **24**, 141–157, (1998).

[12] C. Geib and R. Goldman, 'Plan recognition in intrusion detection systems', in *DARPA Information Survivability Conference and Exposition (DISCEX)*, (2001).

[13] M.R. Genesereth, M:L. Ginsberg, and Rosenschein J.S., 'Cooperation without communication', in *Proc. of the National Conference on Artificial Intelligence (AAAI-86)*, (1986).

[14] Piotr J. Gmytrasiewicz, Edmund H. Durfee, and David K. Wehe, 'A decision-theoretic approach to coordinating multi-agent interactions.', in *Proc of International Joint Cconference on Artificial Intelligence (IJCAI-91)*, pp. 62–68, (1991).

[15] R. Goldmand, C. Geib, and C. Miller, 'A new model of plan recognition.', in *15th Conference on Uncertainty in Artificial Intelligence*, (1999).

[16] M. J. Huber, E. H. Durfee, and M. P. Wellman, 'The automated mapping of plans for plan recognition', in *Proc of the 10th Conf. on Uncertainty in Artificial Intelligence*, (1994).

[17] M.J. Huber and E.H. Durfee, 'Deciding when to commit to action during observation-based coordination', in *Proc. of the 1st Int. Conf. on Multi-Agent Systems ICMAS95*, (1995).

[18] D. Kinny and M. P. Georgeff, 'Commitment and effectiveness of situated agents', in *Proc. of the Twelfth International Joint Conf. on Artificial Intelligence (IJCAI-91)*, Sydney, Australia, (1991).

[19] A. Omicini, A. Ricci, M. Viroli, C. Castelfranchi, and L. Tummolini, 'Coordination artifacts: Environment-based coordination for intelligent agents', in *3rd international Joint Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2004)*, New York, USA, (2004).

[20] M. Piunti, C. Castelfranchi, and R. Falcone, 'Surprise as shortcut for anticipation: clustering mental states in reasoning', in *In Proc. of International Joint Conference on Artificial Intelligence (IJCAI07)*, Hyberadad, India., (2007).

[21] David Poole, 'Probabilistic horn abduction and bayesian networks', *Artificial Intelligence*, **64**(1), 81–129, (1993).

[22] A.S. Rao, *A unified view of plans as recipes*, chapter Contemporary Action Theory, Kluwer Academic Publishers, 1997.

[23] A.S. Rao and M.P. Georgeff, 'Bdi agents: From theory to practice', *Proc. of the 1st conf. on MAS (ICMAS95)*, (1995).

[24] M. Schumacher, *Objective Coordination in Multi-Agent System Engineering*, Springer Berlin / Heidelberg, 2001.

[25] M. Schut and M. Wooldridge, 'Intention reconsideration in complex environments', in *Proc. of the 4th international conference on Autonomous agents-AGENTS '00*, (2000).

[26] Y. Shoham and M. Tennenholtz, 'On the synthesis of useful social laws for artificial agents societies', in *Proc. of of 10th National Conference on Artificial Intelligence (AAAI-92)*, (1992).

[27] L. Tummolini, C. Castelfranchi, A. Ricci, M. Viroli, and A.Omicini, '"Exhibitionists" and "voyeurs" do it better: A shared environment approach for flexible coordination with tacit messages', in *Environments for MultiAgent Systems*, Springer-Verlag, (2005).

[8] www.alice.unibo.it

[9] vsis-www.informatik.uni-hamburg.de/projects/jadex/

Derivation of Minimal Mental Models

David V. Pynadath and Stacy C. Marsella
USC Information Sciences Institute,
4676 Admiralty Way,
Marina del Rey CA 90292 USA
{pynadath,marsella}@isi.edu

1 Introduction

A teacher deciding how to maintain discipline may find it useful to keep track of which students (dis)like each other. In general, enriching the mental models that the teacher has of her students enables her to make better decisions. On the other hand, it is harder for her to maintain correct beliefs over the richer models. Intuitively, we expect a diminishing return on enriching the mental models, where adding more details offers less gain in accuracy in beliefs and less benefit in decision-making quality, while incurring additional overhead in maintaining those beliefs. For example, while the teacher could also keep track of her students' musical performances, she would expect little benefit to doing so. In contrast a student may expect considerable benefit in keeping track of other student's musical interests.

This basic issue of forming and maintaining models of others is not unique to human social interaction. Agents in general face the challenge of forming and updating their mental models of each other in a wide range of multiagent domains. Research in plan recognition has produced an array of techniques for modeling a planning agent and forming a belief about what its goals and intentions are, so as to predict its future actions [4, 6]. User modeling faces a similar problem in trying to understand and anticipate the needs of human users interacting with a software system [2]. Agents working together as teams must maintain beliefs about their teammates' status [3]. Social simulation of human social behavior may require agents with a theory of mind about the other agents in their society [5]. In games of incomplete information, each player faces uncertainty about the payoffs that the other players will receive [1].

In these domains, forming mental models is typically treated as a separate subproblem outside the decision-making context of the agent. The modeling agent starts from an initial set of possible models for the other agents, whether in the form of plan libraries in plan recognition, possible mental models in social simulation, private types in games of incomplete information, etc. As the modeling agent interacts with the other agents, it updates that belief based on its observations of their behavior. The modeling agent then uses its mental models of the other agents to make informed decisions based on expectations of what they will do.

In this paper, we observe that we can quantify the tradeoff by taking the problem of modeling others out of its isolation and placing it back within the overall decision-making context of the modeling agent. Doing so allows the agent to automatically derive a space of mental models according to an informed analysis of the cost-benefit tradeoffs.

Our approach comprises three methods: *Behavior equivalence*, where the modeling agent clusters models that lead to the same behaviors in its decision-making context; *Utility equivalence*, where the modeling agent clusters models that may lead to different behaviors, but produce equally preferred outcomes with respect to its utility; and *Approximate Utility Equivalence*, where the modeling agent clusters models that lead to performance losses that are below a certain threshold, sacrificing a fixed amount of accuracy.

We envision several benefits from these approaches. In most multiagent domains, agents can expect that this analysis will allow them to drastically reduce the original full mental model space, without overly sacrificing performance. Additionally, in simulation research on human social interaction, it establishes a normative baseline for the simplifications and distortions in people's mental models of others or theory of mind.

2 Modeling Other Agents

Across the various multiagent domains already mentioned (and even within each domain itself) researchers have applied a wide variety of possible modeling frameworks. We present a methodology using an abstract agent framework that is general enough to cover these approaches, as well as other decision-making procedures in the literature. When applying our methodology to a specific domain, these components would become specialized to the particular framework used for the agents in that domain.

2.1 Agent Notation

In general, an agent consists of its beliefs (including those about other agents), its actions, and its preferences. We use the same structure to represent both the actual agents and the mental models they have of each other. Thus, we represent the multiagent system as a set of real agents, $\{m_i\}_{i=1}^N$. Each such agent includes possible beliefs over mental models, M_{ij}, that represent what agent i can think of agent j. The modeling agent wishes to minimize this space, M_{ij}. In particular, we want an algorithm that computes the expected utility derived by modeling agent i when using the set of mental model spaces, $\{M_{ij}\}_{j=1}^N$, for all of the agents j in the system. We define the behavior of an agent as a policy, $\pi : B \rightarrow A$, out of a set of possible policies, Π. Any agent architecture will include an algorithm for translating an agent into such a policy, π. We will abstract this procedure into a generic function SOLVE: $M \rightarrow \Pi$, that takes an agent model (whether real or subjective) and returns that model's policy of behavior.

2.2 Example Domain

We have taken our example domain from a scenario in childhood aggression, modeled within PsychSim, a multiagent social simulation tool [5]. There are agents for three students: a bully, his victim (i.e., the student he focuses his aggression on), and an onlooking student to whom the bully looks for affirmation. There is also a teacher who can deter the bully from picking on his victim by doling out punishment. We focus on the problem facing the bully agent, whose decision on whether or not to pick on his victim must consider the possible punishment policy of the teacher.

2.2.1 Utility

PsychSim uses a decision-theoretic model of preferences, so the bully agent decides whether to pick on his victim through maximization of his utility, which has three components: (1) a desire to increase his power, which decreases when he is punished; (2) a desire for affirmation from the onlooking student, which increases when the onlooker laughs along; and (3) a desire to decrease the victim's power, which decreases when the bully picks on him (as well as when the onlooker laughs at him). The bully's utility function is a linear combination of these three components, so that we specify his type as a triple of coefficients, each in $[0, 1]$. Thus, to simulate the behavior of a bully whose aggression is intended to gain the approval of his peers, we would use an agent with a higher weight for the second component. On the other hand, to simulate a more sadistic bully, we would use a higher weight for the third. The teacher's utility also has three components, corresponding to her desire to increase the power of each of the three students. She thus has a disincentive for punishing anyone unless doing so will deter acts that would reduce the victim's power even more. A fair teacher would give equal weight to the three students' power. A bully feeling persecuted by the teacher may think that she favors the victim's power over his own. On the other hand, a bully may feel that the teacher shares his dislike of the victim, in which case he may model her as having a lower weight for the victim. We focus on the bully's modeling of the teacher, so we fix the onlooker to value his power (i.e., he does not want to be punished), while also wanting to decrease the victim's power out of dislike (i.e., he enjoys laughing at the victim when the bully picks on him).

2.2.2 Actions

The teacher has 7 options in her action set, A_T. She can do nothing; she can scold the bully, onlooker, or the entire class; or she can punish the bully, onlooker, or the entire class. Punishing a student causes a more severe decrease in a student's power than simply scolding. The onlooking student has 2 options in his action set, A_O: laugh at the victim, or do nothing. The bully has 2 actions in his action set, A_B: pick on the victim or do nothing.

2.2.3 Policies

To reduce the domain to its most essential, the bully's policy, $\pi_B : M_{BO} \times M_{BT} \rightarrow A_B$, is a function of his mental model of the onlooker and teacher. Given that the onlooker has only one possible mental model, the policy space for the bully, Π_B, contains $|A_B|^{|M_{BT}|}$ distinct policies. Thus, the complexity of the bully's problem of choosing his correct policy is highly dependent on the number of mental models that he must consider for the teacher. Similarly, the onlooker's policy, $\pi_O : M_{OB} \times M_{OT} \rightarrow A_O$, depends on only his mental model of the bully and the teacher. In this current investigation, we focus on only one entry in π_O, namely the one where $m_{OB} = m_B$ and $m_{OT} = m_{BT}$, where there are only two possible values: laughing at the victim or not. We must also specify what the bully expects the teacher to do, which depends on not only her mental models of the students, but also on the prior actions of the students ($\pi_T : M_{TB} \times M_{TO} \times A_B \times A_O \rightarrow A_T$). In other words, the teacher may perform a different action when the bully picks on the victim than when he does not. The bully assumes that the teacher knows the correct model of him (i.e., $m_{TB} = m_B$) and shares his mental model of the onlooker (i.e., $m_{TO} = m_{BO}$). Even with our simplifications, there still remains a large space of possible behaviors for the teacher: $|\Pi_T| = |A_T|^{|A_B|\cdot|A_O|} = 2401$.

2.2.4 Solution Mechanism

We use boundedly rational agents, so the bully's SOLVE algorithm performs a forward projection over his possible actions and chooses the action with the highest expected utility. The forward projection includes the bully's action, the onlooker's subsequent response, and the teacher's resulting punishment decision. To determine the teacher's policy, the bully applies a SOLVE method from the teacher's perspective that exhaustively tries all policies in Π_T, computes the best-response policies for the bully and onlooker, and then chooses the best policy based on her expected utility. Given the teacher's policy, the bully and onlooker can then choose their best-response policies. We can specify the bully's mental model of the teacher in terms of the three utility weights that the bully attributes to her. In other words, our initial space of possible mental models, M_{BT}, contains one model for every vector of weights, $\vec{w} = [w_B, w_O, w_V]$. For the purposes of this paper we discretize this space to contain the vectors $[0.0, 0.0, 1.0]$, $[0.0, 0.1, 0.9]$, $[0.0, 0.2, 0.8]$, ..., $[1.0, 0.0, 0.0]$, with a total size of 66 possible mental models that the bully can have of the teacher (i.e., $|M_{BT}| = 66$). The bully agent's decisions are highly dependent on what he expects the teacher to do. For example, if he picks on the victim, he is more likely to be severely punished by a teacher for whom the victim is a pet (i.e., for which w_V is high), but he would be more likely to escape punishment if he himself is a favorite of the teacher (i.e., if w_B is high). Thus, there is clearly some value to be gained by maintaining differential mental models of the teacher. However, from a psychological point of view, it is unlikely that real-life bullies juggle 66 possible mental models of their teachers in their heads, so the space is a good candidate for reduction.

This scenario is illustrative, and there are clearly many dimensions along which we could enrich it. For example, we could introduce state dependencies (e.g., the weaker the victim, the more damage done by picking on him). However, while these additional wrinkles would change the particular answers provided by our methodology, they would not change the *ability* of the methods presented in the following sections to provide such answers. Our core methodology presents a very general approach to quantifying the value of different mental model spaces even in the face of these additional complications. Therefore, we have removed as many extraneous domain features as possible, so as to be able to provide the clearest illustration of the methods and how they can be applied to any multiagent domain.

3 Behavior Equivalence

The modeling agent's goal is to find a minimal set of mental models that it needs to consider for the other agents. In looking for possible bases for such minimization, we observe that the modeling agent's decisions often depend on only the *behavior* of the agents being modeled. Agents model the hidden parameters of others so as to generate expectations of their resulting behavior, but given the behavior of others, an agent's decision making is conditionally independent of the parameters behind it. For example, in agent teamwork, the mental states of the individual members have no direct effect on performance; only the decisions (actions, messages, etc.) derived from those mental states matter. Similarly, in games, the payoffs received by the agents depend on only the moves chosen by the players. In social simulations, the agents cannot read each others' minds, so they can base their decisions on only their observable behaviors. Therefore, regardless of what underlying parameters govern the modeled agent's decision-making, its eventual behavior is what has an impact on the modeling agent.

3.1 Behavior Equivalence Algorithm

This observation forms the basis for our first method for reducing the space of mental models. If two mental models produce the same behavior for the modeled agent, then making a distinction between them does not help the modeling agent. Therefore, it can safely remove one of them from consideration. It can do so by computing the policies corresponding to the possible mental models and clustering all that generate the same policy. The modeling agent then chooses one representative model from each cluster and removes all other models in the cluster from the overall space.

Algorithm 1 BEHAVIOREQUIVALENCE(M)

```
1: for all $m_1 \in M$ do
2:     for all $m_2 \in M$, $m_1 \neq m_2$ do
3:         if SOLVE($m1$) = SOLVE($m2$) then
4:             remove $m_2$ from $M$
```

For many domains, the repeated invocations of the SOLVE function can be computationally intensive, but there is plenty of opportunity for specialization of Algorithm 1. For example, if the mental models correspond to points in a utility space (as in our social simulation domain), it should be possible to compare mental models to only their immediate neighbors. Furthermore, even if specializing the algorithm is insufficient, there are many opportunities for approximation as well. For example, one could easily re-write the loops in Lines 1 and 2 to implement a sampling algorithm that compares randomly selected pairs for behavior equivalence.

3.2 Behavior Equivalence Results

The bully agent starts with 66 possible mental models for the teacher in M_{BT}. It can apply behavior equivalence to reduce the size of that set, but the policy chosen by the teacher also depends on her model of the bully. For example, different bullies may be more afraid of a teacher punishing the whole class because of him than of being punished by himself. We thus performed a behavior equivalence reduction of the mental model space across different types of bullies. To do so, we discretized the space of possible (real) bullies in the same way that we discretized the space of possible mental models of the teacher. Thus, we represent different types of bullies by different vectors of utility weights, $\vec{w} = [w_B, w_O, w_V]$, and discretize the set of possible types into 66 distinct such vectors, $[0.0, 0.0, 1.0]$, $[0.0, 0.1, 0.9]$, $[0.0, 0.2, 0.8]$, $\ldots$, $[1.0, 0.0, 0.0]$. Each of the 66 possible bully types started with an initial space, M_{BT}, of the 66 possible mental models for the teacher. We gave the teacher and onlooker the correct model of the bully and of each other. 8 types of bullies reduced the number of mental models of the teacher from 66 to 4. The other 58 types of bullies reduced the number of mental models of the teacher from 66 to 5. Behavior equivalence provides a clear benefit to these bully agents. In particular, it is notable that, although the 66 types of teachers had 2401 policies to choose from, a specific bully could expect to come across only 4 or 5 distinguishable teacher behaviors. In fact, looking across the results for all of the possible bully types, there were only 8 policies that were *ever* selected by the teacher in the $66 \cdot 66 = 4356$ bully-teacher combinations. The reason that so much of the teacher's policy space is undesirable for her is that the bully's behavior is constrained by his utility. For example, regardless of where in our utility space he is, the bully always prefers not being punished to being punished. Therefore, it would never make sense for the teacher to adopt a policy of punishing the bully if he does nothing to the victim and doing nothing to him if he does.

4 Utility Equivalence

There are some multiagent domains where the modeling agent derives some direct utility from the values of the intrinsic parameters. For example, in our social simulation, the teacher may prefer being liked by her students, rather than feared, even if both cases produce complete obedience. In such cases, using behavior equivalence may over-prune the mental model space. However, it is still safe to assume that the modeled agent matters only in so far as it affects the modeling agent's expected utility. The modeling agent is thus completely indifferent between different mental models that produce the same expected utility in its own execution.

4.1 Utility Equivalence Algorithm

This observation leads to our second method for reducing the mental model space. If the modeling agent does not lose any expected utility when using a particular mental model when the correct model is actually another, then distinguishing between the two does not help. Therefore, the modeling agent can compute its expected utility derived based on the policies corresponding to each of the possible mental models (of the modeled) and clustering all of the models that generate the same value when mistaken for each other. It then again chooses one representative model from each cluster.

Algorithm 2 UTILITYEQUIVALENCE(m, M)

```
1: for all $m_1 \in M$ do
2:     for all $m_2 \in M$, $m_1 \neq m_2$ do
3:         $\pi_1 \leftarrow$ SOLVE($m1$), $\pi_2 \leftarrow$ SOLVE($m2$)
4:         $u_{\text{right}} \leftarrow EU[\text{SOLVE}(m|m_2)|\pi_2]$
5:         $u_{\text{wrong}} \leftarrow EU[\text{SOLVE}(m|m_2)|\pi_1]$
6:         if $u_{\text{wrong}} - u_{\text{right}} \leq 0$ then
7:             remove $m_2$ from $M$
```

While behavioral equivalence requires only the modeled agent's policy, utility equivalence requires the further computation of the modeling agent's own best response to that policy. Line 5 shows that the modeling agent computes the expected utility (u_{wrong}) it will derive if it solves for its policy assuming that the modeled agent is of type m_2, when it is actually of type m_1. Line 4 computes its expected utility (u_{right}) when using that same policy when m_2 is the correct mental model. If the first is no lower than the second, then the agent can feel free to use m_1 in place of m_2. Line 6 accounts for the possibility that the utility loss might actually be negative when the agent being modeled, in turn, has an incorrect model of the modeling agent. Over time, if the agent being modeled updates its belief about the modeling agent, then such a utility gain is unlikely, because the modeled agent could eventually settle on a best response to the modeling agent's misconception. However, in the transient behavior, the modeled and modeling agents may inadvertently act in ways that improve the modeling agent's utility, despite the error in mental models.

Algorithm 2 adds another round of calls to the SOLVE function beyond what behavioral equivalence requires. The additional cost comes with the benefit of lossless reduction of the mental model space that sacrifices no utility to do so.

4.2 Utility Equivalence Results

To cluster the bully's mental models of the teacher according to utility equivalence, we followed the same experimental setup as for behavior equivalence. The 66 types of bully agents ran Algorithm 2, starting with the full space of mental models, M_{BT}. For this scenario, behavior equivalence implies utility equivalence, as the bully derives no direct utility from the teacher's intrinsic parameters. We can thus cluster the utility equivalence results according to the further reductions in mental model space achieved from M^b_{BT}. Of the 58 bully types with $\left|M^b_{BT}\right| = 5$, 11 types of bullies reduced the number of mental models of the teacher from 66 to 2, while the other 47 types reduced the number of mental models of the teacher from 66 to 4. Of the remaining 8 bully types with $\left|M^b_{BT}\right| = 4$, all of them reduced the number of mental models of the teacher from 66 to 3. Furthermore, for every type of bully, the mental model spaces reduced by utility equivalence (denoted M^u_{BT}) are all strict subsets of those reduced by behavior equivalence.

Some of the clustering occurs for bullies with extreme utility weights. For example, to a bully who cares about only hurting the victim (i.e, $\vec{w} = [0.0, 0.0, 1.0]$), mental models that differ on whether he himself gets punished are equivalent, because he does not care about the decrease in his own power. However, mental models that differ on whether or not the *onlooker* gets punished are not equivalent, because he desires the onlooker to laugh at the victim as well, to maximize the damage inflicted on the victim's power. Some of the clustering in this experiment arises when using an incorrect mental model of the teacher increases the bully's expected utility. For example, two mental models of the teacher may differ regarding whether punishment of the onlooker. From the bully's point of view, if the onlooker laughs regardless of the teacher's policy, then the bully does not care whether the onlooker is punished. Thus, while these two mental models produce different teacher behaviors, they produce the same expected utility to the bully, who is then justified in ignoring the distinction between them.

5 Approximate Utility Equivalence

The reduction of mental model spaces according to utility equivalence is lossless with respect to the modeling agent's decision making. Any further clustering of mental models will cost the modeling agent utility. However, the modeling agent can reduce its cost of maintaining beliefs over the mental model space by also clustering those models together that sacrifice a small amount of utility.

5.1 Approximate Utility Equivalence Algorithm

This observation leads to our third method for reducing the space of possible mental models. We can easily adapt Algorithm 2 to be tolerant of any utility loss below some positive threshold.

Algorithm 3 UTILITYAPPROX(m, M, θ)

```
1: for all $m_1 \in M$ do
2:     for all $m_2 \in M, m_1 \neq m_2$ do
3:         $\pi_1 \leftarrow$ SOLVE$(m1)$, $\pi_2 \leftarrow$ SOLVE$(m2)$
4:         $u_{\text{right}} \leftarrow EU[\text{SOLVE}(m|m_2)|\pi_2]$
5:         $u\text{wrong} \leftarrow EU[\text{SOLVE}(m|m_2)|\pi_1]$
6:         if $u\text{wrong} - u_{\text{right}} \leq \theta$ then
7:             remove $m_2$ from $M$
```

This approximate algorithm is no more complex than that for utility equivalence. In fact, we can perform a reduction using utility equivalence by passing in a threshold $\theta = 0$ to Algorithm 3.

The pseudocode in Algorithm 3 is written to support execution with a fixed threshold in mind. Alternatively, one could perform Lines 1–5 and *then* choose an appropriate threshold, θ, to reduce the space to an appropriate size. In other words, one would first profile the possible errors that would be derived from incorrect mental models before choosing a clustering. One could also easily vary the computation to use error measures other than expected utility. For example, one might be interested in worst-case utility loss instead of expected-case. Simply replacing the expectation in Lines 4 and 5 with a maximization would make the desired adjustment. There are any number of variations that would similarly modify the optimality criterion used in weighing the utility lost from the mistaken mental model.

5.2 Approximate Utility Equivalence Results

Figure 1 shows the results across our three methods for mental model space reduction. Each path from left to right represents the size of the mental model space for at least one possible type of bully as we raise its tolerance for utility loss. At the y-axis, all of the bully agents have the original mental model space of size 66. Then we see that these agents can reduce that size to either 4 or 5 models, using only the behavior equivalence method. The next point shows that the bully agents have spaces of 3–5 mental models when using only the utility equivalence method. Continuing along a path to the right represents the further reduction in the mental model space that comes with clustering mental models that cost less than the given threshold of expected utility.

As another example, there are 7 bully types that follow a path that leads to a mental model space of size one with only 10% loss of expected utility. If bully agents of this type are willing to tolerate a small utility loss, they can do away with modeling the teacher altogether! At the opposite end of the spectrum, there is one bully type that follows the upper envelope of the graph. For this bully type, utility equivalence allows for a mental model space of size 4, down from the size 5 of the space using only behavior equivalence. However, we see that even if the bully is willing to tolerate a loss of 25% of its expected utility, it still needs this full space of 4 models. If it wants to reduce its mental model space by even only one element, it can incur an up to 50% loss in expected utility if it is wrong. This bully type is also one of 14 in our sample space for which even tolerating 100% utility loss is not sufficient to warrant eliminating mental modeling together. In these cases, using the wrong mental model will lead to

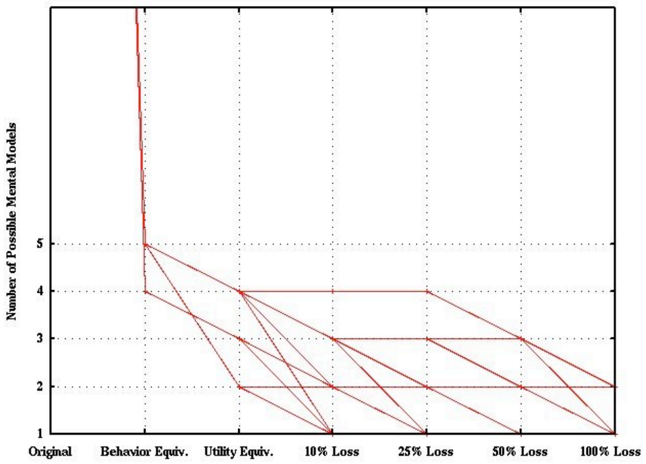


Figure 1. Size of model spaces vs. increasing leniency for utility loss, across all types of bully agents.

negative utility, so the bully has a strong incentive to do at least binary modeling of the teacher.

6 Discussion

While the exact graph in Figure 1 is specific to our example domain, it provides a concrete demonstration of our general ability to quantify the value of mental models to the modeling agent. To make the final decision, the agent must consider the computed value of the mental model space along with the cost of performing the actual model update and decision making during execution. As already described, the policy space of a modeling agent can grow exponentially with the number of mental models to consider. Furthermore, although we did not include the model update subproblem in our experiments, in most real-world domains, its complexity is highly dependent on the size of the mental model space. For example, probabilistic approaches, which compute a distribution over the possible mental models, can have a time complexity that is exponential in the size of the space. By finding minimal mental model spaces, an agent can apply more accurate belief update techniques that would have been computationally infeasible on larger spaces.

This methodology can also potentially create more psychologically plausible social simulations. In our experiments, the bully agents who were more attention-seeking (i.e., higher w_O) derived less value from the more complete mental model spaces for the teacher. Our characterization of bully types is consistent with the psychological literature that one can characterize different types of childhood aggression by the different goals that bullies have [7]. Thus, we can use our algorithms to explore the mental model spaces that we derive from those different goals and validate them against experimental data. Having validated the agents against such data, we can generate more confidence in the realism of the simulation.

We can also apply our algorithms to larger and more complicated domains. For example, our experiments have so far investigated the case of one agent choosing a space of mental models for only one other. Most multiagent domains will have multiple agents creating mental models of all of the others in the system. While our general methodology still applies in such cases, the additional interdependencies may lead to instabilities (e.g., an agent may be able to use a reduced space of mental models of another without utility loss only if the other uses a reduced space of mental models of him in return). Equilibrium concepts would provide one possible solution, but it may also be possible to re-cast our algorithms to simultaneously consider mental model spaces over the entire multiagent system, rather than over one modeling agent at a time.

While we deliberately designed this paper's domain to be simple enough to support a clear exposition and demonstration, we hope to learn more about the impact of design choices in mental modeling spaces and algorithms when we extend the analysis to richer domains. To support such domains, we will most likely have to implement some of the specialization and approximation techniques suggested in this paper. Once in place, we would be able to draw additional general conclusions about the impact of mental modeling choices as a function of fundamental properties of the multiagent system, and we expect that such general relationships may emerge on a richer class of domains.

7 Conclusion

At a higher level, the result of this investigation provides a key insight into the impact of social interaction on the design of multiagent systems. As designers, our immediate reaction is to view such interactions as complicating the problem of deriving appropriate multiagent behavior. However, as our results show, the interplay between the decision-making and modeling efforts of the individual agents is also highly *constraining* on that behavior. For example, out of the 2401 possible policies for the teacher, only 8 were ever desirable when interacting with our 66 types of bullies. When we view the problem of modeling other agents through the subjective lens of the modeling agent's own decision-making, we gain a utility metric that we can use both to restrict the scope of the modeling problem and to derive algorithms to solve it.

We used this metric to design algorithms that can quantify the value of distinctions made within the space of possible mental model space, and that then reduce that space accordingly. An agent can also use this same metric to derive a mental model space from scratch, simply by quantifying the value of *adding* mental models to the space of consideration. In this manner, our metric allows an agent designer to isolate those aspects of the mental models that are most relevant to the agent. We expect the algorithms to give such designers novel insight into the nature of their domains and to minimize the computational complexity of modeling other agents in all multiagent domains where such modeling is beneficial.

REFERENCES

[1] D. Fudenberg and J. Tirole, *Game Theory*, MIT Press, 1991.

[2] Anthony Jameson, 'Numerical uncertainty management in user and student modeling: An overview of systems and issues', *User Modeling and User-Adapted Interaction*, **5**(3-4), 193–251, (1995).

[3] Gal Kaminka, David V. Pynadath, and Milind Tambe, 'Monitoring teams by overhearing: A multi-agent plan-recognition approach', *Journal of Artificial Intelligence Research*, **17**, 83–135, (2002).

[4] Henry A. Kautz and James F. Allen, 'Generalized plan recognition', in *Proceedings of the National Conference on Artificial Intelligence*, pp. 32–37, (1986).

[5] David V. Pynadath and Stacy C. Marsella, 'PsychSim: Modeling theory of mind with decision-theoretic agents', in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1181–1186, (2005).

[6] C.F. Schmidt, N.S. Sridharan, and J.L. Goodson, 'The plan recognition problem: An intersection of psychology and artificial intelligence', *Artificial Intelligence*, **11**, 45–83, (1978).

[7] David Schwartz, 'Subtypes of victims and aggressors in children's peer groups', *Journal of Abnormal Child Psychology*, **28**, 181–192, (2000).

From Language to Thought: Inferring Opinions And Beliefs From Verbal Behavior

Fiorella de Rosis and **Nicole Novielli**[1]

Abstract. In this contribution, we describe our ongoing work in the direction of studying how negative or positive opinions may be recognized from language and how beliefs may be dynamically inferred from expressions of opinion. We begin by considering the language processing methods which have been applied to 'sentiment analysis' to show the results they produced and their limits, and then reflect on how beliefs may be inferred gradually, in conditions of uncertainty and by carefully considering various forms of context.

1. INTRODUCTION

Humans may express their opinions with several means: with actions, body attitudes and language: they may patently shiver, close the windows or say *'Cold today, isn't it?'*, to manifest their opinion that the temperature at home is not adequate. Opinions may be about the environment (as in the previous example), about other people or about themselves. However, the relationship between beliefs and actions, attitudes and language is not so strict: I might simulate shiver, close the window or say *'Cold today isn't it?'* for reasons of politeness, because I presume that my partner living with me feels cold. Considerable efforts are being made towards inferring goals from observation of nonverbal behavior (see, e.g., [6]). Language is particularly difficult to interpret, as an expression medium: humans may more easily lie or simulate their beliefs by speaking than with their body expressions. And still, language will be, probably for years, one of the most common communication media with smart environments. In this short contribution, we describe our ongoing work on the problem of how negative or positive opinions may be recognized from language and how beliefs may be inferred from expression of opinions. We start from considering the language processing methods which are applied in 'sentiment analysis', to show the results they produced and their limits. We then reflect on the fact that beliefs may be built gradually, both in their strength and their level of certainty. We reason on the factors which may influence masking expressions of beliefs in various contexts and how these factors may be considered in the interpretation of a given sentence.

2. BACKGROUND

Relation between language and thought was the object of philosophers' and psychologists' research since long. Three phases may be recognized in the process of inferring the content of a speaker's thought by a hearer: a. identification of the *meaning* of words used; b. identification of the *proposition* expressed in light of the meaning and the rest of the situation in which the utterance takes place and c. identification of further *implicatures* over and above the proposition expressed [5]. Phase b. is generally made possible only by the analysis of *context*, without which sentences would be not interpretable; the same is true for phase c., in which aspects of the mental state of the speaker, such as beliefs and intentions, are inferred. This inference goes beyond the immediate meaning of the utterance. In a well-known example by Grice [7], A may say to B *'the bus will be here within five minutes'* not just in order to transmit his belief, but in order to put her at ease, because he observed that she is impatient for the bus to arrive. But he might say it for a number of other reasons: to put himself at ease because he is impatient, as a pretext to begin a conversation with B, to justify the bus delay because he feels responsible for this, and so on. He might even say something he doesn't believe, either intentionally or without intentionally misrepresenting himself [8] and computers may imitate this behavior [1]. If B knows about A, for instance because the utterance to interpret was pronounced in the context of an ongoing dialogue, finding the most likely interpretation of A's utterance will be easier but the 'context' to consider will be wider. In defining a communication language among artificial agents, Cohen and Levesque [2] neatly stated the semantics of 'illocutionary acts' in terms of the effects the Speaker intends to achieve with them: the hypothesis was that this effect always consists in 'communicating own mental state', with the Speaker's 'sincerity' as a strong assumption about communication conditions. This work on communication language of artificial agents is of primary importance in the immediate interpretation of a given sentence in terms of an agent's beliefs and intentions; however, it is of more limited use in natural language understanding where (as we said in our previous examples) consideration of the context -in a wide sense- is essential to avoid trivial interpretations.

A rough description of the user's beliefs in a human-computer conversation could be made by just extracting and summarizing the opinions expressed during the dialog [1]. This simple summarization still requires considering the degree of uncertainty in the expression of opinions and of consistency in opinions expressed at different times. A more sophisticated description of beliefs requires, however, a wider consideration of the context in which the opinions were expressed and of other sources of knowledge about the speaker's mind.

3. OUR STUDY

Our long-term goal is to build a dialogue system which provides user-tailored suggestions about healthy living habits. According to Prochaska and Di Clemente's Transtheoretical Model of Change (TMC in [14]), this kind of dialogue should apply a strategy in which the presumed 'stage of change' of the client (from a 'wrong' to a more 'correct' behavior) is considered, to adapt dynamically the information and persuasion plans to the specific situation. By building it in a domain-independent way, our ambition is to apply this dialogue system to any behavior problem occurring in a smart environment (smoking in presence of children, using energy sources in a not economic way etc).

[1] Department of Informatics, University of Bari Via Orabona 4
70126 Bari, Italy - email: {derosis, novielli}@di.uniba.it

The TMC model suggests how stages of change may be recognized from a set of signs which include the *value* given to the 'correct' behavior which is the object of the persuasion process, the *knowledge of* reasons influencing the adoption of a problem behavior, the *belief that* (internal and external) *conditions exist to change this behavior* and the *intention* to change it. Recognizing the stage of change in natural language dialogues therefore requires inferring these aspects of the users' mental state from their verbal behavior. A corpus of dialogues may help to define some recognition methods and validate them. Our corpus of conversations about healthy dieting with an ECA was collected with a Wizard of Oz tool [4]: the examples in this paper are extracted from that corpus. We will focus our reflections, in particular, on the expression of the *belief* that own behavior is wrong.

4. THE METHOD

To guess another person's beliefs and intentions, one may *'perceive situational information about another and use it as input into one's own cognitive structure'* [6]. If the user's beliefs to recognize regard his own behavior (whether it is 'correct' or 'wrong'), perception may be translated into identification of expressions of opinions, while beliefs can be inferred through a mapping with the system's belief structure.

Table 1. Excerpt of an example dialogue from our corpus

N	Adjacent pair	Recognition	Question
1	S: Do you think your diet is correct or would you like to change your dietary habits? U: I think I should control my consumption of sweets.	U declares that his *consumption of sweets is not correct.*	Does U really believe it?
2	S: Have you ever desired to change your diet? U: Yes	U declares he *wanted to change* his dietary behavior *in the past.*	Did U want it really?
3	S: Do you think your weigth is correct or would you like to change it? U: I am 4 kg overweight so I would like to change it. What should I do?	U recognizes he *is overweight*. He manifests his (light) *intention of loosing some weight.* He *asks some support* in this direction.	Does U really believe it? Has he really the intention to change? Does he really need support?
4	S: Why you never tried to do it? U: Because I'm not so patient	U declares he is not in the conditions to change his dietary behavior because of *lack of endurance.*	Does U really believe that he lacks of endurance?
5	S: Do you think you assume all the substances you need? U: I regularly assume carbohydrates, proteins and vitamins. I probably assume not enough minerals and too much fat.	U declares that his diet is: - *well balanced in carbohydrates, proteins and vitamins*; - not *well balanced in minerals and fats*	Does U really believe that his diet is correct in these substances and incorrect in others?
6	S: You don't have to completely eliminate fat foods. You have just to limit their proportions U: I can't organize my diet because I have no time.	U declares that he is not in the conditions to change his dietary behavior because of *lack of time*	Does U really believe that he has no time to organize his diet?
7	S: Sure, organize your diet could appear a challenging task especially when nobody can help you and you have to work or to study. U: I am 90% invalid	U declares that he is not in the conditions to change his dietary behavior because of inability	Does U really believe that he is not able to prepare correct meals?

Let us consider the excerpt from one of the dialogues in our corpus, that is shown in Table[2] 1: we will go through this example (and some variants) to discuss some of the problems in progressively inferring the user's opinion and beliefs as far as this short dialogue goes on.

4.1 Identifying expression of opinions about own behavior

Identification of positive or negative opinions expressed linguistically can be seen in terms of 'sentiment analysis'. This method aims at recognizing the viewpoint underlying a text span: a typical task is the binary classification of texts in order to define their polarity (positive vs. negative, that is 'thumbs up' or 'thumbs down', good or bad). This goal is achieved by applying traditional machine learning techniques to a multidimensional representation of the collection of documents. The definition of the set of features involved in the representation is crucial, and several groups are working on the selection and interpretation of indicators to improve results in terms of accuracy.

In the *bag of words* (BoW) approach, basic features for the vectorial representation are unigrams, bi-grams or tri-grams and the standard approach is to measure the frequency of these elements, or of a group of words of known sentiment orientation, in a document belonging to a given class. Text based features can also be derived from an ad hoc lexicon built in a preliminary phase of the study, by means of thesauri or semantic dictionaries such as WordNet[3]. To improve the accuracy of the classification, BoW are usually enriched with additional features which may be based on the proximity between the items to classify [13], on an 'ad hoc' taxonomy [15] or on the relationship of every word with the previous or the next one, as they appear in the parsing tree for the complete sentence [16].

When attempting to recognize opinions within a dialog interaction rather than from analysis of a single text, more information about the context in which a sentence was pronounced is available. As we will see, on one hand this information makes

[2] Translated from Italian. S stands for 'System', U for 'User'

[3] http://wordnet.princeton.edu/

recognition a more complex task but, on the other one, it provides more opportunities for a correct solution.

Let us consider the various forms of 'context' that occur in the interpretation of a dialogue move:

4.1.1 Local context: the user move

Most work on sentiment analysis was developed on monologs, such as movie reviews. Extending these methods to the analysis of single sentences or brief dialog turns is not immediate. At a first glance, sentiment analysis should work well also in these cases: we might think to simply look at the prior polarity of subjective words such as 'correct' to interpret the polarity of sentences like '*I think my dietary habits are correct*'. However, after looking in our corpus we noticed that, to recognize the polarity of the user move, many other things have to be considered. In general, it has been proven that a word based approach is not powerful enough, especially in non binary classification tasks [16]. Mullen and Malouf [12], e.g., tried to identify the political affiliation of bloggers by analyzing their post on a web forum; purely text-based methods produced, in that case, a low accuracy because most posters from across the political spectrum used common terms such as *'gun control'* or *'abortion'*, regardless of their opinion on those particular issues. These authors concluded that the accuracy could be improved by introducing rules based on the observation of how posters interact with each other, that is by adding information about the *context* in which a post is added to a discussion.

The situation becomes more complex if we want to perform sentiment analysis at the phrase level (short dialogue turns): the majority of problems is related to stop-words elimination involved by the BoW representation [11]. The prior polarity of words can be affected by linguistic factors that modify their '*contextual polarity*' [16]. A typical example is the presence of negations, that may be local (*'I think my dietary habits are not correct'*) or may involve longer-distance dependencies (*'I don't think my dietary habits are correct'*). If we simply rely on a word based approach, we might classify as identical the opposite cases *'I think my dietary habits are correct'* and *'I think my dietary habits are not correct'*. This problem is due to the stop words elimination and has an impact on the recognition of the 'strength' of the opinion expressed: since adverbs are usually taken as stop words, sentences like *'I think my weight is pretty good'* and *'I think my weight is really good'* would be considered as identical.

In [9], the modifiers that change the semantic orientation (negations) of a term or its weight (intensifiers and diminishers) are named *valence shifters*. The cited paper presents a comparison between two approaches: in the first one, positive and negative terms in a document are counted, and the text is classified as having a positive orientation if more positive than negative terms are found (and vice versa) or neutral when the number of positive and negative terms is the same. Polarity of single terms is decided according to a dictionary. The second method takes into account contextual valence shifters in determining the semantic orientation of non-neutral words. A parser is used to determine which modifiers to apply to which terms. The term-counting method has the advantage of not requiring any training phase, since one can simply rely on a lexicon established a priori: however, methods based on shifters evaluation proved to be more effective in terms of accuracy. The case of negation and, in general, of all modifiers, is also discussed in [15]: these authors present a new method for sentiment analysis based on extracting and analyzing *adjective appraisal groups* such as *'really good'* or *'not so bad'*. Appraisal groups include an *head adjective* and an optional list of *appraisal modifiers* with nested scope, each denoting a transformation of one or more appraisal attributes of the head. Four attributes are used to describe every group: *attitude*, which gives the type of appraisal being expressed, *orientation,* which is the polarity (positive or negative) of the appraisal, *graduation,* which is the intensity of the appraisal and its focus and *polarity,* which says whether the group is *marked* as scoped in a polarity marker such as a negation. This taxonomy was employed to tag the lexicon in an enriched BoW representation in which terms were located in the four dimensional space by giving a value to all appraisal attributes.

There are also cases in which investigating the role of modifiers is still not enough. A typical example is: *'I can't resist to a delicious sweet, what should I do?'*. In this example, lexicon with prior positive polarity prevails (*'delicious sweet'*) and the action of modifiers (*'I can't resist'*) does not necessarily produce a negative classification of the turn: on the contrary, the negation of the verb strengthens the appeal of the 'sweet' word. In cases like this, the parsing tree of the sentence should be explored to capture its real semantics by analyzing the syntactic role of every word.

4.1.2 Wider context: dialogue pairs

In all the examples we saw so far, the context to consider in sentence interpretation was limited to a single user move. In other cases, however, knowledge of the previous system's move is essential to recognize the user's expression of opinion (see, e.g., the pair n.2 in Table 1). In our corpus, we found complete expressions of opinion like *'I think my weight is correct'*, but also several sentences such as *'pretty good'* or *'I think it is ok'* after the question: *'What do you think of your dietary habits?'*. In these cases, sentiment analysis may classify the user answer as generically positive or negative, but only thanks to our knowledge about the context we may say something about the user opinion.

Beliefs inferred in the two cases have not the same level of validity: we will name *direct beliefs* those inferred from direct declarations of opinions, and *from-answers beliefs* those inferred from answers to system questions. Although they represent alternative ways of expressing beliefs, the first one is likely to provide a stronger evidence that the second one. An example: if (as in pair n.1) the system's question was "*Do you think your diet is correct or would you like to change your dietary habits?"* and the user answers *"I think I should control my consumption of sweets",* we may infer the user's negative opinion about his own behavior with a lower level of certainty than if the question simply was: *'Tell me something about your diet'*. Strengthening or weakening of the level of certainty about an inferred belief may occur by combining different parts of a given move. For instance, in the pair n 3, the final user question *'what should I do?'* (at move 4) strengthens the presumed U's negative opinion about being overweight that was expressed in the first part of the sentence.

4.2 Progressively inferring beliefs: context is the whole dialog

The problems discussed so far are only related to the task of determining the sentiment orientation of an individual user move and inferring a presumed belief from that local analysis. However, beliefs cannot be directly inferred from a unique expression of opinion. Recognizing the user beliefs relies on consideration of other aspects as well, such as the opinion holder, his status (how much credible, how much competent in the domain he is etc). One might express a personal opinion (*'I think I'm drinking too much'*) or refer others' opinions (*'My wife says I drink too much'*) or ask a question to the system playing the role of an expert in the domain, in order to check whether its beliefs are aligned with his own ones *(Do you think that drinking four beers a day is too much?')*. In the first case, as we said, inference of beliefs from opinion expression is more direct and stronger, while in the second and the third one it is more indirect and weaker. Once we understand that in the sentence *'My wife says I drink too much'* the opinion holder is U's wife, we need to know whether U thinks that his wife is credible

Figure 1. From (uncertain) opinion recognition to (uncertain) belief interpretation

and competent in the domain or whether he thinks she is (for instance) too anxious or oppressive. In the third example, a question to the system may be interpreted in terms of a condition of doubt rather than of a clear belief: overall, whether U considers that source as 'believable'.

We call *from-question* all beliefs generated by this kind of situation, and *indirect* the kind of beliefs that originate from referring an external source's declaration rather than a personal opinion. Figure 1 synthesizes the difference in inference of direct, indirect, from-question and from-answer beliefs, in context-based sentiment analysis. In this figure, 'z' represents a generic fact about the user diet; for instance: 'U is overweight', Overweight(U). The node '(Say U z)' represents a declaration of the type *'I am overweight'*. The node '(Answer U z)' represents an answer '*No*' to the system question *'Do you think your weight is correct?'*. The node '(Say U (Say A z)' represents a declaration of the type *'My wife says my weight is not correct'* and the node '(AskWhether U z)' represents a question like *'Do you believe that 90 kilos are too much for a person of my height?*

As we said in Section 2, a belief may be inferred gradually from a cumulative expression of consistent opinions, and this inference process can be based on a mapping with the system's belief structure. A 'correct' behavior is the result of a number of components: in the case of healthy dieting, a good proportion of vegetables, a right balance of the other components, regularity of meals and so on. Figure 2 represents the relationship about believing that own dietary behavior is balanced (or not) and believing that the components of dietary behavior are correct (or not).

Figure 2. Relationship between generic and specific beliefs about own diet

A system playing the role of an advice-giver in this domain holds its knowledge in a 'consistent' set of beliefs. Recognizing how much consistent the user's set of beliefs appears to be is a dynamic process: the system progressively builds an image of the user mind by updating it after recognizing every expression of opinion, and by considering the strength and uncertainty of opinions expressed. Figure 3 represents the dynamic updating of the system's image of the user's beliefs during the dialogue. In this oriented graph, the relationship between every leaf node and its parent node at time t (Bel U CorrectDiet(U) t) is a function of how important is the variable associated with the child node in defining a diet as 'correct'. The relationship between this last node and its parent nodes represents, in its turn, two effects: i) the progressive refinement of the system's image of the user's mental state, based on the information acquired during the dialogue and ii) the possible change in the user's belief about his own dietary behavior, produced by the system's suggestions and information provision.

Table 2 describes how the system's image of the user beliefs evolves during the dialogue, as soon as new information is acquired. Let us start from time t1 (first dialogue pair). The sentence 'I think I should control my consumption of sweets' is interpreted by the sentiment analyser as a direct statement of belief that he tends to take too much sweets in his diet (Say U MuchSweets(U)); this increases the probability of DirectBel U MuchSweets(U) (as in figure 1) and, consequently, decreases the likelihood that he believes his diet is well balanced (Bel U BalancedDiet(U)) and, therefore, correct (Bel U CorrectDiet(U)) (as in figure 2).

We now go to the next time slice in figure 3 (t2). The sentiment analyser interprets the user move 'Yes' as a display of opinion that the present diet is not correct, although with a lower level of certainty than in the previous move (because it is a 'FromAnswer' type of belief).

Figure 3. Dynamic updating of the user's set of beliefs

The probability of the corresponding node is updated... and so on. In this example, the system progressively acquires new information about the user during the dialogue, but apparently it does not influence the user mind with its moves, if not very slightly (as it just makes questions rather than giving overt suggestions) .

Table 3 shows an excerpt of dialogue between a real human therapist and a subject with addictive behavior related to alcohol consumption [10].

The example shows how an advice-giving dialog system should apply a successful persuasion strategy: the user U gradually moves from the 'precontemplation' stage (move 1), to the 'preparation' one (presumed, as it appears from move 7), passing through the 'contemplation' stage in the central part of the dialog in which the awareness of adopting a wrong behavior gradually emerges, thanks to the ability of the therapist in formulating 'ad hoc' questions. Let's denote with S a dialog system equipped to emulate this behavior: in this example the variation in the system's image of the user belief changes gradually because the user is progressively persuaded by the system's suggestions; the model is identical to the one applied in the previous example but the node named 'SystemMove' in figure 3 contributes in this case to increase the probability of the node Bel U CorrectDiet(U), t.

Table 2. Progressive updating of the system's image of the user's mind during the dialogue

N	Adjacent pair	Recognition	U's beliefs at time ti
1	S: Do you think your diet is correct or would you like to change your dietary habits? U: I think I should control my consumption of sweets.	DirectBel U MuchSweets(U)	↓ Bel U CorrectDiet(U), t1
2	S: Have you ever desired to change your diet? U: Yes	FromAnswerBel U not CorrectDiet(U)	↓ Bel U CorrectDiet(U), t2
3	S: Do you think your weight is correct or would you like to change it? U: I am 4 kg overweight so I would like to change it. What should I do?	FromAnswerBel U not CorrectWeight(U)	↓ Bel U CorrectWeight(U), t3
4	S: Why you never tried to do it? U: Because I'm not so patient	FromAnswerBel U not Enduring(U)	↓ Bel U ConditionsToChange(U), t4
5	S: Do you think you assume all the substances you need? U: I regularly assume carbohydrates, proteins and vitamins. I probably assume not enough minerals and too much fat.	DirectBel U OKCarbohydrates(U) DirectBel U OKProteins(U) DirectBel U OKVitamins(U) DirectBel U not OKMinerals(U) DirectBel U not OKFats(U)	↑ Bel U BalancedDiet(U), t5 ↑ Bel U BalancedDiet(U), t6 ↑ Bel U BalancedDiet(U), t7 ↓ Bel U BalancedDiet(U), t8 ↓ Bel U BalancedDiet(U), t9
6	S: You don't have to completely eliminate fat foods. You have just to limit their proportions U: I can't organise my diet because I have no time.	DirectBel U not HasTime(U)	↓ Bel U ConditionsToChange(U), t10
7	S: Sure, organize your diet could appear a challenging task especially when nobody can help you and you have to work or to study. U: I am 90% invalid	DirectBel U not IsAble(U)	↓ Bel U ConditionsToChange(U), t11

Table 3. Progressive change of the of the user's belief during the dialogue, due to the persuasive strategy adopted by the advisor

N	Adjacent pair	Recognition
1	S: So one thing you've noticed is that you are drinking more now than you used to. What else? U: I can't really think of anything else. It doesn't really affect that much. I don't really get drunk very often.	U declares that he is not concerned about his drinking behavior.
2	S: So, although you know that your drinking has gone up over the past few years, it doesn't really seem to affect you more. U: Right. I can drink all night and it doesn't make me drunk. Other guys have trouble keeping up with me.	U declares he has not problems related to alcohol consumption
[…] *U talks for a little while about his father's drinking and bout problem related to that behavior.*		
3	S: Is there anything else you've noticed, any other way in which your drinking seems like your father's? U: Lately, there has been some times when I can't remember things that happened. I'll be drinking at a party, and the next morning I can't remember getting home. It's not too pleasant to wake up and have no idea where you left your car.	U recognizes he has got memory problems due to alcohol consumption.
4	S: That can be scary, especially the first few times it happens. Give me an example. U: About 2 weeks ago, I was out with Bob and I guess I drank a little more than usual. When I woke up in the morning, I couldn't think of where my car was. I looked out the window and my car was in the driveway, and I guess I drove it there. I felt terrible.	
[…]		
5	S: Your situation doesn't seem bad to you. U: No, it doesn't. I've quit drinking for weeks at a time with non problem. And I have a couple of drinks and leave it alone. I have a good job and a family. How could I be an alcoholic? […] I mean, I've got some problems, but I'm not a drunk.	U recognize he has got some problems with alcohol but he declares he *is not a drunk*
[…] *the therapist shows him results of blood test and explain him how drinking could affect his health.*		
6	U: So I'm driving around legally drunk three times a week? So I have a higher risk, then?	U recognize his problem behavior with alcohol
7	S: That's it U: I guess I have to do something about my drinking – either cut it down or give it up.	U declares he want to reduce or even stop alcohol consumption.

5. OUTLINE OF THE ALGORITHM

The analysis exposed so far suggested us to define a markup language for isolating those dialog turns in which we can find expression of opinions and from which to start for the definition of a method for automatically recognize them. The WoZ corpus has been annotated by three independent raters and by taking as text unit every single dialog turn as adjacent pairs (a couple of adjacent System-User moves in the dialog, as shown in our examples).

The markup language is composed by the following four tags:

- *opinion polarity*: this tag can assume either value of 'positive' or 'negative' according to the polarity of the opinion expressed, or 'neutral' if no opinion is expressed;
- *opinion object*: the aspect of the diet (or the examined behavior in general) which is object of the opinion expressed (consumption of sweets, carbohydrates, vegetables etc…);
- *move type*: whether the opinion is expressed by mean of a direct sentence ('direct opinion') rather than a 'question/answer to the system'. The tag may also assume the value 'indirect opinion' when the user refers to a third person's opinion;
- *believability* of the opinion holder: tag used in case of 'indirect opinion', may assume value in {'high', 'low', 'neutral'}.

Results of the markup experiment showed that data are very sparse and we could not simply rely on classical machine learning techniques for automatically infer tag values during the system usage. Our idea is to combine sentiment analysis techniques, applied at level of dialog turns for opinion extraction, with decision rules, based on information about the context and the dialog history, as observed in our corpus.

We sketched an algorithm which describes the dynamic recognition process we intend to implement and perform during the interaction, at every dialog turn. The algorithm is repeated every time a new user's move is entered, after a system's move, and is mainly organized in the following three steps:

1. A new user's move is entered and treated as input for a module which implements sentiment analysis techniques. The process of opinion extraction also involves information about the move and dialog pairs context, as explained so far, and gives as output the opinion polarity and its object (typically one of the aspect of the behavior considered, as showed in fig. 2);
2. The system infers a possible belief with respect to the output produced at the previous step. The belief recognized could be 'direct', 'indirect' fromQuestion' and 'fromAnswer' and each of them has a different weight in updating the set of user's beliefs. The updating of the particular belief inferred has effect on more general ones, as showed by example in tab. 2;
3. At every dialog step at time t, the system updates the image of the user's mind on the basis of the new knowledge acquired at steps 1 and 2 and on the knowledge at previous slice time *t-1.*

6. CONCLUSION

This contribution is a preliminary statement of the direction in which we are moving in our study about the relation between opinion expression and belief inference. The relationship between beliefs and action, attitudes and language is not so strict and in particular language is not easy to interpret. In this work, we have studied how beliefs can be dynamically inferred, during the interaction, from a set of consistent opinions in the scenario of a advice-giving dialog system in a smart-environment.

Our long-term goal is to build an user-adapted dialog system which dynamically fits persuasion plans to every specific situation and provides user-tailored suggestion about about healthy living habits.

According to Prochaska and Di Clemente's Transtheoretical Model of Change, we tried to define a method for automatically infer information about all beliefs related to the particular user's 'Stage of Change'. The main idea underlying our work is that the system may infer users' beliefs through a mapping with is own belief structure, by using as input for this process the expressions of users' opinions, as they can be observed in their linguistic behavior.

In this work we investigated the state of art in sentiment analysis techniques in order to find the main limitations that we have to cope with when operating in a dialog context. The mark-up of our corpus showed us sparsity of data and this suggested us to sketch an algorithm which combines sentiment analysis for opinion extraction with decision rules based on the observation of the context such as the dialog history or the last system's move.

ACKNOWLEDGEMENTS

This work was financed, in part, by HUMAINE, the European Human-Machine Interaction Network on Emotion (EC Contract 507422).

References

[1] J. Allwood (1981): *Language beliefs and concepts*. In Allwood, Fragsmyr, Hjort and Svedin (Eds): Natural Resources in a Cultural Perspective. FRN-Report 37, Stockholm.

[2] P.R. Cohen and H.J. Levesque (1995). *Communicative actions for artificial agents.* Proceedings of the First International Conference on Multi-Agent Systems (ICMAS'95).

[3] F. de Rosis, C. Castelfranchi, V. Carofiglio and R. Grassano (2003). *Can Computers deliberately deceive? A simulation tool and its application to Turing's Imitation Game.* Computational Intelligence, 19, 3.

[4] F. de Rosis, N. Novielli, V. Carofiglio, A. Cavalluzzi and B. De Carolis (2006): *User modeling and adaptation in health promotion dialogs with an animated character. Journal of Biomedical Informatics, Special Issue on 'Dialog systems for health communications'*. T.Bickmore (Ed), 39, 5, 514-531.

[5] C. Gauker: *Think out loud: An essay on the relation between thought and language*. Princeton University Press, 1994.

[6] J. Gray and C. Brezeal (2005). *Toward helpful robot teammates: a simulation-theoretic approach for inferring mental state of others.* AAAI Workshop on Modular Construction of Human-Like Intelligence, Pittsburgh.

[7] P.H. Grice (1989). *Studies in the Way of Words*, Harvard University Press.

[8] S. Gross (2006). *Can one sincerely say what one doesn't believe?* Mind and Language. 21, 1.

[9] A. Kennedy and D. Inkpen (2006). *Sentiment classification of movie reviews using contextual valence shifters.* Computational Intelligence, 22, 2, 110–125.

[10] W.R. Miller, S. Rollnick, *Motivational Interviewing – Preparing People to Change Addictive Behaviour*, The Guilford Press.

[11] T.M. Mitchell (1997). *Machine Learning.* Boston, MA: McGraw Hill.

[12] T. Mullen and R. Malouf (2006). *A preliminary investigation into sentiment analysis of informal political discourse.* AAAI Symposium on Computational Approaches to Analyzing Weblogs, 159-162.

[13] B. Pang and L. Lee (2004) *A Sentimental Education: sentiment analysis using subjectivity summarization based on minimum cuts.* Proceedings of the 42nd ACL, 271-278

[14] J. Prochaska, C. Di Clemente and H. Norcross (2002) *In search of how people change: applications to addictive behavior.* American Psychologist, 47.

[15] C. Whitelaw, S. Argamon and N. Garg (2005). *Using appraisal taxonomies for sentiment analysis.* Proceedings of the First Computational Systemic Functional Grammar Conference, Sydney.

[16] T. Wilson, J. Wiebe and P. Hoffmann (2005) *Recognizing contextual polarity in phrase-level sentiment analysis.* Proceedings of HLT-EMNLP Vancouver

AI and Narrative Games for Education

There is an increasing interest in the computer games industry in the development of games with emotionally compelling interactive stories. Games designers, screen writers and narrative theorists propose contrasting approaches to engineer satisfying stories in which players can participate for pure entertainment or educational purposes.

Intelligent serious games for education are applications that use the power of AI and the characteristics of games to create educational engaging learning experiences. A game is a system in which a player/s can engage in an artificial challenge that results in a quantifiable outcome. The outcome of a serious game is the achievement of the learning goals set within a realistic context.

This symposium focuses on the application of artificial intelligence techniques, frameworks and theories to the creation of interactive engaging narrative games for education. It will address questions such as:

- How is believable story engineered through games?
- What are the crucial elements of a believable story?
- How can educational goals be achieved through narrative games?
- How should the interaction between the player/s and the game take place?
- How should the characters behave to achieve emotionally convincing stories?
- How can we design interactive stories in which the player's experience is central?
- How can we scale up prototype interactive narrative architectures to meet the requirements of today's game engines?

Themes running throughout the symposium will be the extent to which game engines can be used as research tools and the appropriate methods for disseminating and sharing prototype systems throughout the community. We welcome researchers from academia, education and industry, in particular those involved with the design, development and evaluation of AI based narrative and games. Their expertise could be in a range of areas including: narrative, educational research, multimedia, game design and development, interaction design and evaluation for children and any other relevant area.

Daniela M. Romano, Paul Brna, Judy Robertson & Sandy Louchart (Symposium Chairs)

Programme committee: Ahmed BinSubaih (University of Sheffield); Ana Pavia (INESC-ID/Instituto Superior Técnico); Daniel Kudenko (University of York); Dave Moffat (Glasgow Caledonian University); Doron Friedman (University College London); Isabel Machado Alexandre (DCTI - ISCTE and INESC-ID); Judith Good (University of Sussex); Lisa Gjedde (University of Education Denmark); Marc Cavazza (Teesside University); Marco Gillies (University College London); Maria Roussou (MakeBelieve.gr); Nicolas Szilas (TEFCA Geneva);
Patricia Azevedo Tedesco (Universidade Federal de Pernambuco, Brazil); Peter Wallis (University of Sheffield); Paul Richmond (University of Sheffield); Ruth Aylett (Heriot-Watt University); Stephane Donikian (IRISA); Stephane Bura (Elsewhere Entertainment); Sue Thomas (De Montfort University)

Player Agency in Interactive Narrative: Audience, Actor & Author

Sean Hammond[1] and **Helen Pain** and **Tim J. Smith**

Abstract. The question motivating this review paper is, how can computer-based interactive narrative be used as a constructivist learning activity? The paper proposes that player agency can be used to link interactive narrative to learner agency in constructivist theory, and to classify approaches to interactive narrative. The traditional question driving research in interactive narrative is, 'how can an interactive narrative deal with a high degree of player agency, while maintaining a coherent and well-formed narrative?' This question derives from an Aristotelian approach to interactive narrative that, as the question shows, is inherently antagonistic to player agency. Within this approach, player agency must be restricted and manipulated to maintain the narrative. Two alternative approaches based on Brecht's Epic Theatre and Boal's Theatre of the Oppressed are reviewed. If a Boalian approach to interactive narrative is taken the conflict between narrative and player agency dissolves. The question that emerges from this approach is quite different from the traditional question above, and presents a more useful approach to applying interactive narrative as a constructivist learning activity.

1 INTRODUCTION AND MOTIVATION

How can computer-based interactive narrative be used as a constructivist learning activity? The question is significant because computer-based narrative is increasingly being used in education: in schools, in corporate training, and elsewhere. In the academic literature some theory does exist that allows us to approach the question, yet not much is known about the learning effects of interactive narrative. Pursuing this question will shed light on new approaches to interactive narrative in education and will inform new designs for interactive narrative environments.

For the purposes of this review, a *constructivist* learning environment is one in which active and critical (not passive and receptive) learning is produced, and in which learners construct their own understanding of the content (they are not led to specific truths by the teacher). A constructivist learning environment involves some degree of structure in order to ensure learning objectives are achieved. But within that structure, the emphasis is on maximising free exploration, interaction, and enjoyment for the learner — maximising *learner agency* — to ensure that learners arrive at their own understanding.

The question of interactive narrative as a constructivist learning activity will be pursued by looking at existing approaches to interactive narrative, and using *learner agency* as a key analytical tool with which to formally classify them. Learner agency is a crucial aspect of constructivist learning, and will be shown to be antagonistic to traditional approaches to interactive narrative. The review concludes by proposing a way to resolve this conflict.

2 A BRIEF INTRODUCTION TO INTERACTIVE NARRATIVE

The model of narrative most frequently found in the interactive narrative literature is that of the structuralist approach to narratology. As Lindley explains, "the model is very useful when applied to the analysis and design of interactive narrative and story construction systems, and the identification of several levels of narrative meaning clarifies the relationships between different strategies for interactive narrative and story construction" [11, p.7]. This structuralist model makes a distinction between a *story*, defined as "the narrated events, abstracted from their disposition in the text and reconstructed in their chronological order, together with the participants in these events" [20, p.3] and the *text*, defined as the "spoken or written discourse which undertakes the telling" of the events of the story [20, p.3]. The reader (or listener) does not have direct access to the story, only to the text, and in the text "the events do not necessarily appear in chronological order, the characteristics of the participants are dispersed throughout, and all the items of the narrative content are filtered through some prism or perspective" [20, p.3]. The word 'narrative' is understood to refer to this text: "The text itself is *the narrative*" [11, p.6]. Although narratology traditionally considers spoken or written narrative fiction, Lindley explains that "the concept of a text has been generalised to cover audio-visual media, since many of the ways narrative functions semiotically are the same across different media forms" [11, p.5]. The motivation for this distinction between *story* and *narrative* is to clarify that "the same story may be expressed in many different narratives, either within the same medium or across different media" [11, p.6].

Meadows gives the following definition of *interactive narrative*:

> "An interactive narrative is a time-based representation of character and action in which a reader can affect, choose, or change the plot. The first-, second-, or third-person characters may actually be the reader." [15, p.62]

The key is that 'interactive narrative' is not merely the presence of interaction and narrative in the same experience. An interactive narrative is understood as an experience in which the reader (player), through meaningful interaction, is able to change the events that occur in the narrative. This can mean affecting the events themselves, or affecting which events occur and which do not, or a combination of both. The interaction can be on a moment-by-moment basis as in 'emergent narrative' (see 'Emergent Narratives' in section 3) or can

[1] ICCS, University of Edinburgh, Scotland, email: S.P.Hammond@sms.ed.ac.uk

consist of fewer decisions with longer-term effects as in a 'branching story' (see 'Modulated Plot' in section 3) or a combination of both.

This definition raises the question of how to define 'plot.' The idea of *continuity of action* by means of *causal relations* between the events represented has traditionally been central to the notion of plot, as Forster's definition shows:

> "We have defined story as a narrative of events arranged in time-sequence. A plot is also a narrative of events, the emphasis falling on causality. 'The king died and then the queen died' is a story. 'The king died and then the queen died of grief' is a plot." [4, p.93]

Alternatively, Meadows describes plot as "the author's planned organisation of the events of the story...a planned topology that has an implied opinion and perspective" [15, p.27].

Forster and Meadows describe two different aspects of causality in the definition of plot. Forster focuses on the chain of cause and effect within the narrative: the queen died because she felt grief because the king died. Meadows focuses on the author's role: the queen died because the author required it to fulfil the needs of the plot. In an interactive plot both aspects of causality are present. The defining property is that the plot consists of chronologically ordered and causally interconnected events.

3 PLAYER AGENCY: AUDIENCE, ACTOR AND AUTHOR

A *player* in an interactive narrative can be a spectator in the sense that she is a witness to the dramatic spectacle. She can be an actor in the sense that she plays the role of one of the characters in the narrative. And she can be an author in the sense that she collaborates with the system (and perhaps with other players) to produce the resulting narrative experience. The player is not exclusively a spectator, nor an actor, nor an author, but in any given example of interactive narrative the role of player combines these three traditional roles to different degrees.[2]

Player agency is a concept that is crucial to the formal nature of interactive narrative as a medium, and that relates interactive narrative theory to learner agency in constructivist learning theory. In the context of interactive narrative, Murray defines agency as:

> "the satisfying power to take meaningful action and see the results of our decisions and choices." [16, p.125]

and Mateas as:

> "the feeling of empowerment that comes from being able to take actions in the [virtual] world whose effects relate to the player's intention" [13, p.2]

Mateas further clarifies that agency is a phenomenal category: it depends "on what's going on in the interactor's head, on what's communicated between the technical system and the person, not only on technical facts like counting the number of system actions that are available at each moment." [3]

The form of agency experienced by an audience member, an actor and an author is different:

Audience: an audience member can critically analyse the narrative (she can think about it) but she has no power to act within the narrative.

Actor: an actor can act within the narrative, from the perspective of one of the characters in the narrative, but only within the limits and from the perspective of the role designed for her.

Author: an author shapes the narrative experience from without, acting on the structures and processes that make up the narrative as an artificial construct in order to express some form or opinion. But an author is limited by the tools at her disposal, her distance from the audience, and her reliance on actors to manifest her intentions and on the audience to comprehend her intentions.

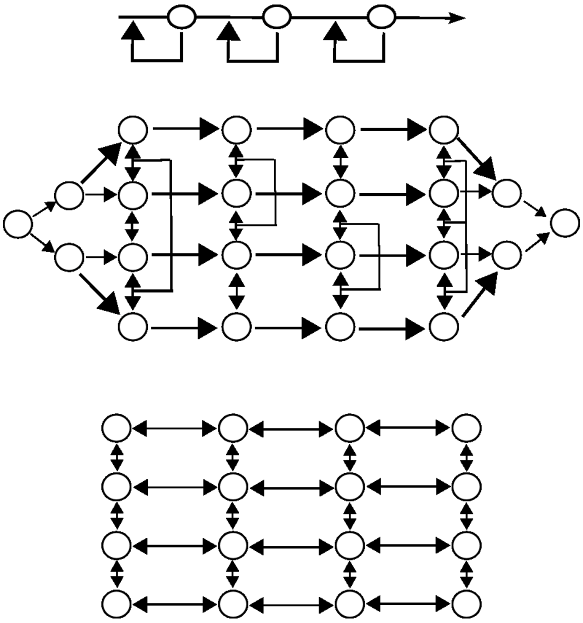


Figure 1. Meadows' nodal (top), modulated (middle) and open (bottom) plot structures [15, p.64]. The lines represent possible plot transitions, the circles represent decision points at which player behaviour can choose between plot transitions.

Meadows and Jenkins provide two classifications of some of the narrative structures and devices available to interactive narrative designers. Meadows describes three plot structures for interactive narrative[4] [15, see figure 1] that exist along a continuum from *Impositional* (the plot is heavily controlled by the game designer, only allowing the player a narrow margin of decisions, or particular moments of interactivity) to *Expressive* (the plot is heavily controlled by the player, the game behaves more like architecture, the player roams freely, explores, investigates, and changes the environment, the breadth of interactivity is much wider but the specifics of a narrative plot are far less defined). The three plot structures represent key descriptive points on the impositional—expressive continuum.

[2] The role of *game designer* is a separate role, distinct from the role of player. Because 'author' is used in this review to denote one of the traditional narrative roles with which the role of player in interactive narrative is described, care has been taken to use the word *designer* rather than author to refer to the procedural designer of an interactive narrative environment.

[3] Michael Mateas commenting on his weblog *Grand Text Auto*, the post is titled *Interaction and Agency* and dated 6th August 2003, http://grandtextauto.gatech.edu/2003/08/06/interaction-and-agency/

[4] Meadows explains that "interactive plot structure is more a system of connections than a curve or arc" [15, p.63], and that these plot structures are an "analysis tool" and "don't have much to do with emotional punch or aesthetic interest." Meadows is aiming to differentiate his plot structures from formal descriptions of plot that focus on dramatic or emotional progression, such as the rising and falling dramatic action of the Aristotelian theory of theatre.

Nodal Plot "a series of non-interactive events, interrupted by points of interactivity" [15, p.64]. This is the most impositional plot structure, with the most support for the classic dramatic arc. Stories of this form have one beginning and two endings. The player fails and must start again from an earlier point in the narrative (this can happen at many points) or the player succeeds and finishes the game. This plot structure provides few affordances for player agency. The player cannot change the direction of the plot, but can only change the pace at which the plot progresses along its linear path. At each decision point, player action decides whether the player fails (and the game restarts from an earlier point in the plot) or succeeds (and the plot progresses).

Modulated Plot player action chooses which path the plot will follow by choosing from finite sets of pre-defined options at fixed decision points in the plot. The player chooses a path through a finite 'plot graph.' These decision points provide affordances for player agency, but their finite nature means that agency is somewhat limited.

Open Plot this structure is "the most expressive for the [player], far less so for the [designer]" [15, p.66], providing the most points of interactivity for the player. The player affects the plot through many small decisions, rather than a few big decisions. The classical dramatic arc may be completely abandoned in the interests of exploration, modification, and investment from the player. The story is usually based on the development of character or the development of environment, or both. The potential for player agency is great. But if the player cannot find meaningful ways to express her intentions on the plot and assess the consequences of that expression, a sense of agency may fail to materialise.

Jenkins describes four devices with which to create "the preconditions for an immersive narrative experience" [7, p.3] in what he calls 'environmental storytelling':

Evocative Spaces an interactive environment can build on stories or genres known to the players, painting the narrative world only in broad outlines and leaving it to the player to fill in the rest. This device provides no affordances for player agency in terms of player *action*, but may provide the player with a degree of agency similar to that of a traditional narrative audience as the player's imagination is given some freedom to help paint the narrative world.

Enacted Narratives an interactive narrative can allow players to perform narrative events. The designer controls the narrative by setting broadly defined goals or conflicts for the characters and inserting localised, non-interactive narrative incidents. The narrative is episodic: "each episode (or set piece) can become compelling on it's own terms without contributing significantly to the plot development" [7, p.6] and within each episode the "sequencing of actions may be quite loose" [7, p.6] allowing for much interaction. This device allows player action to affect the details and ordering of events within an episode, though this freedom is limited by the action constraints of the interactive environment and the higher level plot episodes themselves remain static.

Embedded Narratives Jenkins relates this approach to the traditional detective story. The story is seen "less as a temporal structure than a body of information" [7, p.8]. It is put together, piece by piece, by the player: "narrative comprehension is an active process by which viewers assemble and make hypotheses about likely narrative developments on the basis of information drawn from textual cues and clues." [7, p.8]. The designer controls the progression of the narrative by distributing narrative information throughout the interactive environment. The embedded narrative can be linear while still being closely tied to player agency as the player focuses on discovering and unscrambling narrative elements. The result is two narratives: one controlled by the player as she explores the environment, and another controlled by the designer and embedded in the environment to be discovered.

Emergent Narratives the narrative is not pre-structured but takes shape through game play. The game designer creates "a world ripe with narrative possibilities," "a kind of authoring environment within which players can define their own goals and write their own stories" [7, p.9]. The aim is to provide a form of player agency more similar to that of a traditional author than an actor or spectator.

Taken together the two classifications from Meadows and Jenkins describe a large portion of the approaches to interactive narrative and provide a good introduction to the field.

One way to classify approaches to interactive narrative is to use the concept of player agency to ask to what extent the player is audience, actor, and author in the narrative. In this review these three traditional roles will be used to analyse three theoretical approaches to interactive narrative. Each of the three approaches gives a different way of looking at the three roles, and each positions player agency differently with respect to the three roles.

4 AN ARISTOTELIAN APPROACH TO INTERACTIVE NARRATIVE

Lindley [10, p.2] gives a description of "the central notion of narrative in modern commercial cinema." A narrative of this type has three main parts:

1. A beginning, in which a conflict involving a dilemma of normative morality is established.
2. A middle, in which the consequences of the conflict are played out, propelled by a false resolution of the dilemma.
3. An end, in which the conflict is resolved by an act that affirms normative morality.

Each of these three acts culminates in a moment of crisis, the resolution of which propels the story into the next act (or into the final resolution). The involvement of a central protagonist in the narrative is also key, as is a sense of continuity of action represented by causal connections between events. This narrative structure is known as the three act restorative structure. It is closely related to Aristotle's concept of narrative as an imitation of action that is an organic whole, having a beginning, a middle and an end which fit together naturally and are connected by causes and effects over time.[5] It is also related to Freytag's reworking of Aristotle's model in his Freytag triangle, which expresses a narrative as a function of time in three phases: rising action in which the crisis or complexity of the plot increases, culminating in a dramatic climax, followed by a period of falling action in which the crisis and plot are resolved [6].

In *Poetics* Aristotle organises the different parts that make up a *tragedy*[6] into three hierarchical categories: Objects, Medium and Manner. The objects are the actions (the plot of the drama, made up of causally related events), the characters (the agents of the plot) and the thoughts of the characters that lead to the actions they take in

[5] Aristotle, *Poetics*, 350 B.C.E, available online http://classics.mit.edu/Aristotle/poetics.html

[6] For *Poetics* see previous footnote. *Tragedy* is a form of drama popular in Aristotle's time, involving a conflict between the protagonist and the law, the gods, or society and having a tragic ending.

the drama (if not explicitly described, these thought processes may be inferred by the audience). Medium refers to the medium through which the objects are presented, for example colour and form, voice, rhythm and harmony, or diction and song. Manner refers to the manner of presentation used, e.g. the drama can be narrated or enacted.

With his *neo-Aristotelian theory of interactive drama* [12, 13] Mateas builds on Laurel's application of Aristotle's description of tragedy to human-computer interaction [9] and Murray's description of player agency in interactive narrative [16]. To describe the role of the player in an interactive drama Mateas places *User Action* at the level of character in the Aristotelian hierarchy. That is, the player acts in the drama as one of the characters in the drama, and when the player takes action in the drama "The player's intentions become a new source of formal causation" [13, p.4] in the model that was not present in Aristotle's original model.

To support this, Mateas explains that the player's intentions are constrained by the *material for action* provided by the system "The only actions available [to the player] are the actions supported by the material resources present in the game" [13, p.4] and by *formal constraints* that provide the player with dramatic reasons to want to take particular actions: "the formal constraints afford *motivation* from the level of plot" [13, p.4].

An example from Mateas and Stern's interactive drama Façade [14] will illustrate the Aristotelian approach. In Façade, the player takes on the role of a character in the drama and sees from the first-person view of this character. Dialogue is the main form of interaction: the player communicates with the virtual agents by typing text, the virtual agents communicate by sequencing pre-recorded soundbites and with facial expressions and hand gestures.

The Façade architecture is an attempt to break free of the plot structures and narrative devices described by Meadows and Jenkins (see section 3). Façade dynamically sequences dramatic beats from a large library. Each beat is a small collection of interactive, coordinated behaviours to be carried out by the agents of the drama, and is tagged with preconditions for selection and the consequences of each potential beat outcome on the dramatic arc of the drama. The beats can be reordered in many ways while remaining coherent, and any play of the drama need only contain a subset of the available beats. Façade attempts to select a coherent and dramatically 'good' sequence of beats while remaining responsive to player action.

The premise of the drama is that you (the player) have been invited over to the apartment of Grace and Trip (the virtual agents). The short drama takes place in the apartment, where soon after you arrive it becomes obvious that Grace and Trip's marriage is on the rocks. What happens depends partly on your actions in the 5-15 minutes that make up the drama.

Figure 2 is a transcript of an interaction with Façade [1]. The player is controlling the character named Audrey in the transcript, and Grace and Trip are the virtual agents. There are two things to notice in the transcript. First, when the player types an input that the system does not understand the agents try to gloss over the failure by acting briefly confused, then continuing with the intended narrative, ignoring the unwanted input. Second, as can be seen in the last two lines of the transcript, the agents respond to keyword triggers. The player inadvertently triggers the 'sex' topic. This topic is not supposed to come up until later in the drama, so Trip tries to redirect the player onto the topic of drinks, again trying to continue with the intended narrative despite the unwanted input from the player. If the player persists in her uncooperative behaviour, the agents will close the door on her and the game will be over. As the player who produced this transcript commented, "don't ever go to this apartment in case of emergency." [7]

(Audrey knocks on the front door.)
(Trip opens the front door.)
TRIP: Audrey!!
AUDREY: TRIP I'VE BEEN SHOT!
TRIP: Uh...
TRIP: Well come on in...
TRIP: Uh, I'll – I'll go get Grace...
GRACE: Audrey, Hi! How are you? I'm so happy to see you after so long! – (interrupted)
AUDREY: CALL 911
GRACE: Uh...
GRACE: So, come in, make yourself at home...
AUDREY: OH, F**K THIS
TRIP: Ha ha! Oh I think we're going to need some drinks first if we're going to talk about sex.

Figure 2. An edited transcript of an interaction with Façade [1]

The tendency in the Aristotelian approach to interactive narrative is to try to hide the underlying mechanics of the experience and maintain the player's 'suspension of disbelief.' In this approach, the player's role is something like that of a *passive spectator* and that of a *constrained actor*. The interactive narrative tries to "steer not only a players' action and emotions, but their perceptual behaviour and conceptualisation of events" [18, p.3] and to transport the player into the artificial reality: "the quest is to provide more immersive, more engaging and more affective experiences" [18, p.1].

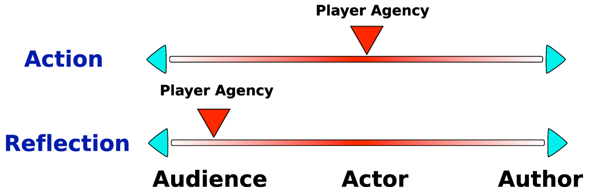


Figure 3. The role of player agency in the Aristotelian approach to interactive narrative

The key aspect here is the role of player agency in the Aristotelian approach, described by figure 3. The player *acts* from the perspective of an actor within the narrative structure with a limited range of actions. The player *reflects* on the narrative as a passive spectator, from a perspective within the narrative, thinking what her character thinks and feeling what her character feels. Player reflection is embedded within the artificial representation of reality that is the interactive narrative. To clarify, imagine the modulated plot structure that was described in section 3. In the Aristotelian approach, the player acts from the perspective of one of the characters in this narrative structure, choosing from finite options at certain points in the plot. The player is an actor within the narrative. But the game designer uses drama and spectacle to try hide this underlying plot structure from the player, so that the player does not perceive the limits within which the experience has been designed for her. Alternatively, in terms of the enacted narrative device (described in section 3), the designer guides the player's progression through the narrative by setting the player's global goals and interrupting free interaction with fixed, non-interactive plot incidents. Again the player acts within the limits defined by the designer, and the designer aims to use drama

[7] Internet forum post, accessed January 2nd, 2007, http://forums.idlethumbs.net/showthread.php?t=2895&page=2&pp=25

and spectacle to prevent the player from becoming too aware of this restriction. In both examples, player reflection on the narrative structures is passive and receptive.

In this Aristotelian approach the balance of power between game designer and player is antagonistic to player agency: player agency is inevitably restricted and the player manipulated to distract attention from this restriction. The player is given a limited role in the experience. Within the Aristotelian approach there is no solution to this problem: as the player's interactive freedom increases, the system needed to support the interaction becomes more complex, and quickly impossible. An interactive narrative cannot "be all things to all players" [1]. To resolve the conflict with player agency, alternative approaches at the formal level must be considered.

5 A BRECHTIAN APPROACH TO INTERACTIVE NARRATIVE

German dramatist Bertolt Brecht Brecht argued that the Aristotelian approach to theatre, by focusing on illusion and empathy and a passive role for the audience, places the audience in a receptive state of mind in which they are encouraged to passively accept a fictional representation of reality. In response, Brecht created a theory of theatre, the *Epic Theatre*, in which the audience are discouraged from becoming empathically immersed with the action and characters on stage, and encouraged to form a distanced, critical relationship with the drama instead. Where Aristotle employs empathy, catharsis and illusion to transport the audience into the drama, Brecht employs techniques designed to prevent empathy and catharsis and break the illusion, to get the audience to reflect on the drama as an artificial representation. Brecht's techniques are used to alienate or distance the audience from the drama, reminding them that they are witnessing an artificial representation, and drawing critical attention to the function of the drama and the real-world issues being represented.

Pinchbeck applies Brecht's thought to modern First-Person Shooter (FPS) computer games. He argues that "Successful immersion implies, by definition, an acceptance of the rules of the artificial experience at a perceptual and behavioural level" and that these rules "are both vastly simplified and highly structured" [18, p.7]. The effect is that "users are steered towards an uncritical relationship with the affordances of the experience, even though these affect the scope of available actions as much as the content" [18, p.7]. To support this, drama is used "to detract attention from the manipulation towards an increased engagement with the reduced corridor of affect of the narrative structure" [18, p.7].

Pinchbeck suggests applying Brecht's theatre techniques to computer-based narrative, embedding devices into the game experience that reveal its innate tendencies without altering its fundamental form. The aim is "to force an audience to consider the implications of the action in the real world by highlighting the artifice and displacement of control within an artificial reality" [18, p.9]. Specifically Pinchbeck suggests pausing the game experience and using in-game narration and music to break immersion and promote critical reflection.

America's Army is an online multiplayer FPS game in which players take on the role of U.S. soldiers from a first-person perspective in combat scenarios. It is an example of Aristotelian interactive narrative, just the sort of thing Brecht might try to subvert. *Dead in Iraq*[8] is an in-progress 'online gaming intervention' being conducted by Joseph DeLappe of the University of Nevada Reno. DeLappe's intervention is an example of how the Brechtian approach could be applied to interactive narrative. DeLappe enters the online gaming environment of America's Army and uses the games text-messaging system, through which players can type messages to each-other as they play, to type the names of U.S. soldiers who have been killed in Iraq. By taking screenshots of the game that show the most recent messages from players at the time of the screenshot, DeLappe collects players' responses to his intervention (figure 4).

```
- i think they are dates of deaths of
  soldiers. are those real people??
- are you enlisted? reserve? have you been to
  iraq?
- u arent encouraging me to join the services
- bin-lad-en: i am srry
- i dunno ..was thinkin of joinin the army
  soon
- its propaganda
```

Figure 4. Selected players' responses to DeLappe's 'online gaming intervention' *Dead In Iraq*.

As the responses show, DeLappe's intervention, considered as an attempted Brechtian technique,[9] has been successful to some extent. The players' comments show some discussion of the real world consequences of the fictional actions, consequences which are not sufficiently represented in the artificial experience. But this approach is limited: DeLappe is not formally modifying the interactive medium itself, he is merely doing something novel within it.

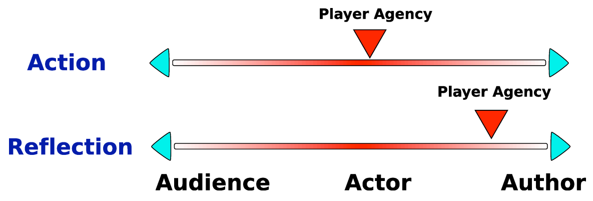


Figure 5. The role of player agency in the Brechtian approach to interactive narrative.

Figure 5 describes the key conclusion: the role of player agency in the Brechtian approach to interactive narrative. The player *acts* from the perspective of an actor within the narrative with a limited range of actions. As in the Aristotelian approach the player may find herself acting from within a modulated plot structure, choosing from fixed options at fixed points in the plot, or she may find herself acting within global goals and fixed plot incidents setup by the designer to guide the experience following an enacted narrative approach. But in the Brechtian approach the player *reflects* on the narrative from a perspective similar to that of an author, from *outside* of the narrative construct, reflecting on the structures and processes that make up the experience as an artificial representation. The player may reflect on the designed plot structure or global goals and non-interactive plot incidents, and the perspective this representation presents of the reality being simulated. The player need not necessarily accept the designer's perspective. The Brechtian approach changes the perspective of player reflection, so that manipulation of the player by the game designer is reduced. But the perspective of player action remains unchanged, so the player remains in a limited role in the experience. Ultimately, this is the limit of the Brechtian approach: the game designer tries to get the player(s) to reflect on the interactive narrative

[8] http://www.delappe.net/

[9] DeLappe himself does not relate his intervention to Brecht

as an artificial representation, rather than to accept it as reality, but retains control over player actions as in the Aristotelian approach. The Brechtian approach does not formally resolve the conflict between narrative control and player agency. An approach that formally modifies the experience is needed to give player agency a greater role in the narrative.

6 BOAL'S THEATRE OF THE OPPRESSED

A theatrical approach that may provide a suitable model for interactive narrative is Brazilian director Augusto Boal's *Theatre of the Oppressed* [2], which is used in radical popular education movements. The aim of Theatre of the Oppressed is "to change the people — "spectators," passive beings in the theatrical phenomenon — into subjects, into actors, transformers of the dramatic action" [2, p.122].

One of the interesting forms of Theatre of the Oppressed is the *Forum Theatre*. An example Forum Theatre, 'It's Too Late,' will illustrate the form. 'It's Too Late' is a short improvisational play. The stage contains three desks, and a clock on the wall. Three actors, 'the oppressors,' play clerks standing behind the desks. A fourth actor, 'the oppressed,' plays a citizen who enters the stage carrying a document, with the goal of using the document to complete a transaction with the oppressors. The rules of the improvisation are that the oppressed must visit each desk in turn and try to enact the transaction with the oppressor. The oppressors must find ways to deny the request based on the idea that 'it's too late.'

A scripted version of the play is first presented to the audience by the actors. This version ends badly — the oppressed is turned away without completing the transaction. In this version, the oppressed makes at least one clear social or political error in trying to solve the oppression. This version, called the 'anti-model,' presents a problematic view of the world to the audience. The audience are asked if they agree with the solutions advanced by the protagonist, with the expectation that they will not.

The actors then act out the play again, but this time audience members are instructed that they may put up their hand at any time to freeze the play and take on the role of the oppressed. An audience member, or 'spect-actor,' goes onto the stage when he or she feels the oppressed is making a mistake and replaces the actor playing the oppressed for a time, to try to enact a better solution to the problem. As soon as a spect-actor enters the stage the oppressors intensify their oppression, responding to the spect-actors solutions with new forms of the oppression. The actor who has been replaced moves to the side of the stage and verbally assists the spect-actor to stay in role and encourages him or her to continue attempting solutions in the face of adversity. The Forum Theatre becomes a creative game or competition which pits spect-actors against actors. The actors try to force the spect-actors to accept the world as it is, as it was presented in the anti-model. The spect-actors try to find a solution, to change the world. A sense of urgency is vital to this game. The actors, when playing oppressors or oppressed, move the narrative toward the same ending as in the anti-model. To prevent this ending the spect-actors must continuously fight the oppression until they break it.

The improvisation may be repeated several times over, and in this way the actors and spect-actors creatively discuss and enact an oppressive problem and potential solutions to the problem. In the example play 'It's Too Late,' potential solutions include: the oppressed demands to be given her rights, the oppressed tries to make friends with the clerks and convince them to give her what she wants, and the oppressed tries to use money to bribe one of the clerks.

The aim is not to produce a well-formed piece of theatre or even a solution to a problem.[10] The aim is to produce a good debate through active, critical thinking, exploration and enactment, and to empower the spect-actors through this enacted debate. The key is to realise that Theatre of the Oppressed is not simply a form of interactive drama. The drama provides a place of fiction in which spect-actors train themselves for action in the real world. As Boal puts it, the aim is "to transform the spectator into the protagonist of the theatrical action and, by this transformation, to try to change society rather than contenting ourselves with interpreting it" [3, p.224].

This approach immediately seems more suitable for the computer-based interactive narrative medium. Aristotle and Brecht's approaches are non-interactive theatre, and as such may not present the most useful models for an interactive medium. Player agency has to be 'incorporated' into the model or 'dealt with' in some way. Boal's is a fundamentally interactive form of theatre, inspired by Brecht's approach, but attempting to go one step further.

In the Aristotelian approach, the fictional character both acts and thinks for the spectator. The effect of a successful Aristotelian experience is to subdue the spectators' desire for agency.[11] In the Brechtian approach the character acts for the spectator, but the spectator thinks for herself, and may "think in opposition to the character" [2, p.122]. A Brechtian experience encourages the spectators' desire for agency: the aim is to produce critical discussion among spectators about the actions and decisions taken or not taken by the characters. Boal's theatre "focuses on the action itself: the spectator delegates no power to the character (or actor) to act or think in his place; on the contrary, he himself assumes the protagonic role, changes the dramatic action, tries out solutions, discusses plans for change" [2, p.122]. In a Theatre of the Oppressed the spectators' desire for agency is not only encouraged but actually *exercised* as spectators act within the safe, fictional environment of the drama. This fictional exercise of agency leaves behind the desire in the spectator to exercise that same agency in real life.

In Forum Theatre, a spect-actor can replace and act in place of any oppressed character[12] at any point in the play, dropping in and out of the characters as she pleases. A spect-actor is not restricted to acting from the perspective of one character, or acting within the role of one character. The role of a spect-actor in Forum Theatre is greater, in terms of agency, than the traditional role of an actor playing a single character.

Each spect-actor is constrained in two ways: by the reactions of the actors and other spect-actors to her actions on stage, and by the facilitator of the forum (the 'joker').

The spect-actors considered as a whole reshape the entire drama over several iterations. They act on the drama from an outside perspective, similar to the way in which a traditional author shapes a drama. But even the spect-actors as a group are limited by the framework set out for them. So it is not accurate to say that the spect-actors have authorship over the narrative. Rather, they have a form of

[10] This does not mean that a Forum Theatre should not be well-formed, Boal says "The most important thing, over and above anything else, is that Forum Theatre should be good theatre; that the model in itself offers a source of aesthetic pleasure. Before the 'forum' part begins, the show itself must be watchable and well constructed" [3, p.277].

[11] Think of watching a good Hollywood movie in the cinema. If you're enjoying the film and are fully immersed in the characters and action, then you don't want it to end. When the film does end and the lights come back on, you have to consciously 'drag' yourself back into reality.

[12] The example used earlier has one oppressed character and three oppressors. But many forum theatres have multiple oppressed characters, and may have characters who are both oppressor and oppressed, and who mutually oppress each other. Usually spect-actors cannot replace purely oppressive characters, as this breaks the game and results in nonconstructive solutions.

agency which has more in common with the agency experienced by a critical author than it does with the agency experienced by a passive spectator.[13]

6.1 A Boalian Approach to Interactive Narrative?

In his thesis *Videogames of the Oppressed: videogames as a means of critical thinking and debate* [5] Frasca envisions a new approach to interactive computer games: “a powerful representational form that encourages critical thinking, empowerment and social change” [5, p.114]. Frasca makes an analogy between Boal’s Forum Theatre and simulation in computer games:[14]

> “Literally, what happens in a [Forum Theatre] session is a simulation. It is not the representation of something, but the simulation of how some situation would happen, depending on many factors. It analyses the world “as it is and as it could be” (Boal, 1992)” [5, p.67].

Frasca further explains that Forum Theatre is “a meta-simulation, an environment where spect-actors can create and question the rules of a simulation” [5, p.73]. Frasca proposes a new approach to interactive computer games in which the players have access to the rules of the simulation, and can alter them. He explains that “Since simulations are representations of the world, they cannot model it without conveying the [designer]’s idea about how the world works” [5, p.79]. Frasca proposes that like the spect-actors in a Forum Theatre construct different ideas about a problem and its solutions in successive iterations of the play, players could discuss a situation by constructing successive simulations that model the situation as a game.[15]

Combining Frasca’s analogy between simulation and Forum Theatre with the review of interactive narrative presented in this paper, a Boalian approach to computer-based interactive narrative can be proposed. A Boalian approach to computer-based interactive narrative would give the player(s) access to the underlying story model to interact with directly and deliberately, to *play* with. It should blur the traditional interactive narrative roles of player and author into one. The player could jump seamlessly and at will between acting within the interactive narrative, in the role of the protagonist (or the oppressed) in the story, and acting on the interactive narrative from outside of it, manipulating the story model underlying the narrative, in the role of author. The player-authors *construct* and *experience* the interactive story at once.

Figure 6 describes the key conclusion: the role that player agency might play if the Boalian approach can be applied to interactive narrative. The player both *acts* and *reflects* on the narrative from a perspective similar to that of an author, from *outside* of the narrative construct, acting and reflecting on the structures and processes that make up the narrative as an artificial representation. Boal writes of turning passive spectators into actors. Here he is referring to the creative, critical, improvisational actors of his theatre of the oppressed. He does not consider passive actors who merely act out a role as written by an author. Applied to interactive narrative, Boal’s passive spectator corresponds to the role of player as passive actor as in the Aristotelian approach to interactive narrative. Boal’s spect-actor (spectator elevated to actor) corresponds to the player elevated to co-author of the narrative with the designer of the interactive environment.

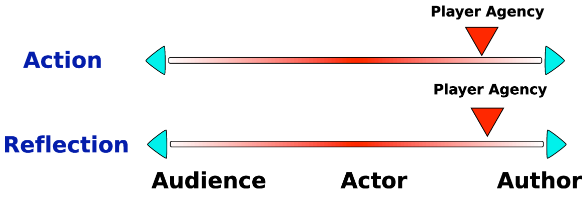


Figure 6. The role of player agency in the Boalian approach to interactive narrative.

A story-model based on a nodal or modulated plot structure (section 3) seems the most obvious candidate for this approach. When in the role of actor, the player controls a character within the narrative, and may make fixed decisions at fixed points within the plot structure that drives the interactive narrative. When in the role of author the underlying plot structure is presented to the player directly, through an interface which allows the player to manipulate the structure itself. The player iteratively constructs or modifies a story by switching at will between these two roles, changing the story model, experiencing the result, changing the story model some more, and so on.[16]

This approach is non-immersive, emphasises the artificial, constructed nature of the interactive narrative, and focuses player agency on the structures and processes underlying the experience. Of the three approaches presented, the Boalian approach seems most appropriate to the constructivist motivation. Because learners are actively involved in constructing an interactive story, the form of learning is the most active and critical, least passive and receptive, of the three approaches. Learners construct their own understanding through exploring and interacting with the system. Not only are they active participants in the narrative, but the learners are fully aware of why they are participating. The Boalian approach is dialectical, not didactic as the Aristotelian approach is. It does not present a solution or model to be followed, instead it presents an anti-model to be debated. Some structure is inherent in the interaction with the envisioned system. The player-author is given a particular plot model and character roles to use as the building blocks of an interactive story, and can only construct what these building blocks, created by the designer of the environment, will allow. Yet by focusing player action *on* the underlying story model, rather than having the player act *within* this structure, player agency is maximised. The inherent conflict between narrative and player agency dissolves.

Such an interactive story player-authoring environment could be used in a *constructionist* [17] approach to learning. Players learn about the models, structures and processes, and modes of authoring that underlie interactive stories through constructing interactive stories. The constructed stories can then be *played* (with the authoring interface disabled) by peers as part of a peer review process.

The application of Boal’s techniques could be fundamental to using this story construction process as a means to collaboratively discuss social issues. This aspect is most clear if you imagine the players

[13] In practice it is sometimes the spect-actors who devise a Forum Theatre for themselves to take part in, so that they have both authorship and agency over the Forum Theatre.

[14] Frasca presents a four-part semiotic model of simulation, which focuses on the process of an observer interpreting a simulation, with which he relates Forum Theatre to simulation [5, p.79].

[15] Specifically, Frasca describes a game derived from the popular series *The Sims* in which players would have access not only to surface characteristics of the game characters, but to the rules that govern character behaviours. Players would use these rules to construct models of problematic social situations and their solutions.

[16] Propp’s *Morphology of the folktale* [19] may provide an ideal basis for constructing a story model for this approach. His description of the plot structure of folktales lends itself well to forming the building blocks of nodal or modulated plots, and he also provides clear descriptions of character roles and their actions with respect to the plot. Kashani [8] provides an excellent example of Propp’s morphology applied to an interactive story environment using a nodal plot structure.

given an interactive story that presents a problem, an oppression of the player/protagonist character of the story. Players then discuss solutions to the problem through a series of modifications to the model underlying the interactive story. The process might be conducted as a workshop, with a person facilitating an interaction between several player-authors and a single interactive story environment.

The intention is not to claim that an interactive story authoring environment which attempts to combine the roles of game player and game designer will be a Boalian Forum Theatre applied to the digital medium. There are many ways in which this learning process will differ from Forum Theatre, and understanding these differences may be more useful than understanding the similarities. The question of how the virtual environment is used in the real world, how the learning experience goes on *around* the artifact, is crucial. The claim here is that computer-based interactive narrative is at the intersection between Boal's Forum Theatre and Papert's constructionism. Applied to interactive narrative, the two provide a promising approach.

7 CONCLUSION

When an Aristotelian approach is applied to interactive narrative the aim is for the system to deliver a well-formed narrative experience to the player. A conflict with player agency that necessitates putting the player in a passive role is inherent in this aim. The player *acts* from the perspective of a constrained actor within the narrative. But the player is encouraged to *reflect* on the narrative from the perspective of a passive spectator. This disparity between the perspectives of player agency in terms of action and reflection necessitates an attempt to maintain the player's 'suspension of disbelief' and to manipulate player perception and action, keeping them within the designed range of possibilities.

A Brechtian approach breaks 'suspension of disbelief' intentionally, aiming to highlight the artificiality of the experience. The player still acts from the perspective of a constrained actor within the narrative, but reflects on the narrative from a perspective outside of it, reflecting on the narrative as an artificial representation of reality.

A Boalian approach builds on the Brechtian approach by changing the perspective of player action to match that of player reflection. The player both acts and reflects on the narrative from an outside perspective, acting and reflecting on the story model from which the narrative is constructed. The aim is no longer to maintain a good narrative experience in spite of player agency, but to provide the player with the narrative construction kit most productive of player agency.

This review argues that the form of player agency in interactive narrative improves, with respect to the motivation of constructivist learning, as we move from an Aristotelian, to a Brechtian, to a Boalian approach.

The traditional question driving research in interactive narrative is: how can an interactive narrative environment deal with a high-degree of player agency, while maintaining a coherent and well-formed narrative? This question expresses the approach categorised here as Aristotelian interactive narrative. If the approach categorised as Boalian interactive narrative is taken, the question becomes quite different: how can an interactive narrative environment provide a story model that supports creative and critical expression through constructing interactive stories? This question motivates further research into four more specific questions: what kind of story model best supports creative and critical expression through constructing interactive stories? How can we design an interface and interface metaphors that allow intuitive interaction with this story model? How can we seamlessly combine the role of actor and author into one role for the player? How can a learning experience be structured within and around this virtual environment?

ACKNOWLEDGEMENTS

I would like to thank Mark Steedman of the Institute for Communicating and Collaborative Systems, Judy Robertson of Heriot Watt University, and John Sykes and the Games Interest Group of Glasgow Caledonian University for feedback and discussion on earlier versions of this paper, and to acknowledge the financial support of EPSRC.

REFERENCES

[1] Ernest Adams, 'A New Vision for Interactive Stories', in *Game Developers Conference, San Jose, California*, (2006).

[2] Augusto Boal, *Theatre of the Oppressed*, London: Pluto Press, 1979.

[3] Augusto Boal, *Games for Actors and Non-Actors*, London: Routledge, 1992.

[4] E.M. Forster, *Aspects of the Novel*, Hammondsworth: Penguin, 1963.

[5] Gonzalo Frasca, *Video Games Of The Oppressed: Video Games As A Means For Critical Thinking And Debate*, Master's thesis, School of Literature, Communication and Culture, Georgia Institute of Technology, 2001.

[6] Gustav Freytag, *Technique of the Drama: An Exposition of Dramatic Composition and Art*, Johnson Reprint Corp (June 1968), originally published in 1863.

[7] Henry Jenkins, 'Game Design as Narrative Architecture', in *First Person*, eds., Pat Harrington and Noah Frup-Waldrop, Cambridge: MIT Press, (2002).

[8] Shahryar Attar Kashani, *Dynamic Storylines in Interactive Virtual Environments*, Master's thesis, The University of Edinburgh, School of Informatics, 2004.

[9] Brenda Laurel, *Computers as Theatre*, Addison-Wesley Professional, 1991.

[10] Craig A. Lindley, 'The gameplay gestalt, narrative, and interactive storytelling', in *Computer Games and Digital Cultures Conference, Tampere, Finland*, p. 203, (2002).

[11] Craig A. Lindley, 'Story and narrative structures in computer games', in *sagas_sagasnet_reader: Developing Interactive Narrative Content*, ed., Brunhild Bushoff, HighText-Verlag, (2005).

[12] Michael Mateas, 'A neo-aristotelian theory of interactive drama', in *Working Notes of the AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment*, AIAI Press, (2000).

[13] Michael Mateas, 'A preliminary poetics for interactive drama and games', in *Digital Creativity, volume 12, number 3*, pp. 140–152, (2001).

[14] Michael Mateas and Andrew Stern, 'Façade: An Experiment in Building a Fully-Realized Interactive Drama', in *Game Developers Conference, Game Design Track, San Jose, California*, (March 2003).

[15] Mark Stephens Meadows, *Pause & Effect: The Art of Interactive Narrative*, New Riders Press, 2002.

[16] Janet H. Murray, *Hamlet On The Holodeck: The Future of Narrative in Cyberspace*, The MIT Press, 1997.

[17] Seymour Papert, *Mindstorms: Children, Computers, and Powerful Ideas*, The Harvester Press Limited, Great Britain, 1980.

[18] Dan Pinchbeck, 'A theatre of ethics and interaction? Bertolt Brecht and learning to behave in first-person shooter game environments', in *"Using drama and storytelling for innovative educational technology," special session, Edutainment 2006, Zhejiang University, China*, eds., Zhigeng Pan, Ruth Aylett, Holger Diener, Xiaogang Jin, Stefan Göbel, and Li Li, volume 3942 of *Lecture Notes in Computer Science*, pp. 399–408. Springer, (April 2006).

[19] Vladimir Propp, *Morphology of the Folktale*, University of Texas Press; 2nd edition, June 1968.

[20] Shlomith Rimon-Kenan, *Narrative Fiction Contemporary Poetics*, Routledge, 1983.

Interactive Generation of Dilemma-based Narratives

Heather Barber and **Daniel Kudenko** [1]

Abstract. This paper presents a system which automatically generates interactive stories. These are focused on dilemmas in order to create dramatic tension. The story designer is only required to provide genre specific storyworld knowledge, such as information on characters and their relations, locations and actions. In addition, the system is provided with knowledge of generic story actions and dilemmas which are based on those clichés encountered in many of today's soap operas. These dilemmas and story actions are instantiated for the given storyworld and a story planner creates sequences of actions that each lead to a dilemma for a character (who can be the user). The user interacts with the story by making decisions on relevant dilemmas and by freely choosing their own actions. Using this input, the system chooses and adapts future story lines according to the user's preferences.

1 INTRODUCTION

In recent years computer games from most genres have included a progressive story line to increase the immersive experience of the user and their enjoyment of the game. However, stories are often linear, and in almost all cases pre-defined, which reduces the replay value of these games. Research into interactive narrative generation (or interactive drama) tries to overcome these weaknesses. Most interactive drama systems (prominent examples include [15, 2, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14]) are focused on generating short story lines and do not adapt to the user (see Section 13 for exceptions).

In this paper, we propose a system that generates interactive stories which are long (potentially infinitely so), and that adapt to the user's behaviour. To add dramatic tension, the story incorporates dilemmas as decision points for the user. These dilemmas are based on the clichés found in many contemporary soap operas, such as the trade-off between personal gain and loyalty to a friend. Overarcing stories connect these dilemmas as points of interaction within a coherent plotline that is dynamically created, based on the user's response and action choices.

Our goal is to keep the story designer's input to a minimum and the user involvement as high as possible. In the proposed system, the story designer provides the story background in the form of character information and other knowledge that relates to the world in which the story is to be created (e.g. the east end of London). The system then instantiates all generic knowledge on story actions and dilemmas accordingly and thus creates the narrative in collaboration with the user's actions. A considerably less interactive version of the system discussed here – with dilemmas only presented to the user – was introduced in [1].

This paper is structured as follows. First a general overview of the system is first given, followed by a discussion of the story background representation. We proceed with a description of dilemmas; the story generator; integrating and responding to user actions; non-user dilemmas; and the user modelling component. The paper finishes with a brief overview of related work and conclusions.

2 SYSTEM OVERVIEW

The interactive drama knowledge base consists of: the storyworld (which contains information regarding the characters); story actions; and dilemmas which can occur in the storyworld. This information is partially genre dependent and provided by the story designer, with the remainder being hard coded. These components are drawn upon in the generation of a narrative through planning. The user is able to interact with the narrative generator, and their actions effect the story experienced. A user model is employed to ensure that the story's dramatic interest is maximised. The interactions between the system components are shown in figure 1. Each of these components is discussed further in the following sections.

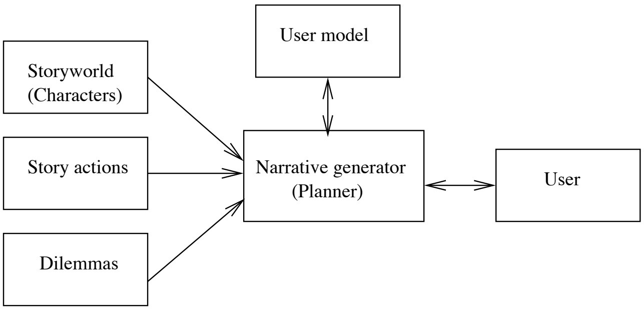


Figure 1. This figure shows the components of the system and how they interact.

3 THE STORYWORLD

The storyworld consists of characters and locations at which the characters can be. These characters have various associated traits, as detailed here.

- Each character's associated domain independent attributes can include information such as attractiveness, gender, sexuality and age group.
- Characteristics are slightly more variable, for example: generosity, morality and selfishness.

[1] University of York, Heslington, York, YO10 5DD email: {hmbarber, kudenko}@cs.york.ac.uk

- It is possible to specify particular personalities, such as bad_boy and busybody. These genre specific character descriptions are those which are not fully deducible from other character traits and which relate to specific storylines within the current domain.
- Characters have storyworld relationships with one another, including friendship and love. They are able to disapprove of one another's partnerships. This can be for any one of a variety of reasons, including an age difference or snobbery. Relationships are unidirectional and have an associated strength, although feelings of one character for another affect the reciprocity.
- The characters hold storyworld principles, such as monogamy, which make their behaviour more believable. Under specified pressures and circumstances, principles can be broken (or their associated strength of belief reduced). Characters also have aspirations, for example wanting a baby. These principles and aspirations affect which actions a character participates in and the dilemmas in which they become involved.

A range of values is associated with each attribute and characteristic. A character's nature affects which actions they can participate in and also, ideally, the user's opinion of that character. The character's personal traits should be apparent to the user from the way the character acts within the storyworld. Each character should act in a manner which is consistent with their traits and how they have acted previously, while at the same time avoiding predictability.

A series of genre-specific locations are required by the storyworld. At any given time in the story, each character is at one of these locations. Direct interactions between characters can only take place if they are at the same location.

4 ACTIONS

Those actions which can take place within the storyworld must be specified for each domain. Every possible action should be included and although these vary between domains there remains a significant overlap.

The domain specific storyworld actions can include characters falling in love, becoming pregnant and being involved in crimes – such as drugging or murder. Each of these actions has associated conditions which must be satisfied before execution (preconditions) and effects which represent changes to the storyworld following execution. For example, the action of a character moving between locations l and k has preconditions of the character being at location l and there existing a path between locations l and k. The effects of this action are that the character is at location k and is no longer at location l. This follows the STRIPS representation.

Before an action is made available to the system for use within a storyline an applicability check is carried out. This ensures that the action is of the type that the acting character is likely to make. For example, a more attractive character can start to fancy a very generous character. A character's attributes, characteristics and personalities affect which actions are possible for that particular character as an action can only be utilised if its applicability is high enough for that character.

The user is able to specify their own actions within the scope of the current genre. This is discussed further in section 7.

5 DILEMMAS

Field [7] states that "drama is conflict", that the dramatic interest in a story centralises on its conflicts. In genres which make use of clichéd storylines these are usually found to be essentially conflicts (or dilemmas). Writers utilise these dilemmas in the creation of stories. A general form of each such clichéd dilemma can be determined, and a computerised storywriter can create an interactive drama around these.

Since the focal point of an interactive drama is the user, each dilemma should represent a conflict to that user. Within the course of the experience, they will be required to make fundamentally difficult decisions which will have negative outcomes whatever choice they make. There may also be decisions in which the user has to decide how to distribute limited benefits in different areas or to different characters.

Our experience showed that when more than two characters were involved in a dilemma, it was either expandable to multiple two character dilemmas, or the characters receiving payoffs naturally divided into two groups with the same resultant utility. Therefore a user decision on a dilemma will involve only two recipients of utility payoffs. Five such dilemma categories were identified. These do not consist of all payoff matrices for two users, as many such matrices would not involve a dilemma for the character making the decision. The relevant categories are: Betrayal (dilemma 1), Sacrifice (dilemma 2), Greater_Good (dilemma 3), Take_Down (dilemma 4) and Favour (dilemma 5). In order to involve a dilemma for the user, these may require characters to be friends or enemies. Where relevant, this is stated with the dilemma utility matrices in dilemmas 1 to 5.

In these dilemmas: A_X represents the decision of character X being to take action A; u_C^i represents the utility of character C for the respective action; and i denotes the relative value of the utility, i.e., u_C^1 is greater than u_C^2.

$$\begin{array}{c|c} A_X & (u_X^1, u_Y^2) \\ \hline \neg A_X & (u_X^2, u_Y^1) \end{array} \wedge friends(X,Y)\ - Betrayal \qquad (1)$$

A character having the opportunity to be unfaithful to their partner is an example of the Betrayal dilemma.

$$\begin{array}{c|c} A_X & (u_X^2, u_Y^1) \\ \hline \neg A_X & (u_X^1, u_Y^2) \end{array} \wedge friends(X,Y)\ - Sacrifice \qquad (2)$$

An example of the Sacrifice dilemma occurs when a character has committed a crime which their friend has been accused of. Here a character has the opportunity to admit to their crime and thus accept the punishment rather than allowing their friend to take the blame.

$$\begin{array}{c|c} A_X & (u_X^1, u_Y^1) \\ \hline \neg A_X & (u_X^2, u_Y^2) \end{array} \wedge enemies(X,Y)\ - GreaterGood \qquad (3)$$

A character deciding whether to give something (such as information or a friend) to their enemy Y in order to save themself (and possibly also their family) would be experiencing the Greater Good dilemma.

$$\begin{array}{c|c} A_X & (u_X^2, u_Y^2) \\ \hline \neg A_X & (u_X^1, u_Y^1) \end{array} \wedge enemies(X,Y)\ - TakeDown \qquad (4)$$

A character deciding whether to injure (or even kill) their enemy in full awareness that they will receive a punishment for this crime would be involved in the Take Down dilemma.

$$\begin{array}{c|c} A_X & (u_Y^1, u_Z^2) \\ \hline \neg A_X & (u_Y^2, u_Z^1) \end{array}\ - Favour \qquad (5)$$

When presented with a favour dilemma the character making the decision will not receive any direct utility from their action regardless of their choice. An instance of this dilemma occurs when a character must choose between potential partners. It is necessary that there is no discernible benefit to the character making the decision of choosing one partner over the other.

As can be seen, dilemmas 1 and 2 are the inverse of one another, as are dilemmas 3 and 4. This means that any dilemma which falls into one of these categories can be inverted to become a dilemma of the other category. All five categories are kept to increase ease of dilemma identification within specific genres. From these categories (as given in equations 1 to 5) dilemma instances can be found and generalised within each domain. From the generalised form of the dilemma the system will be able to create new dilemmas. In the presentation of these to the user wholly original stories are created.

It will not be possible to create great literature in this way – the use of clichéd storylines prevents this. However, such stories are enjoyed by many people and this method is common in such genres as James Bond films, soap operas (soaps) and "chick flicks". The story is built around the cliché, and it is the cliché as well as the story which the audience appreciate, the very repetitiveness and familiarity of the dilemmas adding to the dramatic interest. It can only be imagined how much more enjoyment could arise from the user becoming a character in such domains, and experiencing the dilemmas first hand.

6 THE NARRATIVE GENERATOR

Prior to a dilemma being presented to the user certain conditions must be met within the storyworld. These are the preconditions of the dilemma. It is the task of the storywriting system to achieve these preconditions. This constitutes the build-up – the essence of the story itself. Given actions within the storyworld the system can use planning to satisfy a dilemma's preconditions. In this way a plan to achieve a dilemma becomes a storyline. The interactive drama is made up of a series of such substories, dynamically selected according to dramatic interest.

The system uses a modified GraphPlan planner [3] which utilises a STRIPS-style representation of actions. On being passed a dilemma, the planner finds all plans to achieve this dilemma given the current storyworld state and background knowledge. From these plans, that which is most dramatically interesting can be selected and execution attempted. If the plan is successful the corresponding dilemma is presented to the user. Once the user has made their choice, the system updates the storyworld state in accordance with that choice. The system can then plan from the new state in order to attempt presentation of another dilemma to the user – thus continuing the interactive drama. This sequence of events is demonstrated in fig. 2. From this figure it can be seen that the planner finds all plans in the story dependent on the current state and a given dilemma. If no plan can be found for this dilemma, another is selected. Once all plans have been found, the most dramatically interesting can be followed (providing the user cooperates), resulting in a new state from which the story continues.

The potential consequences of each decision must be clear to the user before they make their choice. Once they have chosen, these repercussions on the storyworld are implemented. The resultant state is thus entirely dependent on the user's decision.

The sequence in which the dilemmas are selected for planning is dependent on the story history, the frequency of dilemma use, dramatic interest and the user model. Dilemmas must depend on what has happened previously and thus become part of a consistent story. Certain dilemmas only occur occasionally, others are more frequent. This will need to be determined for each domain, and considered when selecting each dilemma. It is necessary to plan for dilemmas which have a high dramatic interest, although this largely depends on the current user. The user model is discussed in section 10.

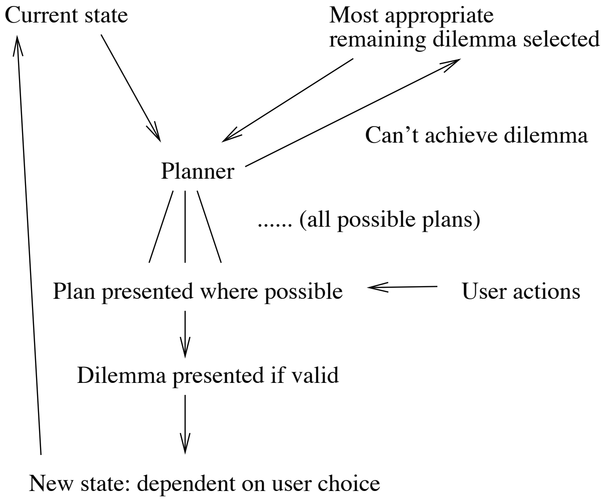

Current state
Most appropriate remaining dilemma selected
Can't achieve dilemma
Planner
...... (all possible plans)
Plan presented where possible
User actions
Dilemma presented if valid
New state: dependent on user choice

Figure 2. This figure gives an overview of the system moving between states dependent on plans, dilemmas and user decisions.

7 USER ACTIONS

This section discusses the methods used to integrate user actions into plans. The user should not have actions imposed on them as this would be very frustrating and unsatisfactory. The system thus allows the user to act as they desire within the storyworld. It must be ensured that the user is as free as possible while still experiencing dilemmas. The substory generated is essentially the same as that with no interaction, although its presentation may not succeed and thus replanning be required even after an appropriate plan has been found.

The planner initially assumes that the user will act in a manner consistent with the way characters with similar traits would act in soaps. Ideally a user model would give a more accurate idea of how the user will act. Once a plan has been chosen it is then presented to the user only inasmuch as is possible before it is necessary for the user to act for the plan to continue. This is the case when a precondition of an action or the dilemma requires a user action to be satisfied. If the user acts in a manner which satisfies the necessary preconditions at this stage then the presentation of the plan continues until a user action is required again. As soon as it becomes possible to present the dilemma this is done.

In its current version the system is control-based. This means that the user selects actions until they choose to pass control back to the system, which then acts until a user action is required to satisfy required preconditions. When the user has control they can take any number of actions. To ensure that the user does not feel constrained every act that other characters within the system can make is available to the user. The user can spend as long as they want considering their options.

The user inputs their action choices as two or three typed words which summarise the action they have chosen, for example 'move club' signifies that the user wishes to move from their current location to the club. The system is capable of recognising a range of possibilities for each action. Additional options available to the user include being able to see the current state of the storyworld and information on other characters.

It is clear that the user will not always act as required by the plan. Any act which satisfies the preconditions of the next stage of the plan is acceptable, but even then the user has a wide range of options. There are various methods which can be used to overcome the problem, for example:

1. Multiple valid plans are maintained. The system only acts in accordance with those which the user is following. As soon as a plan becomes impossible, it will be removed from consideration.
2. An assumption model for the way which the user is likely to act should be created. This is discussed in section 14.
3. In some cases it is possible to adapt the plan to suit user actions, such as by changing the names of characters involved in a plan. Due to the possible actions being so strongly dependent on individual character traits this proved not to be a particularly useful method.
4. Plans with a minimal number of user actions can be chosen. This is not a favourable method as it tends to reduce the user's interaction with the story.
5. Shorter plans are favoured. This means that there are less opportunities for the user to act outside the plan, while still creating plans in which their actions will have an effect. Stories of the same length will involve more drama if plotlines are shorter.
6. The user is cooerced into acting in the way required by the current plan. For example, if it is required that the user moves from location l to location k their friend can go to location l and ask the user to join them in going to location k.

Of these methods 1, 5 and 6 are implemented to good effect in the current system. Methods 3 and 4 were decided not to be of benefit, for the reasons discussed.

As the user may require time to consider their actions, threads are utilised so that planning can take place while the user thinks. Potential plans are added to a list, and the system attempts to integrate the user's actions with the most appropriate valid plan.

8 CHARACTER RESPONSES

If characters other than the user only act within the build-up to a dilemma the experience can become frustrating for the user as they would not see any response to their actions unless they only act within plans. It is also unrealistic, as there are actions which take place in genres involving clichéd storylines which do not have any direct relevance to a specific dilemma. Such actions should be incorporated in the stories produced.

Identifying patterns in large numbers of user actions is complex and requiring this would reduce the extendibility of the system. Therefore a system based on tit for tat reactions and utility scores was designed. In each story state a numerical utility value is assigned to each character. Actions change this value due to the corresponding change to the affected character's score. When the user acts in a way which decreases the score of another character, that character responds by acting to decrease the user's score by the same amount. The use of utility values makes extension to additional actions much more practical, as it requires only the association of a value with each. This method also makes system responses less predictable and more versatile.

An example would occur when a character is fancied by the user, and thus has an associated positive score in that state. If the user stops fancying this character then the character's score is resultantly decreased. In this case it could be that the character responds by ceasing fancying of the user (if this is possible). There are many other action options available to the character, some which are less obvious and possibly more 'revenge' or 'reward' based. These are always consistent with action possibilities for the current genre. In this example, the character might feel rejected and thus encourage (or bully) the user to betray their principle and to steal.

Such responses to user actions take place when the system has not yet presented a dilemma. The system should respond to all user actions since the last dilemma was presented. This is because dilemmas form turning points in the story and are likely to change the direction as dilemma implications cause drastic changes to feelings between characters (including the user). This means that a response to all preceeding acts could well be unrealistic and outdated. If the user's actions have not changed the utility scores of any other characters then there is either no response or a response which is deemed to be the most appropriate, dependent on the user's actions and how they have affected the user.

It is possible for each character to respond to the user's actions towards them with up to two actions. If this was extended the relevance would be reduced as the story would move too far from the focus of a dilemma, thus reducing the dramatic interest of the experience. It is also possible that too lengthy a response would result in the user feeling less involved.

Actions which can occur in response to user actions are not always appropriate as part of a plan. For example it would not be appropriate to have a plan involving a character ceasing liking another character without prior actions to justify this. There are thus certain actions which can only occur in response to user actions. These are used in combination with the basic actions in order to determine appropriate responses to user actions. It is necessary to maintain a focus on responses which don't diverge too far from achievement of dilemmas.

Those actions which are a response to the user's actions will be in accordance with what they require and expect from the story. The responses update the state and thus effect the future path of the story - both immediately and in the longer term. These are not unrelated actions but should become an integral part of the story while serving also to increase the effect of the user's actions.

The interest of the story is increased through use of utility-based responses as the stories and order of dilemma presentation has less predictability when actions are not always in line with the plan. These responses increase the specificity of the story to a particular participant or user. They are likely to encourage the user to act more, as they see an immediate effect of their actions. This may also increase the believability of the characters.

9 CHARACTER DILEMMAS

If characters are not themselves faced with dilemmas they suffer from a lack of depth and interest. This is because they do not participate in the narrative except inasmuch as their actions affect or directly respond to the user. The system therefore allows characters other than the user to be faced with and make decisions on dilemmas.

All of the dilemmas are possible for any characters within the storyworld (given applicability and satisfaction of preconditions) so planning takes place as before. The only difference is that a non-user character is now the deciding participant. As long as the user is not involved in the plan, it is presented as a sequence of actions prior to a dilemma – of which the decision and outcome are shown to the user.

When the plan for a dilemma requires user involvement the issues involved in incorporating their actions into a plan resurface. These are not always negative, as here the user is able to act in a way which could lead another character to a dilemma. This increases the user's involvement as they are able to attempt to manipulate others thus extending the complexity of the world.

The outcomes of dilemmas affecting the user have been adapted. If, for example, the dilemma presented to the user would result in a character choosing to run away with the user then where necessary this now involves the character asking the user to run away with them. This means that the user feels less controlled, although with a developed user model their response should always be predictable.

Once the next most appropriate dilemma type has been identified the system tries to present an instantiation to the user. If this fails an attempt is made to present this dilemma to another character, unless a large number of such dilemmas have just been presented in succession. Planning for character dilemmas takes place as the user thinks, in another thread which continuously updates a list of possible character dilemmas and corresponding plans.

The linearity of the storyline is removed by allowing other characters to experience dilemmas. The user sees that there is more happening in the world as they think and act. In some genres a linear focus on a single character may be more appropriate, as in James Bond films. The proportion of non-user dilemmas can thus be adjusted by the story designer dependent on the genre.

The system is able to create a non-interactive story. This means that there is always a story whether or not the user chooses to act within the storyworld. This creates the illusion that these characters exist outside the user's scope and thus increases their believability. It also gives the user the option of not acting in the world should they choose not to, whether for a long or brief period of time. They experience a story which at any time they have the option to become an active participant in.

When considering the frequency of dilemma use, care is taken to ensure that the user experiences a reasonable proportion and balance of dilemmas while the overall frequency is as would be expected for the genre. The interestingness is also taken into account, although here it is less important as the user is not being presented with the dilemma.

10 THE USER MODEL

The user of an interactive drama system should be modelled rather than controlled. The story should adapt to the user's interactions rather than forcing the user to follow a particular storyline.

The user model is used to identify which dilemmas are going to be most conflicting and dramatically interesting for the current user. There is an "interestingness" value associated with each dilemma. This value is initally fixed in accordance with the values found by a survey of diverse soap viewers. The value will adapt to suit the user and their modelled personality. The system searches for the most interesting story path to a pre-defined fixed depth (dependent on the size of the search space and the speed of the search and planning algorithms).

Each dilemma has associated assumptions as to how the modelled values change dependent on the user decision. Once they have made their choice, the user model is updated accordingly. A selection probability is associated with each criterion, so that the credibility given to the user model depends on how many times it has been updated. It additionally depends on how recently the criterion being utilised was updated – since the user and their opinions are likely to change through the course of the interactive drama. This user model is then be employed to approximate the probability of a user making a particular choice within a dilemma. It then calculates the expected total "interestingness" of that path. The system selects that dilemma which has the highest chance of leading to the most dramatically interesting experience for the user. A section of this search is shown graphically in fig. 3.

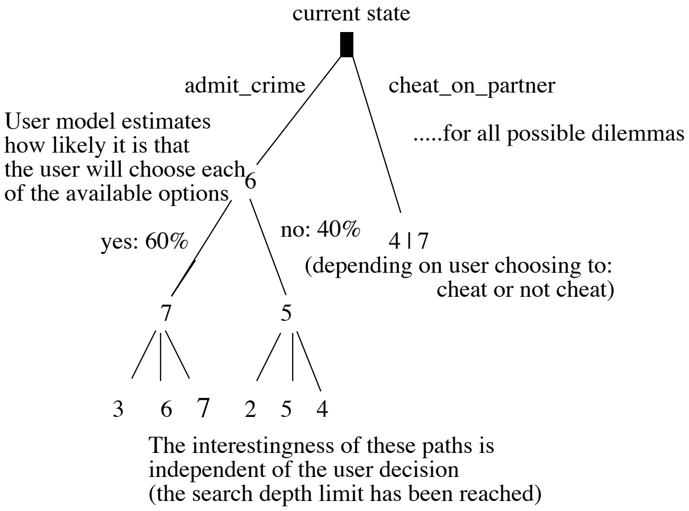

current state
admit_crime
cheat_on_partner
User model estimates
how likely it is that
the user will choose each
of the available options
6
.....for all possible dilemmas
yes: 60%
no: 40%
4 | 7
(depending on user choosing to:
cheat or not cheat)
7
5
3
6
7
2
5
4
The interestingness of these paths is
independent of the user decision
(the search depth limit has been reached)

Figure 3. This figure shows a section of a potential user model. The expected interestingness (for the user) of each dilemma is given at each node, assuming that the highest score will be achieved from the final nodes. The prospected score of the admit_crime dilemma is 12.2 (6 for the admit crime dilemma summed with the expected maximum score for the following dilemma, i.e. 6 + 0.6(max(3,6,7)) + 0.4(max(2,5,4))). A similar calculation can be carried out for each path, and the most interesting subsequently selected.

In this story creation method, care must be taken to ensure that a single dilemma (or group of dilemmas) is not overused. In order to do so, the frequency of occurence for each dilemma (within the specified domain) must be considered.

11 EXAMPLE DOMAIN

The techniques discussed here are applicable in any genre which places a particular emphasis on stereotypes and clichés. It was decided to intially focus on the creation of an interactive soap. This domain does not require an overall story arc but rather involves an infinite series of 'mini-stories'.

The domain of soap operas is commonly understood to revolve around stereotypical storylines. In many cases, these involve a character being presented with a decision likely to result in negative outcomes either way. A range of such dilemmas which characters have faced in recent years from Neighbours, Home and Away, Coronation Street, Eastenders and Hollyoaks have been identified and generalised[2]. These soaps were selected for their accessibility, familiarity and popularity with the general public.

It was found that the soap dilemmas fell into only three of the five possible categories, namely Betrayal (1), Sacrifice (2) and Favour

[2] Thanks to George Barber for his knowledge of soaps and ability to identify such dilemmas.

(5). Figure 4 gives examples of these dilemmas, one of which is generalised in fig. 5.

Hollyoaks: Becca has the opportunity to cheat on her husband Jake with Justin, a schoolboy in her class.
Eastenders: Jane has to decide whether or not to cheat on her husband Ian with the local bad boy Grant.
Coronation Street: Danny has the opportunity to cheat on his wife with Leanne, his son's girlfriend.
Home and Away: Kim has to decide whether or not to cheat on his girlfriend with his best friend Hayley.
Neighbours: Stu has the opportunity to cheat on his institutionalised wife Cindy with a local pretty girl – who previously went out with his brother.

Figure 4. As can be seen from this small sample of similar dilemmas, the plotline of a character being presented with a dilemma involving cheating on their partner has been used in all of the examined soaps. This demonstrates the frequent use of clichéd storylines in soaps.

```
A_X: cheat_on_partner(character(X))
preconditions: partners(X,Y) ∧ loves(X,Z)
              ∧ loves(Z,X)
dilemma: ''Would you like to cheat on your
partner
      character Y with character Z who loves
you?''
if user chooses to cheat:
  add to state: cheating(X,Y,Z)
  update user model:
    honesty - lowered, faithfulness -
lowered,
    value_for_relationship with Y - lowered
if user chooses not to cheat:
  delete from state: loves(X,Z)
  update user model:
    honesty - raised, faithfulness - raised,
    value_for_relationship with Y - raised
```

Figure 5. A dilemma of type Betrayal which is frequently used in soaps (see fig. 1), and can be presented to the user of this interactive drama system.

All domain specific background knowledge was added to the system, including STRIPS-style actions (such as why two characters fall in love) and locations (for example club and house) which appear in the considered soaps. In fig. 6 an action from the system is shown with its pre- and postconditions.

```
Action: X starts to fancy Y
Preconds: fancies(Y,X) ∧ attractiveness X > 1
          ∧ attractiveness Y = 1
Effects: fancies(X,Y)
```

Figure 6. An action in the STRIPS representation in which any characters in the system can participate. Here, an attractive person fancies someone less attractive. In a soap world (where looks are very important) the less attractive character will begin to reciprocally fancy the more attractive.

Figure 7 shows a dilemma to a character other than the user and fig. 8 shows the user interacting with a plan and being presented with a dilemma. For clarity a single dilemma type is used throughout these examples, namely that which involves a character having to choose between potential partners. As can be seen, when the user is involved, they are free to choose their own actions, although they will be encouraged to participate in the plan as shown here. Figure 9 shows a character responding to the user in a manner unrelated to a specific plan.

```
Action is adam moves between park and club
Action is jill gets drunk
Action is adam gets drunk
adam starts to mutually fancy jill

adam has to choose whether to partner joe or
jill, where adam fancies both and the feeling
is mutual.
adam decides to go out with jill, they are
now partners, and joe no longer fancies adam
```

Figure 7. This figure shows the build-up to and presentation of a dilemma in which the user does not participate.

```
Action is john moves between shed and house
Your friend john has come to the house to ask
you to go to the club with them, would you
like to go?
y
You move between house and club
Action is john moves between house and club
Action is joe gets drunk
n
joe offers to buy you a drink. Will you
accept?
y
You accept the drink from joe and get drunk
Action is joe starts to fancy you
fancy joe
You start to fancy joe
Who would you like to partner: adam or joe?
Given that you fancy both and they both fancy
you.
adam
You have chosen adam, you and adam are now
partners.
As a result of your choice, joe fancies you
less.
```

Figure 8. This example shows the user participating in a dilemma plan and then being presented with this dilemma. Where necessary they are encouraged by other characters to participate in the current substory. User input is shown in italics. In the preceeding state the user already mutually fancies Adam.

12 EVALUATION

A sample of 8 people were asked to test the example domain discussed in section 11. Of these 4 were experienced game players

```
flirt adam
You flirt with adam
n
bert asks you to go to the club where they
will buy you a drink. Would you like to take
up this offer?
n
Action is adam flirts with you
```

Figure 9. The user's action decisions here (shown in italics) result in the system failing to present the current dilemma. The utility-based response of flirting with the user is thus created.

(group A), the remainder were not (group B). The users played for an average of 7 minutes. Although the game world was very limited at the time of testing, resulting in a lack of breadth in the stories, this serves to demonstrate the usability and potential of these techniques.

The users in group A found the story to have a reasonable level of interest, rating this and their enjoyment with an average score of 3/5. There was a strong belief that their actions were having an effect on the storyworld. None of these users believed in the storyworld but all felt that they would replay.

It was found that the users in group B struggled with the system. They felt a need for graphical depictions of other characters and their available options. In general, this group felt that the story had low interest and believability and only one enjoyed the experience. However they all felt that their actions were having some effect and all but one would almost certainly replay.

13 RELATED WORK

Other interactive drama systems in existence use planning techniques. Mimesis [15] uses planning to achieve the story goals. This is much longer-term planning and is less flexible around the user's interactions - which are either accommodated in re-planning or intervened with. In the I-Storytelling [4] system, hierarchal task network (HTN) planning is used. Each character is equipped with an HTN to follow in the story, which is defined before the story begins. There is very little allowance for user interactions in this system. In neither system is there any allowance for the story to be dynamically created, but only for it to be dynamically adjusted.

More recent systems use planning techniques to create stories in collaboration with a user. In [14] the planner is used to create each stage of a planning graph. The user is then able to choose from the subsequent options to decide which will appear in the final version of the story. The story presentation will be a mimesis-style experience. Points for re-planning and intervention by the system are specified by the user at the story creation stage, whereever a need is identified by the system. The shortcomings of Mimesis apply here also. The system described in [9] involves goal events which are planned for. The user is able to specify some of these events and to prompt re-planning for any. They may be ignored. The user must then select the final ordering of events - given any constraints. The resulting story is then graphically produced without any interaction, and at a much lower level than that at which the user aided in the story creation.

Fairclough's system [6] utilises planning techniques to dynamically create an interactive story in the fairy tale genre. There are a finite number of subplots and the user's actions determine which is experienced. A plan is then created for the subplot, which consists of a "sequence of character actions" given to the NPCs as goals. The user has a high level of freedom but they are not entirely flexible as they must adhere to a limited number of subplots. In contrast, the system proposed here will allow the user complete freedom. The user is also modelled so that the experience is more enjoyable for them personally. The dilemmas posed to the user in our system will increase the dramatic interest of the stories.

Other systems utilise a user model. In IDA [10] this is used only to direct the user within the story's pre-defined overall plot structure. IDtension [13] uses the user model to determine the user's nature and present dilemmas accordingly. In this system, the user takes turns with the system to choose actions for the story as a whole. If they are modelled to consistently choose actions which avoid violence, the system can present them with a dilemma in which they must choose a violent action in order to achieve the pre-defined goals of the story. The dilemmas here are for the user as an external observer of the system, rather than as a character.

14 CONCLUSIONS AND FUTURE WORK

In this paper we presented an interactive narrative generator that is able to create long, and potentially infinite, story lines that incorporate dilemmas to add dramatic tension. The stories are dynamically created based on user decisions and actions as well as adating to the user's tendencies.

In future work an assumption model will be created based on previous user actions, which will be used by the planner. This will involve an applicability check creating a set of user-specific actions and making these available to the planner. As a result the user should be more able to act naturally and still be presented with dilemmas. As the user model becomes more accurate through the story there will be less need for other methods.

It is intended to extend the preliminary evaluation of the system. This will involve incorporating more participants who will play for longer in an extended version of the system. The results will be statistically analysed.

The possible extension of utility-based responses to use as dilemma implications will be investigated. This would cause actions rather than just character relationship and emotion changes as a result of dilemma decisions. The stories could thus become more interesting. This is not a simple task as determining the exact score changes and maintaining relevance becomes much more difficult.

In the current system all character actions and dilemmas are shown to the user. This has the potential to adversely affect the story interest and change the manner in which the user acts. For example, if a murder is committed and the user sees all acts they will know who the murderer was and the mystery will be destroyed. This removes a wealth of story potential. It would thus be advantageous to decide when information will be presented to the user, eventually revealing everything which is relevant to explain later characters acts and dilemmas. This could also add to the realism as the characters in a story do not always see what happens to other characters. However as viewers usually will it is important to maintain a balance in this.

It may be advantageous to have a less turn-based interface, where the system and user can interrupt one another when acting. It is ultimately intended that these interactive drama worlds will be graphically simulated. In this way the user will see the storyworld as in conventional media but will be a character, and will be able to act as such. In the short term pictoral representations may be possible.

There is additionally the potential for the creation of soap-specific dramas, with characters as in real soaps, for example an interactive *Eastenders* soap.

REFERENCES

[1] Heather Barber and Daniel Kudenko, ‘Adaptive generation of dilemma-based interactive narratives’, in *SAB’06 Workshop on Adaptive Approaches for Optimizing Player Satisfaction in Computer and Physical Games*, Rome, Italy, (2006).

[2] Joseph Bates. The oz project. http://www.cs.cmu.edu/afs/cs/project/oz/web/oz.html, 2002. Bates led the CMU Oz Project.

[3] Avrim Blum, Merrick Furst, and John Langford. Graphplan. http://www.cs.cmu.edu/∼avrim/graphplan.html, 1997.

[4] Marc Cavazza and Fred Charles. Interactive storytelling. http://www-scm.tees.ac.uk/users/f.charles, 2006.

[5] Chris Crawford. Erasmatron. http://www.erasmatazz.com, 2006.

[6] Chris Fairclough, *Story Games and the OPIATE System*, Ph.D. dissertation, University of Dublin - Trinity College, 2004.

[7] Syd Field, *The Screen-writer’s Workbook*, Dell Publishing, New York, 1984.

[8] Barbara Hayes-Roth. The virtual theater project. http://www.ksl.stanford.edu/projects/cait/, 2001. Hayes-Roth led this research group.

[9] Börje Karlsson, Angelo E. M. Ciarlini, Bruno Feijó, and Antonio L. Furtado, ‘Applying a plan-recognition/plan-generation paradigm to interactive storytelling’, in *Workshop on AI Planning for Computer Games and Synthetic Characters*, The Lake District, UK, (2006).

[10] Brian Magerko. Interactive drama architecture. http://www.eecs.umich.edu/ magerko/research, 2006.

[11] Michael Mateas and Andrew Stern. Façade. http://interactivestory.net, 2005.

[12] Nikitas Sgouros. The defacto project. http://www.cslab.ece.ntua.gr/∼defacto/default.htm, 1997. Sgouros led this research group.

[13] Nicolas Szilas. Idtension project. http://www.idtension.com, 2006.

[14] James M. Thomas and R. Michael Young, ‘Author in the loop: Using mixed-initiative planning to improve interactive narrative’, in *Workshop on AI Planning for Computer Games and Synthetic Characters*, The Lake District, UK, (2006).

[15] R. Michael Young and C. J. Saretto. Liquid narrative. http://liquidnarrative.csc.ncsu.edu/people.php, 2001. Mimesis.

From the Event Log of a Social Simulation to Narrative Discourse: Content Planning in Story Generation

Carlos León and **Samer Hassan** and **Pablo Gervás** [1]

Abstract.

This paper presents a proposal for implementing automated story telling of narrative threads within a multiplayer game based on selection and linearization of game logs. Our initial prototype operates on logs generated artificially by a social simulation built by a multiagent system. This provides a log of events for a large set of characters emulating real life behaviour over a certain period of time, with no need to carry out a real game involving several players over an equivalent time. The proposed method addresses tasks of content determination - filtering the non-relevant events out of the total log -, and discourse planning - organizing a possibly large set of parallel threads of events into a linear narrative discourse. Actual sentence planning and realization is not addressed, but rather performed in a crude manner to allow readable presentation of the generated material. Examples of system input and output are presented, and their relative merits are discussed. The final section discusses futures lines of work that may be worth exploring.

1 Introduction

Narrative games used for educational purposes have a great potential for improving the learning experience for students, both in terms of making it more interactive and by providing a strong entertainment component that might act as additional motivation. Part of this potential lies in the fact that there is a story underlying the game. This story is in most cases only implicit, in the sense that it arises as the game goes on. This is what makes it interactive, and it presents advantages from the point of view of entertainment. However, from a pedagogical standpoint, having access to an explicit version of the same story may provide additional advantages. On one hand, it may provide the student with a textual summary of how a particular game or gaming session developed. This may be of use when revising material that has already been covered, or in trying to understand what went wrong. The ability to revise is an important ingredient of the learning experience. If games are to take the role currently played by lectures or laboratory sessions, the explicit narratives of such games might play the part of the notes usually taken by students - as game players are unlikely to take notes as they play. On the other hand, explicit narratives reviewing particular sections of a game may constitute a useful tool in developing functionalities for assisting student/players in succesfully completing the game, maybe by explaining how a particular situation in which they find themselves has come about. It is common for current games to have a set level of difficulty, so that part of their entertainment value lies in the challenge of reaching the level of profficiency required. Players setting off to achieve it from a low level of proficiency may have a hard time at the initial stages, up to the point where many give up before achieving the goal. Providing the system with help facilities based on inserting small narratives explaining particular details required for solving puzzles may be seen as detracting from the challenge the game presents as means of entertainment, but they can be a positive addition from the pedagogical point of view if they ensure that more of the students setting out to solve the game actually reach the final goals. To make the point clear, an example is presented for a particular type of game. Some modern games, like MMORPGs[2], are played by several players over huge maps with many locations and many characters. These games usually have different agents interacting between them, and creating more or less complex relations that could be important for the global story of the gameplay. Non-player characters with coherent storylines, set in motion by the casual presence of one player, may meet other players at a later point. In order to understand their behaviour, this second player may need to know their story. This information is actually available in game logs, and it can be read by game masters, which can then write this data in a human readable form. If the system is to manage this task in an autonomous manner, capabilities for automated story telling must be provided. This paper presents a proposal for implementing such functionality: this text in natural language explaining the most interesting parts of the game can be generated by machines resorting to state of the art natural language generation technologies. The actual sequence of events that have happened is available, stored as a system log or in short-term memory. But telling it in an entertaining way, while at the same time filling in the gaps in the players knowledge of what has happened, is not a trivial task. Research in automated telling of stories attempts to fill this gap. The tasks involved will cover the basic requirements for identifying the most relevant material among a large search space of recorded events, converting a sequence of such events - or various parallel sequences of them - into a story, and presenting this selection to the user, already organised into narrative threads.

In order to avoid the task of collecting real data from massive multiplayer online games, we have based our initial prototype on a social simulation generated by a multiagent system. This provides a log of events for a large set of characters emulating real life behaviour over a certain period of time. The simulation we have used was initially developed for a different purpose in the field of experimental social sciences, and it has been adapted to its current purpose by customising the domain characteristics and the set of possible operations available to the agents to simulate a game-like environment.

We want to simulate a game system with many agents or game characters where the main key is the interaction between them, and

[1] Department of Software Engineering and Artificial Intelligence. Universidad Complutense de Madrid. cleon@fis.ucm.es, samer@fdi.ucm.es, pgervas@sip.ucm.es

[2] Massive Multiplayer Online Role Playing Games

the result is the emergent behaviour as a social group. This behaviour is a full story along a defined period of time, with interesting episodes, boring ones, communities trying to survive, and individual characters doing incredible things. We propose a multi-agent system with social capabilities, emulating a real fantasy medieval game. We have developed a multi-agent system that simulates a community of non-player characters being born, living and dying, where each agent or character saves its history. When all these histories are generated as data logs, we process them to build a structure where only the important facts are told, and that can be easily translated to natural language, freeing in this way the game masters or system administrators from writing this text themselves.

2 Previous Work

In order to develop this system, we have resorted to previous work in the fields of natural language generation and social simulations using multi-agent systems. A brief outline of the relevant studies is given in this section.

2.1 Automatic story generation

The general process of text generation takes place in several stages, during which the conceptual input is progressively refined by adding information that will shape the final text [9]. During the initial stages the concepts and messages that will appear in the final content are decided (*content determination*) and these messages are organised into a specific order and structure (*discourse planning*), and particular ways of describing each concept where it appears in the discourse plan are selected (*referring expression generation*). This results in a version of the discourse plan where the contents, the structure of the discourse, and the level of detail of each concept are already fixed. The *lexicalization* stage that follows decides which specific words and phrases should be chosen to express the domain concepts and relations which appear in the messages. A final stage of *surface realization* assembles all the relevant pieces into linguistically and typographically correct text. These tasks can be grouped into three separate sets: *content planning*, involving the first two, *sentence planning*, involving the second two, and *surface realization*.

The work presented in this paper is related to the first two tasks: content determination and discourse planning. Content determination is known to be always heavily dependent on the particular domain of operation, and tightly coupled with the particular kind of input being processed. Little generalization is possible for this task. Discourse planning determines the ordering and rhetorical relations of the logical messages, hereafter called facts, that the generated document is intended to convey. Most existing approaches to discourse planning are based on either rhetorical structure theory (RST) [5, 4] or schemata [6].

2.2 Social systems

Social phenomena are extremely complicated and unpredictable, since they involve complex interaction and mutual interdependence networks. Sociologic explanations deal with large complex models, with so many dynamic factors involved, they are not subject to laws, but to trends, which can affect individuals in a probabilistic way.

A social system consists of a collection of individuals that interact among them, evolving autonomously and motivated by their own beliefs and personal goals, and the circumstances of their social environment. Due to the mentioned complexity, techniques are required that consider how global behaviour can be derived from the real subjects' behaviours, which are fundamental in any social system. In particular, there is an interest in observing the emergent behaviour that results from the interactions of individuals as a way to discover and analyse the construction and evolution of social patterns.

A multi-agent system (MAS) consists of a set of autonomous software entities (the agents) that interact among them and with their environment. Autonomy means that agents are active entities that can take their own decisions. The agent paradigm assimilates quite well to the individual in a social system. In fact, there are numerous works in agent theory on organisational issues of MAS. Also, theories from the field of Psychology have been incorporated to design agent behaviour, being the most extended the Believes-Desires-Intentions (BDI) model, in the work of [2].

With this perspective, agent-based simulation tools have been developed in recent years to explore the complexity of social dynamics. In this way agents' reactions can be monitored in an observable environment, defining the lines of system evolution. This provides a platform for empirical studies of social systems. And because of that, the specification of characteristics and behaviour of each agent is critical, so it can manage the dimensions of the studied problem. A screenshot of one of these tools is shown in Figure 1.

In the MAS designed, as explained in [7], the agents have been developed with several main attributes: from simple ones such as sex or age, to complex ones, like for example ideology or educational level. The population in the agents' society (as in real societies) also experiments demographic changes: individuals are subject to some life-cycle patterns: they get married, reproduce and die, going through several stages where they follow some intentional and behavioural patterns.

Moreover, the agents/individuals can build and be part of relational groups with other agents: they can communicate with other close agents, leading to friendship relationships determined by the rate of similarity. Or, on the other hand, they can build family nuclei as children are born close to their parents.

Thanks to the underlying sociological model, the parameters of the social simulation system fit all together logically. In this way, the system may be configured to reflect the parameters (such as average number of children per couple, or mean of male average age of death) from a specific country or even import data from surveys that specify the attributes of the agents, reflecting the behaviour of the given population.

Besides, due to the relative simplicity of the agents, the system can manage hundreds of them, reaching the necessary amount for observing an emergent behaviour that results from the interactions of individuals, leading to the appearance of social patterns than can be studied. And for this study, during and after the execution of the simulation tool several graphs may be plotted that reflect the evolution of the main attributes of the social system.

3 Story Generation

Our approach to story generation is based on three tasks: *content determination*, *discourse planning* and *sentence planning*:

- In *content determination* we choose which data is going to be useful for the final narration. In this stage we suppress irrelevant facts present in the log, obtaining a version where redundant or useless data is removed. We can see this step as a "filter" of the log.
- *Discourse planning* consists on identifying a proper order of presentation of the previous data. We apply a particular technique (we

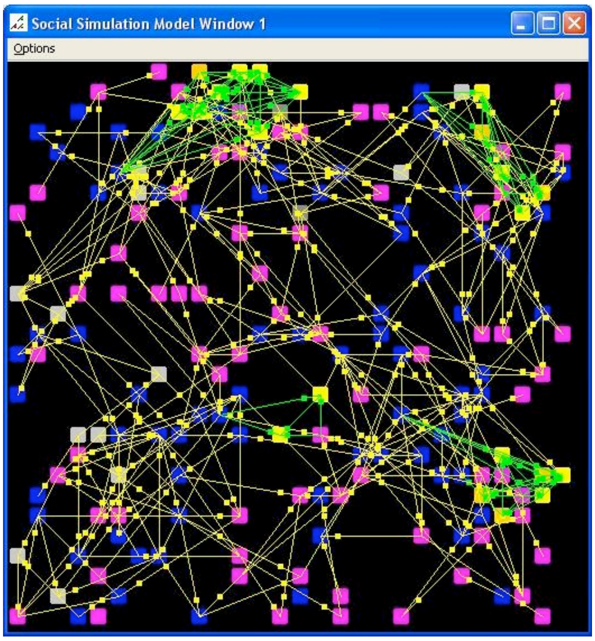


Figure 1. Screenshot of the social simulation

can use several algorithms, later this will be explained), and give the selected data generated in the *content determination* stage a particular order of narration, considered interesting for the readers of the final text.

- Then, we can perform *sentence planning*. This last step is the final process to be done, where the ordered log that represents a story in a structured form is translated to a natural language text.

It is not necessary to run these steps in sequential order. We have decided to join the two first steps into a single one; however, they could be done separated. Next we explain the solutions we have used for this work for each of these previous steps.

3.1 A Manual Story Generation Tool

Before creating a fully automatic system, we want to know which rules we, as humans, apply in story generation. That is the reason why we have created a tool for manual story generation, *Herodotus*. With this tool it is possible, with a simple few mouse clicks, to "draw" a full discourse from the facts and the logs recorded during an execution of a multi-character system.

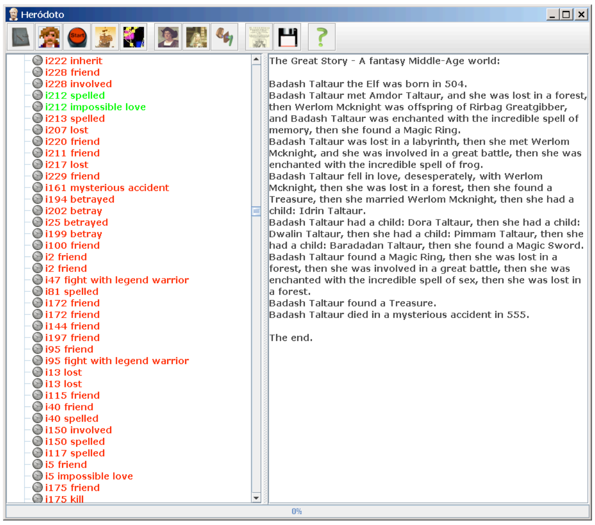
With *Herodotus* it is possible to perform *content determination*, excluding from the final story those facts that we consider to be boring or not relevant; *discourse planning*, creating the components needed to define a particular narration: relationships between facts (nexus between consecutive facts, like "while", "then" or "before that"), *discourse atoms*, or blocks of facts which are a semantic units (can be seen as paragraphs) and start and end points of the story; and simple *sentence planning*, with template-based solutions for transforming facts into text. This tool can also export a file in each step, in this way, for example, we could do *content determination* and *discourse planning*, export the result, and run a different program to generate natural language text, or an animated summarised reproduction of the gameplay.

To use *Herodotus* one only needs to load an XML file from the multi-agent system or from the log of a real game. Then, the full list of logs for each agent/player becomes visible in the main panel, with their facts, ordered by time. Once loaded, the log can be edited just by dragging with the mouse, drawing lines that represent relationships between the facts.

The facts can also be removed from the list, as well as the full logs, just by selecting them by clicking over them with the mouse, and pressing a button on the toolbar. Also, logs and facts can be added by hand, creating new threads of action and new characters.

Once we have connected the facts in order, and having removed those facts that are not important, it is only needed to group the events in blocks, that will be the *discourse atoms*, as we have explained before.

In Figure 2 we can see a screen capture of *Herodotus* working.



```
i222 inherit
i228 friend
i228 involved
i212 spelled
i212 impossible love
i213 spelled
i207 lost
i220 friend
i211 friend
i217 lost
i229 friend
i161 mysterious accident
i194 betrayed
i202 betray
i25 betrayed
i199 betray
i100 friend
i2 friend
i2 friend
i47 fight with legend warrior
i81 spelled
i172 friend
i172 friend
i144 friend
i197 friend
i95 friend
i95 fight with legend warrior
i13 lost
i13 lost
i115 friend
i40 friend
i40 spelled
i150 involved
i150 spelled
i117 spelled
i5 friend
i5 impossible love
i175 friend
i175 kill
```

The Great Story - A fantasy Middle-Age world:

Badash Taltaur the Elf was born in 504.
Badash Taltaur met Amdor Taltaur, and she was lost in a forest, then Werlom Mcknight was offspring of Rirbag Greatgibber, and Badash Taltaur was enchanted with the incredible spell of memory, then she found a Magic Ring.
Badash Taltaur was lost in a labyrinth, then she met Werlom Mcknight, and she was involved in a great battle, then she was enchanted with the incredible spell of frog.
Badash Taltaur fell in love, desesperately, with Werlom Mcknight, then she was lost in a forest, then she found a Treasure, then she married Werlom Mcknight, then she had a child: Idrin Taltaur.
Badash Taltaur had a child: Dora Taltaur, then she had a child: Dwalin Taltaur, then she had a child: Pimmam Taltaur, then she had a child: Baradadan Taltaur, then she found a Magic Sword.
Badash Taltaur found a Magic Ring, then she was lost in a forest, then she was involved in a great battle, then she was enchanted with the incredible spell of sex, then she was lost in a forest.
Badash Taltaur found a Treasure.
Badash Taltaur died in a mysterious accident in 555.

The end.

Figure 2. Screenshot of *Herodotus*

3.2 Adapting the MAS for Story Generation

The ideas expressed above concerning social simulations using multiagent systems are the core of action from which we have built the whole narrative system. Several changes to the original MAS have to be made in the perspective of execution to be able to generate full logs of action which will be the basis for the texts describing the storyline. It is necessary to shift the point of view from data acquisition to log generation. These logs must save the data in such a way that story generation can be carried out as easily as possible. We do not need numerical data, but semantic content that can be interpreted by the rules as we interpret them, because we want the story generation to be as close as possible to what humans might have done faced with similar sets of events.

We changed the meaning of the actions of the agents, not only by changing their names and the sets of them, as explained below in 3.3, but also by changing our interpretation of them, creating in this way a rather different world. For example, a value of "low" in economy has a particular meaning in the social simulation (a small house, no car), but in a Middle-Age time setting, a "low economy" means that the character is a peasant. Following this, the semantics we assign to each fact affect the significance of that fact in particular. A “low economy” character in the medieval setting does not have the same interest than a “low economy” character in a modern society.

3.3 Adapting the MAS to a New Environment

Several minor changes have been introduced in the designed MAS for its adaptation to a new environment: a Fantasy Medieval World far from the previous Post-Modern context. Thus, we have introduced Name and Last Name apart from the ID of each agent, together with the inheritance of the Last Name: this will be useful for telling the stories of lineages, and for personal events. We added a new attribute to each individual: the race, so they can be elves, humans, dwarfs... Thanks to the modular structure of the system it has not been a difficult task to achieve.

Other changes are related to the system structure. One problem was the involvement of non natural deaths, never considered in the old MAS. We added a random possibility of dying for each agent, allowing the possibility that we can relate this early death to the betrayal of a friend, poisoning by a wife, or even a mysterious accident.

The finishing touches arrived with the recording of the sequence of “life events” for every individual. But usual life events, like having friends, finding a couple, or the birth of children, are not interesting enough to build an exciting fantasy adventure. Because of that, we have included new types of events related to this context that will appear randomly. Thus, along his path, the agent can suffer several spells (loss of memory, fireball... or even change of sex!), kill horrible monsters (ogres, dragons), get lost in mazes or dark forests, find treasures and magic objects in dangerous dungeons... In this way we can build a really amazing (and sometimes weird) story, with several characters that evolve and interact among them.

At the end of simulation, this collection of events, together with the agents’ characteristics, is exported to an XML file. The XML-Schema pattern that rests beneath is not context-dependant, so the same format can be applied to other simulation environments. This file will be imported by a tool that will continue with the process of generating a story from the lives of some of these agents: the most interesting ones.

Here we present an explained example of the generated XML with the important information of each agent. In Figure 3 we can see the header of the file.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<Story Id="fantasy">
 <Description>
   A fantasy Middle-Age world
 </Description>
```

Figure 3. XML header of a log

Now the logs of every agent are listed: the initial ones, together with the next generations that appear during the simulation. Here we show just one of these logs: the one corresponding to the individual that will be selected as star of our story.

Each log is divided in two main sections. The first one (Figure 4) corresponds to the characteristics of the agent: each attribute of the character has two parameters, expressed as XML attributes: its ID (identifier of the attribute) and its Value. The value of these keys is, of course, context-dependant: here they represent aspects like its race or how religious the character is.

```
<Log Id="i212">
  <Description>
    Log of a character of the simulation.
  </Description>
  <Attribute Id="name"
      Value="badash"/>
  <Attribute Id="last_name"
      Value="taltaur"/>
  <Attribute Id="race"
      Value="elf"/>
  <Attribute Id="sex"
      Value="female"/>
  ...
  <Attribute Id="religion"
      Value="very religious"/>
```

Figure 4. Attributes of a character

The second main section (Figure 5) is the collection of life events, associated with the time in which they took place. As in the previous sections, XML attributes are context-free, but values of these attributes depend on the context. Thus, we can read in the full log (here we only show a small fragment), that in the year 515, the elf Badash Taltaur suffered a spell that transformed her into a frog. Or, analyzing the chain of events, we can see that the impossible love of her youth was, after she grew to be an adult, her formal couple, giving her many children and living happily... at least for some years.

```
  <Events>
   ...
   <Event Id="e9" Time="515"
      Action="spelled" Param="frog"/>
   <Event Id="e10" Time="515"
      Action="impossible love"
      Param="i229"/>
   ...
   <Event Id="e14" Time="526"
      Action="couple" Param="i229"/>
   <Event Id="e15" Time="526"
      Action="child" Param="i258"/>
   <Event Id="e16" Time="526"
  </Events>
</Log>
```

Figure 5. Events of a character

3.4 A Rule-Based Story Generation System

Given appropriate configuration parameters the social simulation generates a set of results which is sufficiently complex to constitute an interesting challenge for content planning. For this purpose, we have extended the manual story generation tool with an automatic story generation system. This program accepts a thread of facts from each agent of a defined set, and analyzes the connections and relations between these threads in time.

In our current design, we have chosen to perform an iteration through the elements of the log, using a rule-based system. Our first try was to implement the rule system in Jess [3], but, although it did work, the execution was extremely slow, and it required enormous amounts of memory. In contrast, writing rules in Jess is much easier than in Java, language in which we have developed the final version.

We have defined a set of domain-dependent rules for this problem in particular. We want to keep separated the general application of story generation (the editor *Herodotus*, structures for storing the stories, the natural language text generator, etc.) from ad-hoc content specialized for the specific system or a particular game. In this way the only work needed for adapting the application to other domains is restricted to defining the rules that establish which facts are important, and how they are going to appear in the final text, presentation or animation.

We have considered these rules to be expert knowledge. In the domain we are working on we cannot ignore the semantics present in the data saved in the logs during the gameplay for story generation. The meaning of particular attributes, not measured with numerical weights, must be taken into account before narrating a log: killing a red dragon is usually more interesting than killing a little spider. Of course, we can set some numerical values, as "kill-dragon interest", that should be a higher value than "kill-spider interest", but the final discourse will be made interesting with some "hand-made" rules, established by the system administrator, or perhaps the game-master.

3.4.1 Content Determination

As commented before, the first thing to do is to determine which data is not going to be told, and remove it. There are many possible solutions for this problem. The one we have used is to give a *factor of interest* to the characters. This interest factor is only a numerical value that represents how important it is for that character to appear in the story, not necessarily the comparative importance of that character with respect to other characters in the story. The value can represent real interest, coherence, fun, or any other reason why a given element from the logs should appear in the final text. In this way, a unimportant character can have a high *factor of interest*, because it is necessary that such character appears in the story. This factor is divided in two values:

- *Base interest* ($I_b(X)$) is the value we associate with the facts of some character X, and with their attributes. In this way, the character can be easily evaluated. With the attributes we can design a heuristic function h that represents the significance of some fact in the life of that character, given the attributes. It is usual for a man to fall in love, but not for an orc. That is why falling in love is more interesting in an orc's life than in a human's life. The actual method for computing $I_b(X)$ is shown in Formula 1 below,

$$I_b(X) = \sum_{i=1}^{n} f_i \cdot h\left(X, i\right) \tag{1}$$

  where f_i is the interest that we assign by hand for the fact i, x is the character, and $h(x, i)$ is the weight for the appearance of i in the life of x. The value of h is calculated with the type of i (what kind of fact it is) and with the attributes of x (if it is an elf, or an orc).
- *Relationship interest* ($I_r(X)$) is the level of importance of a character X calculated from the interest of their relationships with other characters: friends, foes, offspring, etc. We could not build a good factor of interest by considering only the characters as individuals, so we added this additional value. As before, the attributes of a character determine the final interest. We have a new function, g, depending on the relationship and the two characters, that represents the true interest of a relation: two elfs can easily be friends, but it is very strange (and perhaps we should tell about it) a friendship between an elf and an orc. The actual value is obtained using Formula 2 below,

$$I_r(X) = \sum_{i=1}^{n} I_b(Y) \cdot g\left(X, Y, i\right) \tag{2}$$

  where $I_b(Y)$ is the *base interest* of the character who has the relation i with X, and g is the heuristic function of the relationship i between different characters X and Y. The value of g is calculated with the type of i (what kind of fact it is) and with the attributes of x and y (if they are two orcs, or an orc and an elf, for example).

The final *factor of interest* is, in our current implementation, obtained according to Formula 3:

$$I_f = I_b + I_r \tag{3}$$

Once we have this value calculated, we have a new explicit data that will determine what is going to appear in the final structure. With the "interest" and some rules, like redundancy elimination (delete symmetric data: A is friend of B, and B is friend of A, then delete A or B), omission of irrelevant characters (those that are just born at some stage of the gameplay and then die at some later stage with little intervening activity), and of course, an importance filter (remove those characters whose *factor of interest* falls below a given threshold), we can have a set of facts and characters ready to form part of the final story. With these and other rules and filters, we can determine not only which characters are going to appear, but also which of their facts are going to be shown. The particular solution applied in generating the interest factor ensures that facts that are related to important characters are always included. This is intended to avoid the risk of eliminating non-interesting elements that may be of importance in a plot.

3.4.2 Discourse Planning

In discourse planning, basically we just reorder the facts in the story, and adapt the relationships between them. This is, in terms of computing, an easy task. But the goal of discourse planning is not only organizing the facts stored in the log, but inferring the guidelines of the story, giving them priority, and making them the main structure of the narration.

Several tasks must be accomplished in order to create a meaningful, clear and interesting story. In fact, we have found that these tasks are very dependent on the domain, and on what we want to present in the final story. While, as we have verified, adjusting the *factor of interest* to appropriate values is usually good enough for *content determination*, in discourse planning this is not true. It is very difficult to write general rules that generate different stories for different domains.

What we have done is to define ad-hoc rules for the domain we are working on, to process the particular data we have; and rules to generate the stories that we think that could be interesting for the reader. This rules are based on the three sets of data that we have: *facts content*, *attributes of the characters* and the *time*.

Some of this rules are, for example, to narrate the birth and death date of the main character only, to maintain a more or less time-ordered discourse, to talk about the unusual facts only, and so on. If we wanted to generate stories of fairy tales, for example, we could have omitted the dates, and we could have ordered the facts in a different way, trying to hide data that is only important in the end of the story.

It is important the way we manage time. In [1] we can see many ways of representing time, very related to this work. At this moment we consider that facts are instantaneous, ignoring intervals and time reasoning. We generate the time nexus between facts also with rules, and we have verified that, for simple narrations, this could be sufficient.

Once we processed the initial log, and having performed *content determination* and *discourse planning*, we can generate the final result. This result can be not only text, but also a script that controls an animation, a generated comic, or a summarised reproduction of the gameplay.

3.4.3 Sentence Planning

The final generation of the story is not only a nice way of showing the results. It can make the discourse interesting or boring, even if the order of the facts resulting from *discourse planning* is bad or good, respectively. Thus, we cannot ignore this step if we want to evaluate the generated content. It is not the same to say “Elrond was an Elf. He had a daughter called Arwen. Elrond was friend of Aragorn.”, as to say “Elrond the Elf, father of Arwen, was friend of Aragorn the King”. The final form of sentences not only gives beauty to the text, but may also convey information not actually present in the data structure. We can infer, in the second sentence, that Elrond is somebody important, as Arwen, and Aragorn is going to play a main role in the story. This knowledge is not contained in the first sentence. To achieve computational modeling of these characteristics is currently beyond the scope of this paper, but we intend to address it in future work.

The actual examples of output text presented in this paper have been generated with the use of a simple template-based surface realizer built on purpose for this particular application, and which produces monotonous text with little inflexion and no concern for literary style. This is because the main concern of the research reported here has been the succesful completion of the content determination and discourse planning tasks. For this purpose, such output texts are sufficient, and yet considerably easier for the reader to understand than the corresponding XML output files. The final result in terms of stories to be read by humans may be considerably improved by resorting to an existing sentence planning application. In future work, we intend to address this problem by integrating the present work with the PRINCE generator [8].

3.5 An Example

Now, we show a real example of our application. The multi-agent system is capable of running parametrized simulations, changing the number of characters, probabilities of the facts, years of simulation, and all other attributes of the system. Once executed, the system generates logs in XML, like the ones we have presented in 3.3.

At this stage, the story generation application reads the resulting XML file, and outputs a text. This example is the result of a simulation of the life of 200 initial characters and their descendants over a time span of 80 years. The system has inferred who is the most important character, and it produces the following rendition of her mortal life:

> *The Great Story - A fantasy Middle-Age world:*
> *Badash Taltaur the Elf was born in 504.*
> *Badash Taltaur met Amdor Taltaur, and she was lost in a forest, then she was enchanted with the incredible spell of memory, then she found a Magic Ring.*
> *Badash Taltaur was lost in a labyrinth, then she met Werlom Mcknight, and Werlom Mcknight was offspring of Rirbag Greatgibber, and Badash Taltaur was involved in a great battle, then she was enchanted with the incredible spell of frog.*
> *Badash Taltaur fell in love, desesperately, with Werlom Mcknight, then she was lost in a forest, then she found a Treasure, then she married Werlom Mcknight, then she had a child: Idrin Taltaur.*
> *Badash Taltaur had a child: Dora Taltaur, then she had a child: Dwalin Taltaur, then she had a child: Pimmam Taltaur, then she had a child: Baradadan Taltaur, then she found a Magic Sword.*
> *Badash Taltaur found a Magic Ring, then she was lost in a forest, then she was involved in a great battle, then she was enchanted with the incredible spell of sex, then she was lost in a forest.*
> *Badash Taltaur found a Treasure.*
> *Badash Taltaur died in a mysterious accident in 555.*
> *The end.*

4 Discussion

There are three main points worth discussing in an analysis of the proposed story generation solution: the possibility of evaluating results by comparing with human performance over similar tasks, the possible role of the sentence planning solution employed in the perceived quality of the output, and the particular choice of implementation that has been used.

4.1 Evaluation Against Human Performance

We are not evaluating if the story is interesting or funny, yet. We are only focusing on how similar are the machine generated stories with those stories that could be written by humans from the same source. We will keep on refining, in particular, the *content determination* process, because the output of this step is where we decide the interest of the elements of the story.

It would be interesting to compare the resulting work of the application of *content determination* and *discourse planning* in a log from a gameplay presented on this paper with a manual generation of the same log. In this way, we could see if the rules that we have applied in the code (filtering, ordering, connections between events) are those which would be applied by a human narrator. This task is, of course, possible, but the cost in time and human effort is very high. To perform the previous tasks by hand, over a log of 500 characters, could mean several days of work.

This prevents us, in principle, from evaluating how correct our application is, but it is an indicator of the utility of this work. This kind of story generation is very hard to do by humans, and can be easily done by machines. However, one possible evaluation of the system could be to ask a group a people to write a text describing a small set of facts of the log. This would provide an evaluation of the discourse planning stage of the system, but only partially address the evaluation of content determination - unless an evaluator chooses to omit a fact included in the selected set. In this way, we could compare human generated texts with machine generated ones.

4.2 The Effect of Bad Sentence Planning on Perceived Quality

Relative to the final output of the present work, it is obvious that the final example of generated text that we have presented does not have a nice form, and the narration is a little boring. The reason is that the *sentence planner* we are using is a skeleton implementation not even intended to be passably correct at its task.

This can be easily illustrated by a close analysis of the sentence planning tasks that are performed poorly in the given example, and considering how the text might have improved if those tasks were actually addressed in the implementation.

An important issue is how the sentence planner decides to represent the fact that a particular set of facts have been grouped by the discourse planner into a block of related events, to be narrated as a distinct thread within the discourse. In the current implementation this is simply solved by chunking all such facts into a single sentence, clumsily linked together with discourse markers indicating some kind of sequence. This can be seen in the example above in fragments such as:

> *Badash Taltaur met Amdor Taltaur, and she was lost in a forest, then she was enchanted with the incredible spell of memory, then she found a Magic Ring.*

This could easily be improved if, for instance, a simple sequence of sentences where used:

> *Badash Taltaur met Amdor Taltaur. She was lost in a forest. She was enchanted with the incredible spell of memory. She found a Magic Ring.*

However this obscures the fact that there are indeed chronological relations linking these particular facts with one another. A complex sentence planner would have to take this into account, and possible decide to give up the chronological information in favour of more fluid text.

Another related problem concerns sentence aggregation. The current sentence planner is incapable of detecting that a fragment such as:

> *...then she married Werlom Mcknight, then she had a child: Idrin Taltaur.*
> *Badash Taltaur had a child: Dora Taltaur, then she had a child: Dwalin Taltaur, then she had a child: Pimmam Taltaur, then she had a child: Baradadan Taltaur,*

might be considerably easier to read in a form like:

> *She married Werlom McKnight. They had five children: Idrin Taltaur, Dora Taltaur, Dwalin Taltaur, Pimmam Taltaur and Baradadan Taltaur.*

This transformation seems simple but involves at least an abstraction that is not trivial: the fact that a set of facts with the same predicate can be regrouped as a single predicate with a plural compound second argument.

This same example illustrates a different problem, that of referring expression generation. The sentence planner does indeed address this task in a clumsy manner, deciding at different places in the discourse to refer to a given character either by its full name or by a pronoun. This could be greatly improved, especially if it were considered in its interaction with elements such as additional sentence boundaries arising from a more refined realization of narrative threads. Additional issues related with this task arise from the fact that, if they are mentioned in close proximity, knowing the surname of the parents one may omit the surnames of all their children. This could lead to an even more refined version of the example above:

> *She married Werlom McKnight. They had five children: Idrin, Dora, Dwalin, Pimmam and Baradadan.*

4.3 Implementation Issues: Modularity vs. Efficiency

Relative to the implementation, it is also worth discussing the efficiency problems we have encountered using a declarative rule definition system like Jess. We first tried to build the whole rule system, and the evaluation of every fact present in the log, just using an implementation written in Jess. But it has problems of efficiency, because the algorithm behind Jess (the *Rete* algorithm), works in a way that is not optimal for our problem in particular.

We could have, then, implemented a hybrid system, and, while this is possible, the remaining content that could have been written in Jess was very reduced and easily translatable to Java. For that reason, we decided to stop using Jess, at least for this work.

As an example of rule, we present a definition of a simple filter that removes from the list of facts, those whose interest is equal to zero.

In Figure 6 we show the code as we implemented using Jess. The line "`(event (type ?type)(interest 0))`" means "that event of a defined type that has no interest". The other conditions in the rule are needed for the interface with Java (with the data structures). The resulting action of the rule is to remove, from the story, that fact.

```
(defrule remove-non-interesting
    (story (OBJECT ?story) (facts ?facts))
    (fact (type ?type)(OBJECT ?fact))
    (test (?facts contains ?fact))
    (event (type ?type)(interest 0))
    =>
    (?story remove ?fact)
    )
```

Figure 6. Rule implemented in *Jess*

The corresponding code in Java is the one we show in Figure 7. This implementation is much faster. If we add more rules to the system, and make them sequential in a Java program, it will be even more efficient than if we implement the rules in Jess.

```
ListIterator<Fact> it = facts.listIterator();
while (it.hasNext()) {
  Fact h = it.next();
  if (h.getInterest() == 0) {
    it.remove();
  }
}
```

Figure 7. Rule implemented in *Java*

5 Conclusions

We have presented a system where interactions between agents over a long period of time can be told in natural language automatically. With this work MMORPGs can generate texts describing the gameplay for different audiences and purposes. The text could be generated at the end of the game or while a player is still playing, or it could be the script for a 3D, or a generated comic.

We have shown a particular way of generating the stories, based on rules. We have explained a three-step process for performing this task, and we have verified that for *discourse planning*, the rule-system is very dependent on the domain, and the desired type of story.

Although the implementation includes an application for the manual development of narrative structures from a log of events, it has proved impossible to contrast the results generated by the application with any manually obtained equivalent due to the sheer size of the input logs that the application is currently handling. The effort involved for human evaluators is too large for voluntary participation.

The results of the system are less impressive - when rendered in a readable text format - than they might have been if the system included an elaborate sentence planning module. The current version is just a skeleton implementation that lets down an otherwise acceptably selected and planned discourse.

6 Future work

We plan to empower the multi-agent system, through several lines of evolution. The main point where improving is always required is to build a more interesting story. The introduction of random events was a huge step in this direction, and more improvements in this field can have incredible results.

We can add more characteristics to the agents, selecting the most attractive for the context. For example, including the profession or role of each agent could be a great idea for improving the story told: knight, king, princess, wizard, priest, peasant... If a peasant kills a dragon, would be much more heroic than if a knight does so. Another good characteristic to be introduced is geographical position. In our social simulation there is a graphical visualization of the agents, distributed in a space. If we parse this (x, y) positions dividing the space into countries, we would have knights that come from a far kingdom to save the princess.

Adding characteristics is now a particular field of the agents... but what about if we give "personality" to the inanimate objects? If we give an ID and a Name to the objects of the events, we would have events like: "lost in the Lorien Forest", "found the Anduril sword", or "killed by the dragon Smaug". These events can be analyzed to generate stories in which the dragon Smaug killed three knights (with their names), but the fourth one, Aragorn, at last killed him and freed the Gondor kingdom.

The relationships between agents represent another sector where we can add complexity. New type of relations could be included: hate (natural feeling between orcs and elfs), complex family relationships (like cousins), to belong to the same religious order...

The most part of the fantastic life events (like killing a dragon) are generated randomly in every agent. Thus, the events are particular for each agent. A new type of event could be generated: a common random event, which could affect to lots of agents at the same time (maybe to the whole world, maybe to just one kingdom). For example, a huge battle in the year 527 between dwarfs and orcs, killing lots of them, harming others, killing loved ones... and even it can lead to a prince that inherit the crown of his dead father.

Other improvements are planned for the story generation tool. A new objective can be to find a more efficient alternative to the one we tried with Jess only, perhaps a hybrid implementation between the speed of a procedural language, and the flexibility and power of a rule definition language, so the tool can be built in a more modular way, and also having the benefit of an easier to write system. Of course, another line of evolution is to enlarge the amount of rules that control the rule-based system, so more precise and complex knowledge can be used.

Another important objective is to apply more sophisticated time representation and reasoning concepts for fact and block nexus. It is very important to focus on how we narrate the story in terms of choosing what should be told before, and how we connect it with the rest of the discourse.

Different approaches to story generation are planned, and future comparisons between this work and them. An interesting line of research that is contemplated is to consider whether a Case-Based Reasoning solution, applying in discourse planning a set of patterns learned from the way humans have told similar sequences of events in human-generated stories, might compete with the simple rule-based solution.

ACKNOWLEDGEMENTS

This work is partially supported by the Spanish Ministry of Education and Science project TIN2006-14433-C02-01, and research group grant UCM-CAM-910494, jointly funded by Universidad Complutense de Madrid and the Comunidad Autónoma de Madrid (Dirección General de Universidades e Investigación).

REFERENCES

[1] James F. Allen, 'Time and time again: the many ways to represent time', *International Journal of Intelligent Systems*, **6**, 341–355, (1991).

[2] M. E. Bratman, *Intention, Plans, and Practical Reason*, Harvard University Press, Cambridge, MA, 1987.

[3] Ernest Friedman-Hill. Jess, the rule engine for the java platform. http://herzberg.ca.sandia.gov/jess/, 2006.

[4] E. Hovy, 'Automated discourse generation using discourse structure relations', *Artificial Intelligence*, **63**(1-2), 341–386, (1993).

[5] W. Mann and S. Thompson, 'Rhetorical structure theory: Towards a functional theory of text organization', *Text*, **3**, 243–281, (1988).

[6] K. McKeown, 'Discourse strategies for generating natural language text', *Artificial Intelligence*, **27**, 1–42, (1985).

[7] J. Pavon, M. Arroyo, S. Hassan, and C. Sansores, 'Simulacion de sistemas sociales con agentes software', in *Actas del Campus Multidisciplinar en Percepcion e Inteligencia, CMPI-2006*, volume I, pp. 389–400, (2006).

[8] Francisco C. Pereira, Raquel Hervás, Pablo Gervás, and Amilcar Cardoso, 'A multiagent text generator with simple rhetorical habilities', in *Proc. of the AAAI-06 Workshop on Computational Aesthetics: AI Approaches to Beauty and Happiness, July 2006*. AAAI Press, (2006).

[9] E. Reiter and R. Dale, *Building Natural Language Generation Systems*, Cambridge University Press, 2000.

Effects of Narrative Levels on Comprehension : Theoretical Framework and Methodology

Baptiste Campion[1]

Abstract. Studying educative interactive narrative, we define the deep level as characterized by a conjunction between the storyworld and comprehension macrostructure ; we define the surface level as characterized by a disjunction between the storyworld and comprehension macrostructure. Both are often used in interactive designed for children. The goal of this contribution is to present work in progress that intend to evaluate educative effects of both levels. First, we will present the whole research and its theoretical bases ; second we will present what it is set up for empirical evaluation.

1 INTRODUCTION

Narration is often used in edutainment products. Sometimes, it seems that it's only for 'packing' (presumed) boring educative content. Narrative is supposed to be attractive despite difficulties linked to the educational content. In other cases, narration results from a scenarisation process of the hypermedia (for example an hypermedia structured around a quest). And there are another cases, when narration and educative content seem be set up together (i.e. due to structural convergences, like for historical contents). All these examples show that there are different uses of narration in educative interactive documents. And it shows also that if narration constitutes a structure for all these documents, narration can imply very different documents and, thus, different comprehension processes for a given reader/user.

For these reasons, distinguishing between all these situations is important. Distinctions must furthermore be used for setting up some reception models focusing on possibilities of different ways of using narration. These models should be useful for researchers in education, but also for designers. If we can prove there are some significant comprehension differences between different ways of using narration in educative narrative, you will not write the same story if you want to focus reader's attention on one aspects more than another one. Results should be valuable as well for 'classic' (linear) narratives as for interactive narratives or narrativised educational games.

We will present in this paper some elements of an undertaken research about educative use of narrative, especially in interactive narrative. Because this research is still a work in progress, this contribution intends focus on theoretical and methodological issues with broader interest. But this aspects will be enlighted by some empirical elements. We will focus on a single assumption but this research counts other dimensions we will not discuss here. This focused assumption concerns what we called the 'level' of narration use in educative narratives, which is illustrated by previous examples. First there will be a short presentation of research theoretical framework. Then experimental design will be presented and discussed.

2 THEORETICAL BACKGROUND

2.1 General context

The general purpose of our research is to investigate educative effects of narrative use for science pupularization. Can we learn something when explained with narratives ? 'Effects of narratives' are defined in terms of cognitive effects : how can subjects use narratives in order to understand parts of the world narratives are talking about ? Indeed, we make some distinction between understanding a story and learning somewhat from a story. Only this last case is called 'comprehension'.

First, we will describe how people understand narrative. Then, in next section, we will see how we can consider narration as an cognitive resource for readers. By this way, we will have at disposal some model describing how storytelling can be used in education.

2.2 Narrative comprehension

Following van Dijk and Kintsch [17], we define discourse comprehension as the constitution (by the receiver) of a mental representation integrating and articulating inputs. Following this theory, readers 'comprehend' a discourse (we generically call 'text') through a double process of construction of a coherent representation of discourse and construction of a model of the situation this discourse is speaking about. This process results from an automated (mental) strategie. Schema theory can be use for describing the integration/organization of picked-from-the-text elements in a coherent mental representation [2], [15], [16].

What about comprehension of narrative ? Narrative comprehension is basically a discourse comprehension operation even if narratives are particular discourses. In narratology —with the *story schema* theory [13], [14]— 'schema' definition remains ambiguous because it can either refers to mental structure or parts of the story (semiotic structure) [3, p.381]. So, we prefer describe these mental structures with the mental models theory [9], [10] (which is not incompatible with the schema theory). According to this model, various cognitive operations result from (non propositional calculation) operations carried out on the basis of running a 'mental model'. This model of the world is far away from the syntactic structure of narrative sentences, even it's based

[1] Groupe de recherche en médiation des savoirs (GReMS), Université catholique de Louvain, Belgium, email: campion@reco.ucl.ac.be

on a narrative and is a model of the world narrative is speaking about. Signification cannot be reduced to a purely intra-linguistic operation [10].

If we follow Herman's cognitive narratology [8], [6], narratives suppose a double mechanism of story comprehension and construction of a situation model similar to this postulated by van Dijk and Kintsch, and which can be completed in terms of mental models. Herman considers that comprehension of a narrative passes by the constitution of a 'storyworld' [7], i.e. a mental model of situation defining some elements useful to locate, contextualize and interpret the narration. The storyworld is built from the narrative text when the reader articulates bottum-up and top-down operations in two stages. Level of the microdesign (bottom-up) for the reader consists in establishment of an inventory at the local level while concentrating on 'What's going on ?'. The macrodesign (top-down) level refers to integration of these various parameters in a higher level whose result will consist in a mental model of situation.

2.3 Can narratives be used for comprehension ?

Lots of works have shown such comprehension mechanisms. But what it is interesting is that we can use the constitution of a given mental model by the narrative reader, to present the assumption that this mental model —the storyworld— can be used for later cognitive operations based on this model. Herman, following Vygotsky's 'cognitive artifact', considers narrative as a general cognitive tool : "I argue that stories provide crucial representational tools facilitating humans' effort to organize multiple knowledge domains, each with its attendant sets of beliefs and procedures. […] My hypothesis is that stories provide, to a degree that needs to be determined by future research, *domain-general* tools for thinking" [8, pp.157-159]. This postulate enables studying the knowledge and the comprehension of the world conveyed through narration, or more exactly through the mediation of a storyworld built on the narration.

This not only happens in *educative* narrative, but potentially in all kinds of narrative. But because we wants precisely see how narrative can be used as tools for learning, specific inquiry must be set up.

There are no reasons of thinking that this is not true for interactive narrative or even some narrative games (due to narrative structure of most of them based, for example on a quest schema), even if it's possible to formulate opposite assumptions about the effective effects of interactivity and non-linearity[2].

3 THE 'LEVELS' OF NARRATION

So, readers constitute a mental model of what they have read [7], and this mental model can be used by people for later mental operations (for example : inference). The question now is : when narrative contains specific educative stuff (explanation of a scientific phenomenon, historical precisions, etc.), how is it implemented to the storyworld ? Or : has the specific educative content a different place in reader's storyworld in different narratives ? More concretely, designers will ask how to implement educative content in a narrative so that the narrative will encounter the (correct) planned educative effect.

The concept of storyworld allows to define different ways using narrative in educative interactive documents. We call these ways 'levels' even if there is no normative judgement about it. We define two opposite levels of using narration : a 'surface level' and a 'deep level'. In both cases, new knowledge must be extracted from narrative, but we assume that the way it is done differs from one case to another. Last, these two cases can be viewed as extreme poles of a continuum on which we can place most of educative narrative productions.

The surface level appears when one gives a 'narrative packing' to some educative content in order to transmit educative information to the reader. In this case, the storyworld does not relate to the field of knowledge which one wants to speak about in the narrative, but it refers to the situation of the narrative (characters, actions, etc.). In this case we assume that understanding a narrative is not sufficient to reach comprehension. Readers must integrate specific integrative information in another mental model : the storyworld doesn't help for integration.

The deep level consists in using the narration structure itself to transmit the matter. There is a stronger integration between the field of knowledge and narration ; the storyworld can be used as basis for real appropriation and integration of this knowledge. Readers can base their comprehension of educative content on the storyworld, even if abstraction/extraction work has probably to be done for total integration of new knowledge.

The main consequence of this assumption is that formal aspect of a narrative should directly influence comprehension of educational data integrated to the narrative. Effect depends on reader's focus which depends on used level. Reader's capacity of extracting and integrating new data should be greater with deep level. In surface level case, disjunction between the story itself and educative stuff should cause integration (to a coherent mental model of the explained situation) problem. But that does not mean that first case is *better* than the second one : it depends on the planned/desired effect. We test here comprehension, not memorization, for example.

4 CURRENT EXPERIMENTATIONS

4.1 Research assumption

This framework leads us to the following research assumption : deep level narrative should lead subjects to build to a relatively unified representation. On the contrary a surface level narrative should oblige subjects to work with two levels of representation : one for the story itself, and the other for the educative contents.

This assumption is currently being quasi-experimentally tested with specific educative interactive narrative explaining to children a scientific phenomenon. We speak about 'quasi'-experimentation [4] because it will be performed in schools rather than in real lab conditions.

The dependent variable is thus the coherence of the mental model/representation of the scientific phenomenon. The explicative variable is the level of narrative use (deep/surface). Other variables will be controlled as much as possible. In particular, we will neutralize the 'interactive' or non-linear variable[3] : all experimental document will be strictly linear for this quasi-experimentation. Finally, our population sample can be considered as 'equivalent' in terms of scholar skills because we will carry this out in classrooms in the same degree.

[2] These assumptions about interactivity effects are for example partially developed in [3].

[3] This quasi-experimentation is a part of a broader research for which we also test effect of linearity/non-linearity with similar interactive documents. For this specific test, we don't use any non-linear document.

4.2 Methodology

We will compare representations of a scientific phenomenon acquired by two groups of children from a deep level narrative and from a surface level narrative. We will control these results with those of two other experimental conditions : a group who read a non-narrative explanation, and a control group without any explanation about the phenomenon. (This last is set up only to control children skills about the matter.)

The comparison will focus on children ability to synthetically explain the scientific phenomenon explained in interactive document. We indeed postulate that discourses held by subjects contain 'traces' of mental model used by subjects to understand the situation they are speaking about. We need this postulate in order to consider any empirical experimentation about such phenomena. It is consistent with works about language postulating and/or highlighting linguistic traces of the subjacent cognitive activity[4].

So, our data will consist in written discourse held by subjects as they were answering a research questionnaire after reading the interactive document. This questionnaire contains four questions. One is a recall question (they have to explain what they remember about what's explained in the document). One another is a problem-solving question (subjects have to solve a problem which need a good comprehension of the scientific phenomena). Third is a 'drawing' question (subject have to make a schema of the phenomenon). The last one consist in words explanation ('what's a bacterium ?', etc.). These questions should enable us to sketch central dimensions of the subject's mental model (storyworld).

Our indicators are :

- Elements and relations between elements (spatial relations, inclusion, exclusion, superposition, motion…) in pictures ;
- Specific vocabulary used by subjects when describing the scientific phenomenon, especially action verbs, personification, names, etc. ;
- Conjunction or disjunction between answers ;
- Subjects ability to abstract and re-use gathered info (in problem-solving question).

All groups will have the same questionnaire, behalf the control group (condition without any document) where the recall question (that makes no sense) is suppressed.

4.3 Experimental material

We will work with around 100 children of Belgian 5th year elementary school (± 11 years old). They will each read one version of the experimental interactive documents built for the experience. These documents are HTML pages These documents explain a simple 'scientific' phenomenon : how do tooth decay develop in the mouth ? Three versions of the experimental document have been built[5]. They are partially derived from a former study on narration and memorization [5] because it showed they were suitable for 11 years old children.

Two versions of the explanation are defined as 'narrative', following Adam's six criteria. It's indeed difficult to characterize exactly a text as 'narrative' even if everybody know spontaneously what a narrative is. So we use Adam's criteria [1]. It's not the only way to define a narrative and each criterion could be discussed, but we assume that if each criterion is individually respected, the text can surely be considered as a narrative. For Adam, a narrative is characterized by : (1) a temporal succession of actions, (2) a thematic unity, (3) predicates transformation, (4) a process, (5) narrative causality-consecution in dramatization and (6) a final evaluation [1, pp. 92-110].

Both experimental narratives are written following all six items, but in two different ways. The first one is defined as a 'deep level narrative' : scientific content is narrowly integrated to the story (it's the story of a bacterium who tries to perforate a tooth in the mouth). We consider there is a narrow integration because characters (bacterium), processes (transformation of sugars into acids) and other agents are the same for understanding narrative *and* understanding how does a decay develop. The other one is defined as a 'surface level narrative' where we maximized disjunction between the story (it's the story of a boy who musts go to dentist before a match play) and scientific content (how does tooth decay develop). These two versions correspond to modalities of 'level use of narration' variable.

The third (and last) version is defined as a 'non-narrative' condition : that's an explanatory text where we paid attention not to follow Adam's criteria when it make sense. For example there are no characters, no predicates transformation, no dramatization.

All scientific (i.e dentistry related) information has been controlled so that it is strictly equivalent between conditions. Each condition will count around 25 pupils.

4.4 Forthcoming results

Data acquisition is currently under way. Some data were already collected in two schools. The full tests should be performed for april-may 2007.

5 CONCLUSION

The main goal of this research is to enlighten the presumed role of what we called the level of narration use in comprehension of a phenomenon. Even if we conclude with significant results, that will not mean there is a *normative* difference between levels of narration use. We hope this experimentation will provide sufficient data in order to perform additional qualitative and comprehensive interviews with other subjects. The purpose of this forthcoming phase will be enlightening elements required for a better integration of so-acquired knowledge.

If our assumptions about surface and deep level are verified, further works should focus on precise effects of these levels in terms of comprehension in relation with hypermedia elements that enable (or prevent) conscious use of one level or another. In particular, it will be useful to focus on the mechanisms of extraction of scientific information in the two configurations.

Another axis of investigation is the interaction between levels of narrative use and reader's implication, especially in interactive stories and games. We can for example presume that improving reader's 'first person' central experience increase effects of deep level because it's own experience is mobilized in defining a mental model of the matter.

[4] See for example the cognitive grammar of Langacker [12] or the works about metaphor of Lakoff and Johnson [11]. These authors show (each one on their specific object) how the language contains traces of mental operations and structures on which would be based our knowledge of the world.

[5] These can be read for a while at following URLs (all documents are in French) : *http://www.comu.ucl.ac.be/reco/grems/batweb/expe/site2/* for deep level narrative ; *.../site4/* for surface level narrative ; and *.../site3/* for non-narrative condition.

ACKNOWLEDGEMENTS

Many thanks to Estelle Fierens who allowed me to use and transform her original experimental story [5] for this current experimentation.

REFERENCES

[1] J-M. Adam, *Le texte narratif : traité d'analyse pragmatique et textuelle,* Nathan, Paris, 1996.

[2] F.C. Bartlett, *Remembering : a study in experimental and social psychology,* Cambridge University Press, Cambridge, 1995 (new edition of his 1932 book).

[3] B. Campion, 'Theoretical Framework for Construction of Representation Through Interactive Narrative', in Z. Pan & al. (Eds), *Edutainment 2006*, LNCS 3942, Springer-Verlag Berlin Heidelberg, 380-388, 2006.

[4] T.D. Cook, D.T. Campbell, *Quasi-experimentation. Design 1 Analysis Issues for Field Settings,* Rand Mc Nally, Chicago (IL), 1979.

[5] E. Fierens, *Le récit au service des documents socio-éducatifs. Vérification de l'apport de la forme narrative en tant que stratégie communicationnelle sur la construction des connaissances du public,* Undergraduate thesis, Université catholique de Louvain, Faculté des sciences économiques, sociales et politiques (Département de communication), Louvain-la-Neuve, unpublished, 2004.

[6] D. Herman, Narratology as a cognitive science, *Image [&] Narrative (Online Magazine of the Visual Narrative, K.U. Leuven),* **1**, 2000.

[7] D. Herman, *Story Logic. Problems and Possibilities of Narrative,* University of Nebraska Press (Frontiers of Narrative), Lincoln and London, 2002.

[8] D. Herman (ed.), *Narrative Theory and the Cognitive Sciences,* CSLI Publications, 2003.

[9] Ph. Johnson-Laird, *Mental Models. Towards a Cognitive science of Language, Inference and Consciousness,* Cambridge University Press, Cambridge, 1983.

[10] Ph. Johnson-Laird, 'La théorie des modèles mentaux', in M-F. Ehrlich, H. Tardieu, M. Cavazza, M. (eds), *Les modèles mentaux. Approche cognitive des representations,* Masson, Paris, 1993.

[11] R.W. Langacker, *Foundations of Cognitive Grammar, vol. 1, Theoretical Prerequisites,* Stanford University Press, Stanford, 1987.

[12] G. Lakoff, M. Johnson, *Metaphors we live by,* University of Chicago Press Chicago (IL), 1980.

[13] J.M. Mandler, *Stories, scripts and scenes : aspects of schema theory,* Lawrence Erlbaum Associate, Hillsdale (N.J.), 1984.

[14] D.E. Rumelhart, 'Note on a schema for stories', in D.G. Bobrow, A. Collins (eds), *Representation and understanding. Studies in cognitive science,* Academic Press, New York San Francisco London, 211-236, 1975.

[15] D.E. Rumelhart, A. Ortony, 'The representation of knowledge in memory', in R.C. Anderson, R.J. Spiro, W.E. Montague (eds), *Schooling and the acquisition of knowledge,* L. Erlbaum, Hillsdale, 99-135, 1977.

[16] D.E. Rumelhart, D.A. Norman, 'Representation in memory', *Technical Report CHIP 116 Steven's Handbook of Experimental Psychology,* University of California at San Diego, 38-48, 1983.

[17] T. van Dijk, W. Kintsch, *Strategies of Discourse Comprehension,* Academic Press Inc., New York, 1983.

Towards a classification of Video Games

Djaouti Damien[1], Alvarez Julian [2], Jessel Jean-Pierre [3], Methel Gilles[4] and Molinier Pierre[5]

Abstract. This paper is part of an experimental approach aimed to raise a video games classification.
Being inspired by the methodology that Propp[3] used for the classification of Russian fairy tales, we have cleared out recurrent diagrams within rules of video games, named "Game Bricks". The combinations of these different bricks will allow us to represent a classification, in accordance to their rules, of all the video games.

In this article, we will study the real link between these bricks and the rules of video games, trough realisation of an experimental ”brick-only” based game.

1 INTRODUCTION

The idea of classification of video games is not a new idea of course. Le Diberder brothers[4], or Stephane Natkin[5] have already raised classifications.

But, in all these works, even though they are references, we have rapidly found absences or slants. These facts are denounced by Mattieu Letourneux[6] in his article “The question of the kind of video games”: To him, any video game classification is condemned by its very nature to the obsolescence, because games technological evolution also modifies the chosen criterions.

How define what a video game is, if its classification is rapidly wrong?

Being inspired by the Propp’s methodology[3], we have exposed in a previous article[1] the genesis of this project that leads to the development of "V.E.Ga.S", a tool to index and analyse video games. Influenced by Salen & Zimmerman we focused on the game rules[14].

With this tool and a list of 588 video games we have proposed a first step[2] of the development of a classification criterion: we have emphasized the "Game Bricks"(figure 1), the "fundamental elements" whose different combinations seem to correspond to different rules and aims of a video game ("Game" aims to the "game rules" notion, referring to Gilles Brougère).

Figure 1: The Game Bricks known up today

The number of "different combinations" thus obtained was rather high, but we have noticed that some pairs of bricks, named "Metabricks" (Figure 2) were recurrently found in a large number of combinations.

After analysis [2], we have realized that these "MetaBricks" really seemed to outline an encouraging path towards a classification of video games.

Figure 2: The two MetaBricks discovered up till today

To summarize, we have identified "Game Bricks" that represent "tasks to carry out" within the video games. Based on these bricks, we have updated a classification based on groups of video games into "families" having identical combinations of "Game Bricks", these families could be regrouped by the presence or not of some pairs of bricks named "MetaBricks".

For example, the Game Bricks featured in “Pac-man” are : “MOVE”, meaning player can move an avatar, “AVOID” for the Ghosts you have to avoid, “DESTROY” for the dots you have to eat, and “POSITION” because you have to reach each dot’s spatial position to destroy it.

But you can also find these Bricks in the race game like “Need for Speed”: MOVE a car, AVOID opponents, and POSITION on checkpoints you have to DESTROY. When reached a checkpoint becomes “out of the game” and is not reachable anymore, so it can be considered “destroyed”, just like a dot eaten by Pacman.

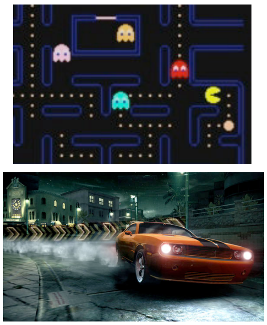


Figure 3: From the outside, nothing seems to rely Pacman (Namco 1980) and Need for Speed Carbon (E.A. 2006).

As both games feature the same bricks, they are classified in the same family, one of the game families featuring the “DRIVER” MetaBrick (MOVE+AVOID).

There are nevertheless problems left to be resolved, that we wish to solve to make an improved analysis tool.

We have to try to reduce the part of subjectivity which appears during the valuation of a video game. Two complementary approaches appear then to us:

- A quantitative approach, which notifies several entrances for each game, thanks to contributions.

[1] IRIT, Université Toulouse III & LARA, Université Toulouse II, France ; daminous@gmail.com
[2] IRIT, Université Toulouse III & LARA, Université Toulouse II, France ; alvarez@irit.fr
[3] IRIT, Université Toulouse III, France ; jessel@irit.fr
[4] LARA, Université Toulouse II, France ; methel@univ-tlse2.fr
[5] LARA, Université Toulouse II, France ; pierre.molinier@univ-tlse2.fr

- A qualitative approach, which eliminate at the most the subjective aspect of the definition of the "Game Bricks".

On the other hand, the definitions of some bricks like ANSWER are in a lack of precision. This problem is due to the fact that we are still not able to fully answer the question: "What do really the bricks represent concerning the video games?"

The aim of this article is thus to propose a formal definition of "Game Bricks".

At first we will introduce an experimental validation work about bricks, followed by thoughts about the very nature of the bricks and their relationships to the rules of video games.

These two steps will allow us to propose a positive definition of the bricks, considered as criterions among a classification of video games in accordance to their rules.

2 EXPERIMENTAL VALIDATION

2.1 Specifications

In order to test the pertinence of our bricks, we have elaborated an application and the target is to help us to see how, on a data basis, the "Game Bricks" are put together in a video game.

Ideally, it would be an application allowing us to add or remove "Game Bricks" in order to be able to observe the impact on the game. This stage implies a finite definition of the bricks, in order to be able to insert them in a program.

Being inspired by the works of Raph Koster[7] and Stéphane Bura[8] who both try to elaborate a grammar of video games in the shape of diagrams, we have thus formalised diagrams as definitions of our bricks.

With the idea to handle the rules of a video game on a data basis, we have thus thought of a game representation model in an algorithmic way.

We have been inspired by the works of Michael Thielscher [9] in "the General Game Playing", who creates programs for games being able to play games with rules that are initially unknown. His team has developed in particular a language, the GDL (Game Description Language), which allows representing a game in a logical way by describing its rules and its initial state.

We have also been very inspired by the “games creation softwares”, like those created by Clickteam [10]: "Klik n’Play", "The Games Factory" and "Multimedia Fusion". These softwares are an aid in the creation of video games: they withdraw the technical part and allow the Game Designer to focalize on the rules of the game, the graphics and the sounds, as well as the control of the interfaces. The construction of levels and game scenes (level design) is also easier by using these tools.

2.2 Conceptual representation of a game

We rely on the definition of a game according to Katie Salen and Eric Zimmerman [11]: "An activity with some rules engaged in for an outcome".

Katie Salen and Eric Zimmerman thus consider a game as an activity defined by two elements: The rules and the result, the last one according to a previous goal.

2.2.1 *The game rules: "some rules".*

If we consider that a video game takes place in a virtual universe and that it is composed by several "elements", in a large point of view, then these different elements are submitted to "rules", in accordance to the game like the elements composing our own universe which are governed by physical and behavioural rules.

For example, the universe of the game "Pong" is composed by the following elements: The racket of the player, the adverse racket and the ball. The area of the game (the size of the screen) can also be considered as an element, even though it doesn't have a graphical representation, it does "exist" within the rules of the game.

These elements are submitted to different rules like “Each frame, the ball element moves according to an (x;y) vector”, or further on ,"if the ball touches a racket, then its vector of movement (x;y) becomes (-x;y)".

Analysing this last rule, we will realise that it is composed by two parts:

- The "trigger": "if the ball touches a racket, "
- The "effect(s)": "then its vector of the movement (x;y) becomes (-x;y)".

We will call "targets", the elements to which are applied those rules.

We will notify a similitude between this conceptual representation and the algorithmic or even programming on the whole: a condition ("if") driving to the production of a succession of instructions ("then").

2.2.2 *The objective of a game: "an outcome"*

In the same logic, the aim of a game can also be described by its rules, for example by Pacman: "if all the pastilles have been eaten, then the level is "won"". It is all about a rule having an effect corresponding to "the game has been won" (moving up to the following level, end of the game...), associated to a condition formalizing a target to be obtained.

At this level, we consider that it's logic to include "the objective of the game" into "the whole of the game rules", the "Game" part of a video game.

2.2.3 *Conceptual Diagram*

We will then obtain a model permitting us to describe a game by enumerating the elements of its universe, elements applied to the whole of the rules, including the objective of the game.
These rules are composed by different triggers and effects (figure 4).

Figure 4: Conceptual diagram of a game

2.3 The modifiable game: “Gam.B.A.S.”

Starting from this design, we have programmed a whole of "elements", "triggers" and "effects". The elements position is randomly chosen, we do not include any aspect of level design in this experimental game for now.

Figure 5: Gam.B.A.S. (Game Bricks As Switches)

Further on , we have programmed triggers like "Always", triggered on every frame, "MouseDown", triggered when you push the left button of the mouse, or even "Collision", when two elements collides.

These triggers are linked to one or several effects such as "CreateElement", "DestroyElement" or also "Move Element" applied when the condition of the trigger is "true".

Figure 6: Simplified Class Diagram of Gam.B.A.S.

We have then been able to gather these elements, triggers and effects in order to realise basic video games: a game where you have to collect some elements and avoid others, recalling "Pacman", or even a game where you have to destroy elements that you don't have to touch, shooting on them, recalling the famous "Space Invaders".

2.4 The very nature of "Game Bricks"

At this stage, we have decided to set up in these "games" the "Game Bricks", based on the logic diagrams being defined in the “Specifications” phase *(see 2.1)*.

In order to simplify, we will not set up neither the bricks of "TIME" and "SCORE", nor the brick of "TOY" because of a lack of satisfying diagrams.

We then realized that the "POSITION" brick is composed by a "Collision" trigger between two elements with spatial coordinates. The "SHOOT" brick features a "CreateElement" effect, and the brick "DESTROY" is composed by a "DestroyElement" effect applied to every element of the scene except of those relied to the player.

We finally observe that it is possible to build our bricks by assembling elements based on the previous definitions: the triggers and the effects.

These two being "the construction elements" of the rules, we realise that the "Game Bricks" can thus be translated into "game rules".

We also notice that the bricks definition diagrams can not be translated directly into rules: actually, there are within these definitions "areas of liberty", especially about the elements that are targeted by the rules. For example the definition diagram of the "Move" brick specify its effects are applied on "element relied to the player", but it doesn't specify the number of these elements: Is it about one unique piece or a whole army of mutant orcs?

The translation of definition diagrams into rules needs to answer this kind of questions.

2.5 Statement of the experiment

For the needs of this experiment we had to:

- Define a model of the representation of a game: a universe composed by elements to which rules are applied.
- Define "elements of construction" for the game rule: they are composed by two elements, the "triggers" and the "effects".
- Establish definition diagrams for Game Bricks.

At this stage we will define the "Game Bricks" as "a canvas of rules", a diagram to follow in order to build a rule or a group of rules in a video game.

Nevertheless, if we observe the games obtained by the successive realisation of different bricks, even though they unquestionably remind us the basic principles of the classified games, we realise that we don't obtain precisely one of them.

For example, after having activated the bricks of the game of "Pacman", it seems that there still is a "lack of rules" compared to the original game: there are no "special dots" that make the "ghosts" edibles, the ghosts/elements to avoid don't move.....

We finally realise that all the rules of a game are not covered by the bricks. This "no-exhaustiveness of the video game rules" finds its answer in the objective of bricks, which intend to be a criterion to a classification, but will return to this point further on.

Figure 7: Conceptual view of a game: rules covered by a Brick are created from its template, unlike uncovered rules.

3 A VIDEO GAMES CLASSIFICATION ACCORDING TO THEIR RULES

The objective of the study of the "Game Bricks", according to the previous articles [1] and [2], is to achieve a definition of criterions for a classification of the Video Games. The "Game Bricks" should thus be these criterions, as their association into "Metabricks" will allow us to obtain "families" recalling those of the Russian tales classification by Propp[3].

The works on the very nature of "the Game Bricks" described previously have permitted us to achieve the following observation: the bricks represent "diagrams of game rules", translated into rules by the specification of "areas of liberty" present in their definitions.

These "areas of liberty", generally relied to the elements targeted by the rules or "feedbacks" definitions, have been included intentionally within these bricks.

Actually, a precise definition for an effect like “the Pacman element moves 15 pixels north" or "the Pacman element moves 12 pixels east", yet matches exactly to the rules of Pacman, but would be completely unusable for a classification : the number of rules and thus of bricks would be extremely large with such precise definitions.

The combination of bricks allows us to represent the whole of the games being indexed, but it doesn't represent them in an exhaustive manner: numbers of rules are not included in the definitions of the bricks.

It's a choice made in order to limit the number of the bricks, and thus the criterions of our classification in order to enlarge the performance of it.
We therefore have concentrated our efforts on representing the rules related to the actions of the player with the "Game Bricks".

In accordance with the methodology described by Propp[3] and his classification, we have decided criterions of our classification, the "Game Bricks" form the video games : we have observed indexed games, and we have identified "recurrent rule diagrams". These different "rule diagrams" are, as it has already been said, the definition of the "Game Bricks".
The bricks we have identified at present are the issue of 588 games being indexed in a first version of V.E.Ga.S. and are the result of an iterative approach, as former versions of bricks were created from smaller corpus[1].

The differences between games featuring the same bricks, strictly concerning the rules, are coming out from these two choices of limitation of the precision of the bricks with the aim to obtain a relevant classification.

According to the former version of bricks[2] we notify the disappearance of the "ANSWER" brick, because its definition was too large, and which intend to be replace by two new bricks : "SELECT" and "WRITE".
We also notify the removal of the "SCORE", "TIME" and "TOY" bricks, which weren't directly related to rules, and an enlargement of the definition of the "POSITION" brick which becomes the "MATCH" brick.

For instance, within the limits of the game rules, we notice great resemblances with the game of "Pacman" and a racing game such as "Need for Speed Carbon": in both games you have to move an element (Pacman/car >> "MOVE" brick), and avoid others (Ghosts/Rivals >> "AVOID" brick) that it is possible to destroy

("DESTROY" brick), and finally passing by several succeeding spatial positions (Swallow dots/Pass the checkpoints >> "POSITION" brick);

Nevertheless, even within their rules, these two games are different: the movement and thus the "MOVE" brick has two dimensions in “Pacman”, but three in “Need for Speed Carbon”, the number of checkpoints to pass in the last one is much smaller than the numbers of dots that Pac-man has to swallow and the movement of the elements to avoid is different in the two games....

These differences between these two example games are the issue of different implementations of "rule diagrams" from the bricks they are sharing, but are also due to the use of rules are not covered by the bricks, as the frequency of these rules in our corpus were to weak for us to index them as a “rule diagram” in a “Game Brick”.

4 GAME BRICKS DEFINITIONS

We will here introduce the diagrams of the different "Game Bricks" that we identified up till today.
These bricks will be used as criterions of classification in a further version of V.E.Ga.S., our tool of video game indexation and analysis.

5 CONCLUSION

We hope that we have clarified by this article **the very** nature of "the Game Bricks" having been clear about the choices at the time of their construction in the target to use them as criterions of a classification of video games according to their rules;

Answering the article by Matthieu Letourneux, "The question of the kind of video games"[6] that points out the short life of the video games classifications due to the lack of “no-evolution criterions”, we consider that the game rules of the video game seem to be an interesting criterion by the fact of its obvious redundancy between different games. We also notify that this aspect of the video game doesn't seem to be submitted to an evolution as quick as the one concerning for example the control devices or the graphic aspects, which make “the rules” particularly interesting for a classification criterion.

We can nevertheless establish a relationship between the "The game rules" and the "middleware". The "middleware" corresponds to different "engines" (game, graphics, physics, sounds....) sold separately and that permit the creators not to reprogram the redundant parts of their different games.
These games engines are generally distributed with the pre-programmed rules, rules that you will find in the important lines in all games of the same "kind", according to the classifications by the specialized press (Shoot'em up, FPS, RTS,…)

We consider this as a real example of the small variation of "game rules" between the games considered as being the "same style", when these same games offers different graphics or controls.

This article helped us to reconsider what is a "Game Brick" in accordance to a game: a rule diagram, or more precisely "recurrent game rules diagrams". We realize then, that the choice of the creation or not of a brick relies on the evaluation of the pertinence of the diagram as well as the definition of its "areas of liberty".
As we previously have explained, the bricks that we have identified up till today are the result of an heuristic approach from 588 games. We pretend neither to have identified all the bricks, nor to have identified the more pertinent diagrams.

We are aware of the fact that the planned increase of our indexed games will lead to an refining of the bricks definitions, or maybe even be the discovery of new bricks or meta-bricks.
The "Game Bricks" showed in this article, along with their definitions, will be used as criterions for the classification being included in the second version of "V.EGa.S”. As we have mentioned in the introduction, we wish to decrease the part of subjectivity during the evaluation of the games, done by the human being, thanks to an approach of quality as well as a quantitative approach.

While the current paper is a part of the qualitative aspect, the quantitative aspect is related to the opening to the public of our V.E.Ga.S database. We will thus apply for a contribution concerning the inventory and the evaluation of the games, the bricks featured in a game will then be chosen according to the statistics of the different evaluations that the game received.

You may offer, evaluate or get informed about a game in the online version of our classification:

http://www.gameclassification.com

Nevertheless, it is obvious that a game is not made only of rules, it also features a graphic aspect, interfaces, and a content. Talking about content, the work presented here get a broader meaning when focused on “Serious Games”.

The current article was focused on the “Game” part of Serious Games, and need to be related with the work on the “Serious” part presented in our second paper [12]. This one started from the analysis of five Serious Games areas: Edutainement, Advergaming, Edumarket Games, Political Games and Training games.

This analysis led us to conclude that these Serious Games are composed of two main categories:

- Serious Games based on simulation which present a “world”, with its “rules” and where there is no objective imposed by the application.
- Serious Games based on video games which propose a “world”, with its “rules” and implemented objectives that the user has to reach.

We can observe both categories of Serious Games seem to take place in a “virtual world with its rules”, thus we can see the role of the “rule analysis” work presented here.
But we can also notice that some Serious Games features an “objective to reach”, whereas the first category, based on simulators, doesn’t impose any objective.

Can the “Game Bricks” be applied on the “Game” part of both Serious Games categories?
Or does the lack of objective of the first category imply its games will use a different set of bricks?

We will try to work on theses questions on our future works.

6 GREETINGS

The authors wish to thank Jean-Yves Plantec and Martial Bret from the "Iode" Society, for their point of view concerning the notion of "Brick", as well as Stéphane Bura, Art Director at 10Tacle Studios, who have let us know a great number of references.
We also wish to thanks a lot Annika Hammarberg for the translation of this paper from French to English, and Rashid Ghassempouri for his general help and thoughts in the earlier works about this game classification.

7 REFERENCES

[1] Alvarez, J., Djaouti, D., Ghassempouri, R., Jessel, J.P., Methel, G., V.E.Ga.S.: A tool to study morphology of the video games, Games2006 Portalegre - Portugal (2006).

[2] Alvarez, J., Djaouti, D., Ghassempouri, R., Jessel, J.P., Methel, G., Morphological study of the video games, CGIE2006 Perth - Australia, (2006).

[3] Propp, V., Morphologie du conte (1928), Seuil, (1970)

[4] Le Diberder A. et F., Le Diberder F., L’univers des jeux vidéo, Ed. La découverte, (1998)

[5] Natkin S., Jeux vidéo et médias du XXIè siècle, Vuibert, (2004)

[6] Letourneux, M., from the Genvo, S. book : Le game design de jeux vidéo, L’Harmattan, (2005)

[7] Koster R., A Grammar of Gameplay, http://theoryoffun.com/grammar/gdc2005.htm

[8] Bura S., A Game Grammar, http://users.skynet.be/bura/diagrams/

[9] Thielscher M., Game Description Language, http://games.stanford.edu/language.html

[10] Clickteam, Game Tools, http://www.clickteam.com/

[11] Salen K. and Zimmerman E, The Rules of Play, MIT Press (2003).

[12] Alvarez, J., Rampnoux O., Jessel, J.P., Methel, G., Serious Games: just a question of posture?, AISB’07 NewCastle - Scotland (2007), to appear.

[13] Zyda M., From Visual Simulation to Virtual Reality to Games, IEEE Computer Society, 2005

[14] “Looking at games rules means looking at games as formal system, both in the sense that rules are inner structure that constitute the games and also in the sense that rules schemas are analytic tools that mathematically dissects games.” (p 104).
Salen K. and Zimmerman E, The Rules of Play, MIT Press (2003).

Serious Game: just a question of posture?

Alvarez Julian[1] and Rampnoux Olivier[2] and Jessel Jean Pierre[3] and Méthel Gilles[4]

Abstract
This article explains the difference between a large variety of Serious Games and tries to propose a classification to understand this type of video games. We explore the connection between the goal of the game designer, the objective of the game and the posture of the player. Finally, we explore how we can create some serious game to make corporate communication or educative programme.

Introduction

Great numbers of Serious Games are proposed in various fields of application like health, army, education or communication...Facing this diversity, are we really in the presence of various categories of Serious Games or is it just a variety of fields of application? If this is the case, which are the elements being characterized by each of these categories and which is the part of marketing of each variety?

In the first part of this paper, we will introduce elements that characterize a Serious Game and thus index five big categories. In the second part we will estimate the relevance of these different categories and lead a reflection to see if transmitting a message by a Serious Game is just a choice of posture that the creator of the application or the mediator tries to get adopted by the user. In fact, in some special circumstances, the players, especially the children, don't have a direct access to Serious Game, but the game might be introduced by an adult, according to Vygotsky's theory. For example, at school or in a youth center, the child does joint activities or mediatized activities. (La Ville, 2005).

1 HOW TO CHARACTERIZE SERIOUS GAME?

In its article "From Visual Simulation to Virtual Reality to Games", Mike Zyda proposes the following definition for Serious game: “A mental contest, played with a computer in accordance with specific rules, that uses entertainment to further government or corporate training, education, health, public policy, and strategic communication objectives.” (p. 26) In other words, the vocation of Serious Game is to invite the user to interact with a data-processing application whose intention is to combine at the same time teaching, training, communication, or information aspects, with ludic mechanisms based on video game. The purpose of such an association is thus to give attractive shapes or plots (Game) to didactic contents (Serious).

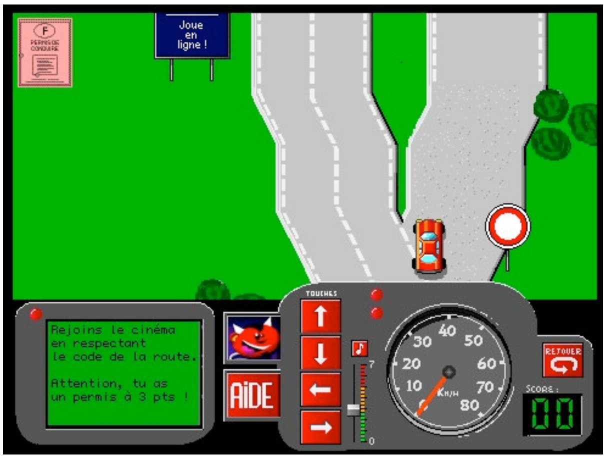
Zyda indexes a broad range of the applications concerned with Serious Games as David Michael and Sande Chen do also in their book “Serious Games:Games that Educate, Train, and Inform” (2005). In this enumeration, it is important to raise a major distinction between the applicability concerned with "health", law and order, or engineering and the categories of intentions such as “Communication Strategy" or "Education". The fields of application are too many and too subjective to be able to build a resistant typology contrary to the categories of intention which are simpler to identify and to formalize.

We propose 5 categories to classify the Serious Game: Edutainment, Advergaming, Edumarket game, Political games, and Training and simulation games.

1.1 Edutainment

The ambition of an edutainment is to transmit knowledge or training by a ludic approach. The game “Auto junior” from the French multimedia magazine “Mobiclic” n°6 of October 1998, (editions Milan-Presse interactive) (playable on the website www.ja-games.com), invites the user to drive a car. The objective is to reach an open air cinema while respecting the Highway Code and being careful about speed. The game thus proposes a random series of tests (avoid an elk which crosses the road, not to cross a solid white line, stop at the halt sign…) which insist on a rule to respect. Each mistake is given an explanation and punishes the player by drawing points away from his driving license. The faster the player will drive, the more he will be exposed to the traffic accidents. We are facing a game whose scenario is made to give an educational message: to drive prudently by paying attention to the speed and to respect the Highway Code. This game is classified in the category of edutainment products.

This game’s production and realization constraints require to find an equilibrium between the “educative” and the “ludic” components. The game aspect can easily get the upper hand hiding all educative or informative aspect. In the same way, the too strong formative aspect brings the product closer to a quizz. The users are not taken in and they reject the product (Kellner, 2006)



Rejoins le cinéma en respectant le code de la route.

Attention, tu as un permis à 3 pts !

Figure 1: Auto Junior (Editions Milan/Ja.Games – 1998)

In the line of this paradigm, the MIT and the University of the Wisconsin joined to develop a research program named "Education Arcade" (http://www.educationarcade.org). The two terms

[1] IRIT, Université Toulouse III & LARA, Université Toulouse II, France ; daminous@gmail.com
[2] IRIT, Université Toulouse III & LARA, Université Toulouse II, France ; alvarez@irit.fr
[3] IRIT, Université Toulouse III, France ; jessel@irit.fr
[4] LARA, Université Toulouse II, France ; methel@univ-tlse2.fr
[5] LARA, Université Toulouse II, France ; pierre.molinier@univ-tlse2.fr

"Education" and "Arcade" are put here together to emphasize the idea to conceive education systems built on great ludic principles.

1.2 Advergaming

"Ponkey Bong" from the website www.spirou.com, presents two characters, Parker and Badger, created by Cuadrado and published by Dupuis Editions. In this video game, the player controls Parker and has to deliver his friend Badger. This one is attached on a rocket ready to take off! An angry site foreman, who looks like a gorilla, located at the top of a metal structure, throws barrels which roll along the various scales (fig.1). The gameplay of this game parodies "Donkey Kong" imagined by Shigeru Miyamoto (Nintendo) created in 1981 (fig.2). The objective of "Ponkey Bong" is here to transform a game into a tool of communication: to make the children play with the two characters of comic strips. This type of Serious Game, called "advergaming", is based on the "ludic culture" of the players. The idea is to release them from the training of the game play so that they are focused on the peripheral elements. We are in the same situation as an add for children where peripheral elements become more important because the narrative structure is quickly taken in.

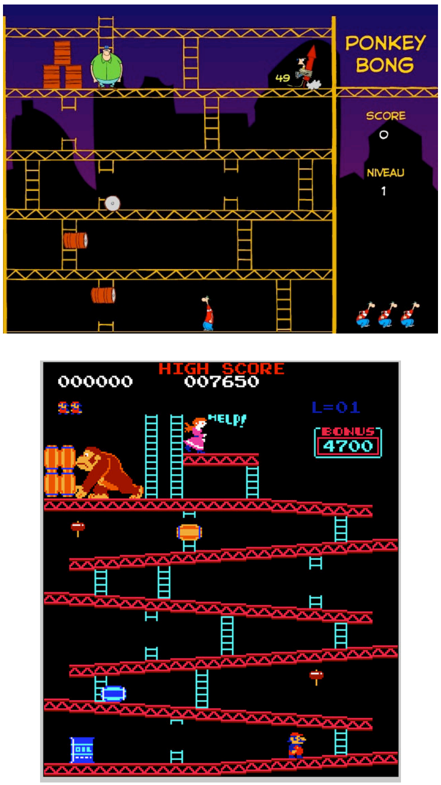
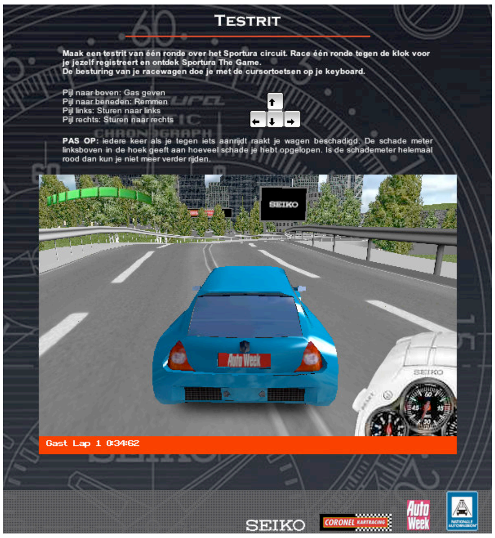


Figure 2: Ponkey Bong (Editions Dupuis/Ja.Games – 2002) and Donkey Kong (Nintendo/Miyamoto – 1981)

The video game "Sportura the game" http://www.sporturathegame.nl/public/testrit.php (Nonoche.com, 2004) plunges the user into a race car game. The goal is to be the fastest.

Brougere, in "Jouer/Apprendre" defines ludic culture as « a combination of procedures which make game possible" (p 106). He writes about a "personal ludic heritage […]: young adults remain marked, for some of them, by videogame which belongs to their culture, their story. They discovered it during childhood, but many of them kept it in their personal ludic heritage" (p 113). Brougere evokes the young adults audience but "that can be applied to all the players socialized through videogames practising and who would share perception and action habits coming from common ludic paradigms" (p 8)[1]



TESTRIT

Maak een testrit van één ronde over het Sportura circuit. Race één ronde tegen de klok voor je jezelf registreert en ontdek Sportura The Game.
De besturing van je racewagen doe je met de cursortoetsen op je keyboard.

Pijl naar boven: Gas geven
Pijl naar beneden: Remmen
Pijl links: Sturen naar links
Pijl rechts: Sturen naar rechts

PAS OP: iedere keer als je tegen iets aanrijdt raakt je wagen beschadigd. De schade meter linksboven in de hoek geeft aan hoeveel schade je hebt opgelopen. Is de schademeter helemaal rood dan kun je niet meer verder rijden.

Figure 3: Sportura the game (Nonoche, 2004)

The required reasoning is similar to a process largely used in the cinema, "the placement of products" (Galician, 2004). This term indicates the positioning of brands, logos or even products in the scenery of a videogame. In all the phases of play thus appears a Seiko watch and the road is strewn with posters pointing out this brand. The back number plate of the car is used to display the name of an automobile magazine. Lastly, on both sides of the game are posted the whole of the partners' logos which allowed the production of this title (fig.3). The exact term used by the communication agencies to indicate the placement of products in a videogames is "in-game advertising". This marketing concept can be pushed a little further and become interactive. In the MMORPG (Massively Multiplayer Online Role Playing Game) Everquest II, there is now an option to order true pizza pies to Pizza Hut Company online!

1.3 Edumarket games

This section gathers applications with an educational purpose, or at least applications aimed to make its users (especially children) sensitive to an educative message through video games. This different way of communication allows to change children's sensitivity, in order to help them having a better understanding of social stakes. For example, these social stakes can be durable development, school orientation, labour market, humanitarian aid... Edumarket games are tools aimed to communicate on a video game

[1] Personal translation by authors

basis while integrating an educational aspect.

For example, in this section we can find the game called Food Force (www.food-force.com), released by the United Nations in 2005, freely downloadable on Internet, with country-specific translations (Italy, France, Poland, China, Japan,...), and which is intended to make children sensitive to humanitarian missions made by the United Nations in their daily fight against starvation. On the website, we can find a special area for teachers, in order to help them building teaching lessons aimed to strengthen children's knowledge by complementary activities linked to the theme of this Serious Game.

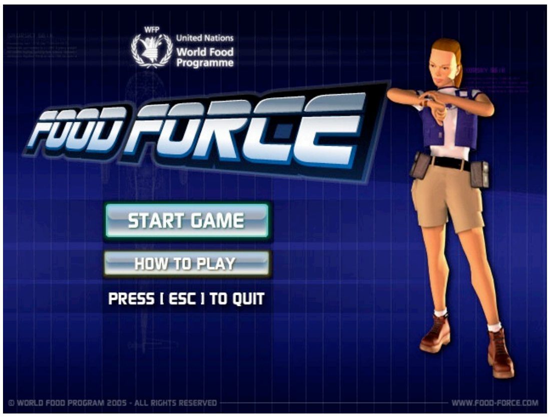


Figure 4: Food Force – Introduction

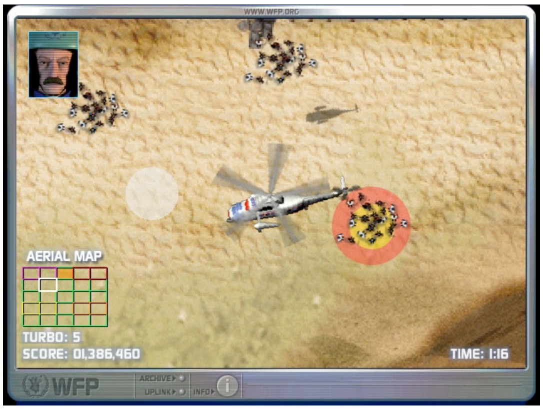


Figure 5: Food Force – Example of game

This title features six different mini-games, each representing a different aspect of the humanitarian aid, linked to a global objective: help a disaster victim area to recover. These games show the difficulties encountered by the different humanitarian workers. Each game is introduced and explained, including problems and game rules, by a 3D character seeming to come straight from a video game, such as Lara Croft.

When the mission is over, a short movie looking like a journalistic report shows real images of the tasks pictured in the game. When the global mission is over, the player can check his ranking on an online score table. The score table is of course intended to invite the player to improve his or her performance, but also helps to develop a reflexion about the community of players who devote themselves to "Food Force".

1.4 Political games

In the first level of the video game "Darfur is Dying" (http://www.darfurisdying.com), the user is a child from Darfour who must go and seek water for his family. On his way, he crosses dead animals and must avoid being captured by the militia (fig.6). The goal of this Serious Game is to denounce in a direct way the problems which currently strike Darfour. Gonzalo Frasca, a researcher at the Center for Computer Game Research of the IT University of Copenhagen, Denmark, calls this kind of video games "Political games".

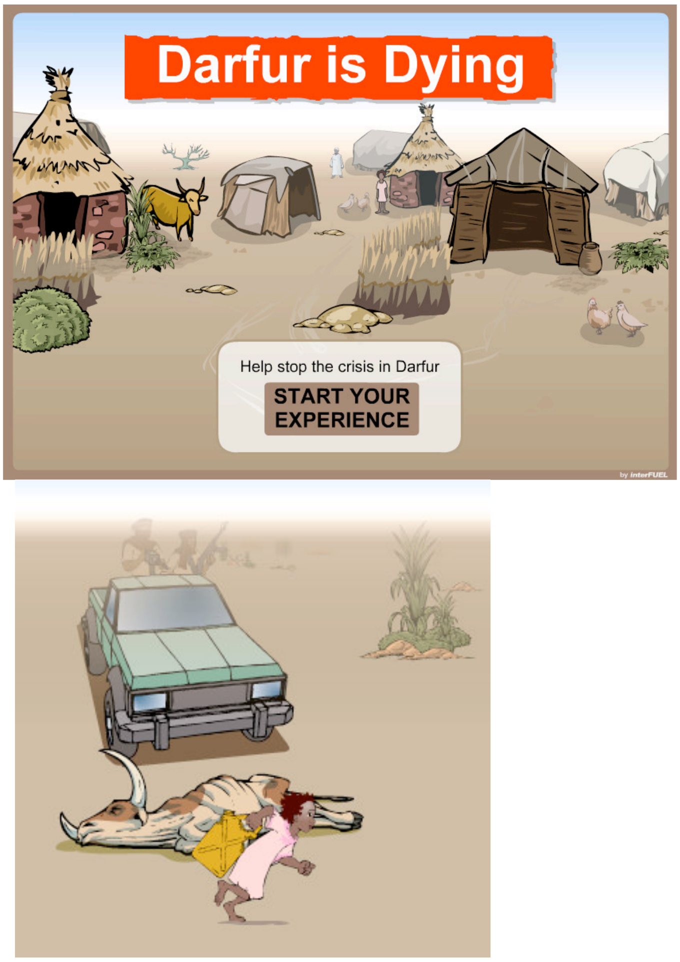


Figure 6: Darfur is Dying (MTV Networks On Campus Inc)

The line followed to carry out such plays consists in mobilizing in a diverted way the ludic mechanisms of the video game within a politically engaged situation. This diversion can be done on two levels:

- ✓ By modifying the rules of the game: For instance, "Antiwar Game" (http://www.antiwargame.org) prevents the player from winning if this one adopts the tactics which lead to the victory in a traditional videogame: to develop a powerful deterrent force, or to pile up many resources... Here on the contrary, these strategies lead on to the defeat or a state of stagnation. To make progress, military budgets will have to be replaced by social development in the end.

- ✓ By transforming the graphics and sounds of the game, following the example of advergaming. For example, the patch "Velvet-Strike" (http://www.opensorcery.net/velvet-strike) allows players to tag the walls of the Counter Strike FPS (First Personal Shoot), with pacifist graffiti.

These two aspects are not exclusive. There are patches, which not only modify the graphics or sounds of the game but also modify its rules. That's called Mods, abbreviation of Modifications. For instance "Escape from Woomera" (http://escapefromwoomera.com) is a Mod added to "Half-Life", a futuristic FPS (Sierra Studios/Valve Software), to transform it into a refugee camp called Woomera which really exists today and which is located in the south of Australia. The objective is to make the player sensitive to the problems of the asylum seekers in Australia and to take a critical look on the solutions applied by the government.

The website Sklunk which devotes a file to the diversion of the videogames (http://www.sklunk.net/Detournez-the-plays-video) indexes a whole of political games. It is striking to note that out of about fifteen games presented, eleven denounce violence or war. Knowing that many commercial titles mobilize this principle in the gameplay, it is also a militant act to want to modify the structure of it; we even think that it is a form of *reductio ad absurdum* and the provocation which encourage to act.

1.5 Training and simulation games

The most famous Games in this section are "Sim city", "The Sims" and "Flight Simulator". These applications allow the user to build and look after a virtual city, a virtual family, or to fly virtual planes based on real physical models.

The purpose here is not to win, but simply to have fun or to reach some "user-generated objectives", as Frasca explained in the second chapter of his thesis "Videogames of the oppressed: Videogames as a means for critical thinking and debate". He first reminds us that the Le Diberder Brothers define simulators as a virtual world, where attention to detail is a major feature, and with no clear objectives stated. The lack of objectives allows the user to switch as he wants from a playing purpose, called "paidea" (according to Roger Caillois's taxonomy) to a gaming purpose with precise rules, named "ludus".

Figure 7: Sim City 4 (Maxis/EA)

Frasca takes the example of "Flight Simulator" in which no precise objectives are stated. The player can enjoy "free-flight" (paidea) or decide to reach an imaginary aim such as flying under a bridge without crashing himself (ludus). Frasca concludes with the following: "The designer might suggest a set of rules, but the player has always the final decision."

Figure 8: The Sims 2 (Maxis/EA)

Microsoft Flight Simulator 2004 - A Century of Flight
Flights Aircraft World Options Views Help
SIMULATION PAUSED - Press P to continue.

Figure 9: Flight Simulator 2004 (Microsoft)

2 JUST A QUESTION OF POSTURE?

2.1 Reduction of the number of Serious Games' categories

In the first part, we have identified five categories of Serious Games: Edutainment, Advertainment, Edumarket game, Political games and Training and simulation game. When we analyse the nature of the first four categories, we realise that the method used to conceive them always consists in diverting, not in an exclusive manner, either the rules or the "cosmetics" as Chris Crawford says (graphics and sounds) of the video games. We also notice that these four categories share the same purpose that consists in delivering a message. Finally, it seems that it's only the very nature of the message that makes the difference between these first 4 categories. At a formal point of view we are thus in front of the same collection and the target is to deliver a didactic message or information. Only the latest category of "Training and simulation games" seems to be distinguished by relying exclusively on simulations which are cut out to pass down a knowledge first of all, leaving the player free to choose the way he wants to proceed.

It is also important to notice that simulation games just as the other categories of Serious Games have a system of values. The psychiatrist and doctor, Director of the Marmottant Hospital in Paris, Marc Valleur denounces the Sims as having consumerist values from North America. The richer one player is, the more friends he has. Actually, being wealthy make the social activities as well as the relationships easier between the actors in the game. But, Will Wright, the author of the Sims has made a place for money like Molière in The Miser. Money is a part of our Western Society and has its own function. It makes relationship "smoother" between people (Kauffman). It thus makes the exchange easier, even though it "decreases and simplifies" the very nature of the relationships. Starting from this analysis, the questions show that a simulation could also be a support for the distribution of a message. .

2.2 The message diffused by a simulation game

For Frasca, in Sim City, a simulation videogame, the user builds his own rules and objectives. For instance, to develop the largest, the smallest or the richest city but also to set fun challenges like deciding to make the most aesthetic city. However, we remain here exclusively within the framework of the game. For Genvo, to play is also a choice of posture that the user adopts. Indeed, by using Sim City, a trainer fixes the objectives in adequacy with a teaching progression, the player adopting a posture of learning, according to the context defined in the set objectives: for instance, to understand and to analyse the reactions of a population if the city does not have any shopping centre, or to observe the impact of road infrastructures ill adapted to the economic development of the city.

Thus, it is very simple for a user to switch from the paidea to the ludus, but also from a ludic posture to a didactic posture with a simulation. As Brougere explains to us in "Jouer/Apprendre" by using the concept of "frame" developed by Goffman (p.45), to adopt a choice of posture depends on the context within which the use is (home, school, institution...), if the user is alone or not. All of these notions are also mentioned by Katie Salen and Eric Zimmerman and regrouped in one of their three “primary schemas” named “Culture” (p.102 to 105).

If simulation can take an educational function, it also can take an advergaming function. For that the game designer just has to introduce advertising posters or commercial products into Sim City. To introduce video reports on the trades of town planner, architect, mayor to each annual balance sheet for example would make it possible to bring an Edumarket game dimension to Sim City... Lastly, for the political aspect the game designer just have to add tags or political posters on the walls or to introduce situations of play around poverty (Homelessness, impoverishment, excessive debt). The incidence of the user’s political choices makes it possible to insufflate some not disguised criticisms on the policy of urbanization and economic development currently carried out by the rich countries. Board games like “Tiers Mondopoly” (Orcades Editions) come from the same reflection.

Consequently, we can deduce that a simulation can diffuse all types of messages and objectives like video game does, according to the posture that the user chooses to adopt and to the ingredients (rules and design) which the game designer decides to introduce in the “world”.

2.3 Can the video games permit to train like simulations?

We have just seen that simulation can diffuse a message as well as the first four categories of Serious Games founded on videogames. At this step the added value of simulation would be, if compared to the video game, to offer a training to the user. This thus leads us to know if the video games can do the same.

The answer is obviously related to the posture that the user decides to adopt with his video game. If the video game is essentially an invitation with ludic, Michael Stora in his book "Guérir par le virtuel", explains to us how he uses video games as a therapeutic tool to cure a child’s behavioural troubles. It is here necessary to insist on the place that the adult occupies within the relation which is established between the child and the video game: He is engaged in order to modify the intention and the posture of the child player. In the same way, Shawn Williams tells us in his article « Learning the gaming way » (The Escapist, n° 59), how video game is used daily by his wife, who has a degenerative disease, to preserve her health. The video game thus offers the same properties as simulation.

Thus, we can conclude that Serious Games are composed of two main categories defined as follows:

- ✓ First Serious Games, based on simulation which present a “world”, with its “rules” and where there is no objective imposed by the application.
- ✓ Second, Serious Games, based on video games which propose a “world”, with its “rules” and implemented objectives that the user has to reach.

To diffuse a message and to let the user the choice to adopt ludic, didactic or training posture are possible with the two categories that we have identified, the fields of application being similar.

2.4 To implement objectives, is it an added value to spread a message?

We have just identified in 2.C. that the difference between the two main categories of Serious Games lay only in the presence or not of objectives implemented in the application. Now, the question is to know if the presence of objectives laid down within an application constitutes an added value to spread a message or not.

An experiment carried out in September 2006, in collaboration with the Vortex team of the Toulouse Institute of Computer Search (IRIT) makes it possible to lay down some orientations for future research. Within the framework of the centenary celebration of the discovery of Garges’ cave, three multimedia devices were set up. The idea was to present to the public, through this numerical process, the inaccessible places or restricted areas in order to preserve the cave.

The first device is a simulation which invites the user to locate and raise the layout of various animals on the wall of the cave. The device is composed of a multimedia table on which a video is projected representing the wall of the cave where engravings illustrating the animals are tangled up. The user, thanks to a light pen, draws the contour of some animals which he has to locate first. To accompany him, an organizer guides his browsing and gives explanations (fig.10).

The second one is a traditional computer connected to a video projector which presents a simulation in three dimensions of the hands’ sanctuary. The user can look at each recess thanks to a spherical panoramic that he can move with a mouse. Here, an organizer is present too, to explain the vocation of the numerical set and to comment on the pictures (fig. 11).

The third one is a multimedia video game whose goal is to invite the player, in less than 3 minutes, to locate and draw with a mouse one animal’s contour on the same wall of the cave that is presented in the first numerical set. The effigy of the animal is permanently presented on screen. Here there is no organizer in charge of explaining the contents and the rule of the game (fig. 12). However, when an organizer was present, the users only questioned this one about how to play.

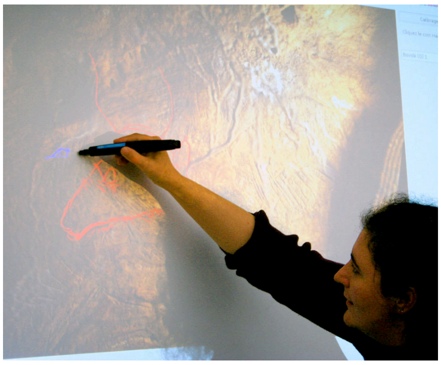


Figure 10: Set on multimedia table

During the day, we interviewed three twelve-year old children, having used the three numerical sets, in order to collect their perceptions and their feelings. Concerning the first device, the spectacular dimension, to draw with a light pen, is arisen in an obvious way. Concerning the reception of information, the children are able to enumerate the animals that they had recalled. The children were fascinated by this imaginary and futuristic activity. The technology generated by itself a ludic and emotional dimension which resulted in a gathering around the set. Even some seniors have approached chairs. The performance of the volunteers who came to draw was a true show for them (fig. 13). The second set was mainly described by the explanations given by the organizer. The children explain the vocation of the device and the nature of the pictures displayed. The global intention and the organizer's remarks are well restored. As for the multimedia video game which represented the third set, the children described it only with the ludic challenge which it proposed: "In this game, you have to recall the animal's shape before it is too late!" The children neither evoked the name of the animals that they had to draw nor formulated questions or comments about the difficulties that the scientists had to face when they listed all the shapes on the real walls.

Figure 11: Organizer presenting the device of the "Hands 'sanctuary" displayed on traditional screen and video projector.

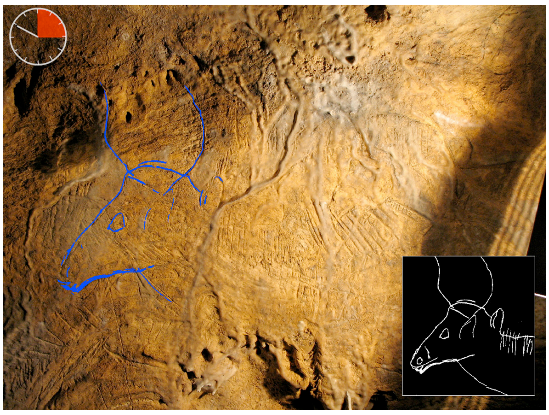


Figure 12: The video game "Gargas" cave" (Vortex/Ja.games – September 2006).
http://dreampict.free.fr/Gargas/Gargas3.swf

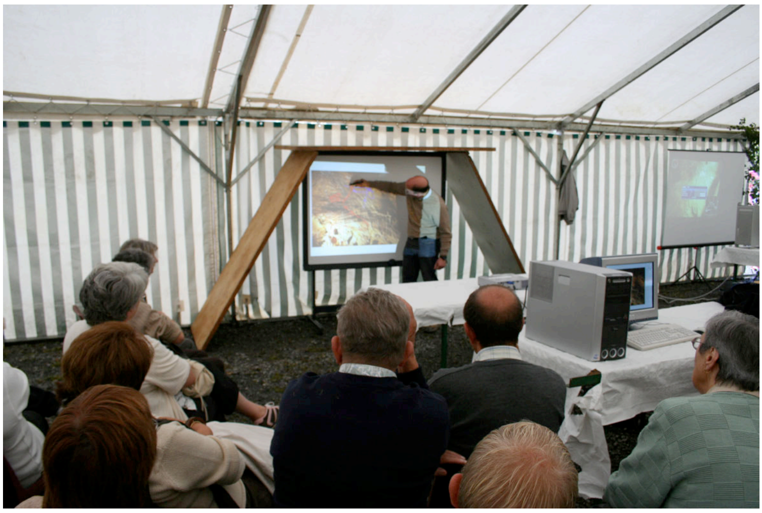


Figure 13: For the first computer set, the seniors sitting attend the performance of the volunteers who drew on the multimedia table as if it was a show.

These three devices highlight that the simulation accompanied by an organizer more often invites the user to adopt at first a didactic posture. Conversely the game, especially without an organizer, naturally invites the user to adopt a ludic posture. In this context, according to the way in which an organizer or a teacher wishes to diffuse his message, it directly influences the users's posture. Thus the mediatized activity takes a different experiential dimension.

A short term memorizing is effective in the experiments one and two, which lets us think that the presence of a mediator reinforces the potential trainings around Serious Game. This assumption is under evaluation in our current search on Edumarket Games.

2.5 Beyond posture, some marketing aims to take into account

The challenge for the industrialists or the institutionalists who wish to use Serious Games as marketing strategy or communication tools, is to manage to offer products which take into account a child's educative and also playful environment. The objective is then to manage to develop products corresponding to the cultural referents of the aimed market. To reach such a goal, it is necessary to go beyond the mere integration of its brand in the existing game play. A heavy adaptation of the original concept can't be avoided because of a different sociocultural context. This process implies a complete transformation of the product by teams understanding as well the cultural stakes as the technological challenges linked to the game and to the brand. It is the only way for the industrialists to be able to settle on markets on a long-term basis and to avoid emergent resistances from consumers who are more and more aware and critical about new advertising strategies.

The implication of both the educative relation and the pleasure dimension corresponds to this wish to build a clear understandable message. This specificity reinforces the idea that the marketing action's main line lies in the experiential dimension of consumption (Hetzel, 2002) and of use (Kline, Duer-Witheford. and De Peuter, 2003). Pleasure of telling and acting, confrontation to challenges, interactivity and narrative liberty are communication lines widely mobilized and prepared in advertising campaigns using serious games.

However there are limits in this search for efficiency and result in a communication policy. The attitude towards the brand aspect takes us back in a wider way to the consumer's perception aspect. The individual mustn't be trapped in market logic at the risk of creating forms of resistances. It is then necessary to build well-balanced plurimedia strategies that respect one of the major stakes of society today: provide the consumer with the "keys" of consumption practice and help him to understand things behind offer and the consequences of his choices, that is to say educate the individual to consumption. This doesn't mean to inculcate him in unquestionable behaviour ways (such product rather than such other) but rather to help him to build a common reference of skills, that is to strengthen the resources that can be called up when he meets the product and its graphics. Thus Serious Game as a communication tool has an importance in diverting and educative principles, even if the posture choice remains unknown for the user in the end.

CONCLUSION

The reflexion in this article has allowed us to discover that there are two kinds of Serious Games: Those based on the Video Game proposing a target that the player has to obtain, and those based on simulation without a special aim. This fact leads us to consider that the different categories of Serious Games being indexed up till today don't find their foundation within a formal constitution, but are a part of a choice of position that the game tries to transmit to the player, by representing "a world" governed by rules as well as graphics and sounds in accordance; The player is always the one that decides about the position to adopt about using the Serious Game.

In order to get to know whether it is better to use an application with available aims, we consider for the moment that the player will at first appreciate to play if the targets are implemented but if they are not, he will get a didactic or training posture. The impact of the distribution of the message is probably depending on how this way is used at the beginning by the game designer or the mediator.

At last we have seen that above all the question of position has to be taken into account when you will construct a strategy of communication with Serious Games. This implies to give "keys" to the user to teach him how to apprehend a Serious Game better over time and to discover its performances.

In that way a Serious Game is a fundamental challenge within modern societies because it reveals ideological models that are hidden and it shows the ambitions of society. This dimension also asks the question about the responsibility of the creators of games because the activity is significant and has a lot of meanings.

REFERENCES

[1] Bura S., A Game Grammar, http://users.skynet.be/bura/diagrams/

[2] Brougère G., Jouer/Apprendre, Economica/Anthropos, 2005

[3] Buckingham D. et Scarlon M. Ed., Education, Entertainment and learning in the home, Open University Press 2003

[4] Caillois R., Man, Play and Games, Free Press, 1961

[5] Crawford, C., On Game Design, New Riders, 2003

[6] Galician, M. L., Handbook of product placement in the mass media, The Haworth Press, 2004

[7] Frasca G., Videogames of the oppressed: Videogames as a means for critical thinking and debate, 1999

[8] Kline S., Duer-Witheford N. and De Peuter G. Digital Play: the Interaction of technology, culture and marketing, MsGill-Queen's University Press, 2003

[9] Koster R., A Grammar of Gameplay, http://theoryoffun.com/grammar/gdc2005.htm

[10] La Ville (de) I., L'enfant consommateur, Vuibert, 2005

[11] Le Diberder A. and F., Le Diberder F., L'univers des jeux vidéo, La Découverte, 1998

[12] Letourneux, M., in Le game design de jeux vidéo, Genvo S., L'Harmattan, 2005

[13] Michael D and Chen S., Serious Games:Games that Educate, Train, and Inform, Course Technology PTR, 2005

[14] Salen K. and Zimmerman E, The Rules of Play, MIT Press 2003

[15] Thielscher M., Game Description Language, http://games.stanford.edu/language.html

[16] Zyda M., From Visual Simulation to Virtual Reality to Games, IEEE Computer Society, 2005

Educational Games: Overview of Shortcomings and Proposed Solutions

Rania Hodhod

Abstract. Educational computer-based games (*edugames*) are games that promote the acquisition of skills and knowledge in a pleasant interactive way. It is well known that not all the users share the same preferences or styles when interacting with a game and solving game-problems. This leads to the importance of adaptation in the sense that behavior of each play-instance of a game depends on the actions of an individual user/player. The major aim for an adaptive game-based learning system is to support and encourage the learner/player/user by considering his needs, strengths and weaknesses. However, the lack of a common design vocabulary has considerably slowed the progress of edugame design.

For this research proposal, we propose to develop a design/methodology for adaptive educational games and to evaluate it empirically by implementing an edugame prototype to practice prolog programming. Evaluation that addresses the new and main aspects in the developed design/methodology will be prominent at the end of the research.

1 INTRODUCTION

With rapid technology development in graphics, sound, and real-time video; electronic games have become increasingly more entertaining and enjoyable for kids as well as adults. Among the various kinds of games, there is a special category, educational games (*edugames*), which have one goal beyond solely entertainment and that is education.

Research in edugames has over time progressed via three separate stages. The first stage perceived the use of computer games as a direct way to change the behaviour of a user through repeated actions. The second stage put the spotlight on the relation between the computer game and the player. The latest stage now includes the context of computer games and how they facilitate learning environments.

Since the 1970's various educational games have emerged and some of them claimed to have educational effectiveness. However, very few formal evaluations [1] have been conducted to evaluate the actual pedagogical values of these games.

Taking into account that different personal interests, different knowledge status, and learning abilities will often lead to different playing patterns implies a factor that must
as will be shown below.
be taken into account in any evaluation of a game. This leads to the importance of a design/methodology on evaluation of adaptation in edugames.

The paper is organized as follows: The next section presents the various aspects and educational needs of games. Following this is a discussion on problems encountered in edugames and some solutions. After which the paper presents a brief introduction to different learning theories and an overview of existing edugames. The paper finally finishes with a research proposal and the conclusions so far reached.

2 GAME ASPECTS AND EDUCATIONAL NEEDS

Games are enticing problem solving environments which the player can explore at will, creating his own ideas of its underlying structure and synthesizing strategies which reflect his understanding of this structure. They are competitive interactions bounded by rules to achieve specified goals that depend on skill, and often involve chance and an imaginary setting [2].

Games have challenges, fantasy, abstract concepts and curiosity that engage the player's attention [6, 7, 8, 14]. To this is added other powerful characteristics such as virtual worlds. These virtual worlds are not just about facts and isolated skills, but embody particular social practices such as developing situated understanding, and experimenting with new and powerful identities [4, 5]. Moreover, games have the potential for motivating drill and practice by providing environments in which students actually enjoy repetition.

Noting the highly motivating nature of games and all the other constructive aspects games can provide, researchers have started to investigate whether these games could be utilized to assist learning [3].

Many (if not most) of the present edugames have not been designed based on any of the existing learning theories [8, 16, 18, 19, 21, 22, 23, 24, 25] but have been designed in an ad-hoc way. Only few designers claim that their games are really effective in education, and even fewer support these claims with results from formal empirical studies [1]. Some researchers such as Klawe [9] consider edugames effective only if the interaction is monitored and directed by teachers, or if the games are integrated with other more traditional activities like pencil-and-paper exercises. Other

[1] University of York, UK, email: rania.hodhod@cs.york.ac.uk

researchers believe that effectiveness of edugames is related to the features, preferences and behaviour of a particular user [3]. We argue that a design bearing the “individualized instruction” feature can be an efficient way to deal with personal differences.

3 PROBLEMS ENCOUNTERED IN EDUGAMES

Empirical studies have shown that one major problem is that while edugames are highly engaging, they often do not trigger the constructive reasoning necessary for learning. Two researchers [10, 8] have argued that students can be successful game players by learning superficial heuristics rather than by reasoning about the underlying domain knowledge; but the lack of a common design vocabulary presents problems in evaluation these claims. (In addition to which is the observation that the evaluation phase has not been a serious factor in present designs of edugames.)

In adaptive edugames more problems are presented such as the real-time adjustment of the background story (dependent on the user interaction), and the expansion of the user model which itself is a key element in the adaptation process as it includes not only the level of student knowledge but also his intentions. These issues (and others) are often missing due to the lack of awareness of existing learning theories; theories which themselves can serve as a template in the design process of edugames. Such awareness in a design of an edugame can lead to achieving higher learning levels implying better educational outcomes.

4 EDUGAMES AND LEARNING THEORIES

Many learning theories exist that edugames research area can utilize to achieve desired educational needs. According to research [11] those of Gagne’s events of instruction [13], Keller’s ARCS Motivational model [11], and Bloom’s taxonomy [15] are the most appealing templates to be used in game design principles, while Reigeluth’s Elaboration Theory can be also be optionally included [12].

• Gagne [13] has developed what is called “events of instruction” which serve as a guide for developing and delivering a unit or units of instruction. His described nine events are: Attention gaining, Objective setting, Invoking of prior learning, Presentation of new material, Created scaffolding, Provision of practice, Feedback, Assessment, and Retention-and-transfer of new knowledge to a real-life situation.

• According to Keller [11], motivation is a necessary but not sufficient condition needed to ensure that learners actually learn something. His ARCS model is represented using the four following classes: Attention, Confidence/challenge, Relevance and Satisfaction/success. In deeper detail, gaining attention is a learning prerequisite while relevance is about what is taught and how it is taught. Confidence is expectancy for success, and finally satisfaction is about how people feel about their accomplishments. Keller’s model is intended to be incorporated in accordance with instructional models like Gagne.

• Bloom [15] has identified six levels within the cognitive domain, from the simple recall or recognition of facts, at the lowest levels, through increasingly more complex and abstract mental levels to the highest orders which are classified as synthesis and evaluation. His theory is further discussed below.

• Reigeluth’s Elaboration Theory [12] proposes several major strategy components: An Elaborative Sequence where good games follow a well-paced sequence progressing from simple (and easy) to complex (and hard). Learning rote sequences is the involving of simplified problems as well as providing suggestions. Summary is something that almost all games provide in the some form of statistics/percentages (e.g., score, health, strength, maps, assets, etc.). Synthesis is building on knowledge gained from previous knowledge. In Analogies players very quickly learn to look for approaches or tactics that are similar to some other game they have played, and will try to apply these in any new context that looks like it might favour this approach. The idea of Cognitive Strategies is the ability to force the player to use strategies invented by the designers in order to achieve goals. Learner Control is the idea that a player/learner is always in control is an obvious requirement for all games since without it a game becomes a non-interactive computer program.

In common to all the above approaches is the need to measure the learning outcomes of edugames. The higher the learning level achieved, the better the edugame learning outcome. As seen in the above mentioned theories, the various components and attributes are shared like: attention gaining, feedback, motivation, relevance, success, summary, cognitive strategies, etc. These concepts should be kept in mind throughout the design and implementation of any edugame.

However, Bloom’s classification of the learning levels can serve as a measurement of the learning outcomes of edugames. The next section introduces the existing edugames briefly and in a way that points out the most important aspects and weak points found. In addition a measurement will be assigned to their learning outcomes according to Bloom’s taxonomy.

5 EXISTING EDUGAMES

Mapping the learning outcomes to Bloom’s learning levels requires first to identify exactly what each level in the taxonomy means, so that a gauge can be calibrated guided by these definitions. Bloom’s taxonomy of learning levels can be defined as follows:

- Knowledge is defined as the remembering of previously learned material.
- Comprehension is defined as the ability to grasp the meaning of material.
- Application refers to the ability to use learned material in new and concrete situations.
- Analysis refers to the ability to break down material into its component parts so that its organizational structure may be understood.
- Synthesis refers to the ability to put parts together to form a new whole.

- Evaluation is concerned with the ability to judge the value of material for a given purpose.

An early educational game, such as *How the West Was Won* [16] was developed in 1976 to teach mathematical expressions. It has an embedded user model that leads the student through the game while identifying the student's weak points. Another edugame developed at this time (1977) to teach logic and probability is the *Wumpus* game [17]. Wumpus has an embedded user model to identify the player's logical problems. Both edugames reach the Application Level.

The embedding of agent technique with user modelling can be seen in the edugame *Easy Math* [18] (developed in 2000). This embedded user model helps in identifying the misconceptions of individual students. Although this edugame has a puzzle game as one of its exercises, it lacks many of the game features which affect its success as an edugame. This edugames reaches the Knowledge Level.

The *Aqua Moose* edugame [19] (developed in 2002) to teach mathematical functions through visualization. This edugame proves that a fantasy story line or a good interpreted background story can have priority over graphical issues in the edugame environment. However lack of a user model prevents the edugame from tracking the player performance. This game reaches the Comprehension Level.

Prime climb edugame [10] (also developed in 2002) to teach number factorization. This edugame also shows the importance of having a well structured story line to engage the student. If such a story line is absent, the player will be distracted from the main purpose of the game by trying to find other joyful objects in the playing environment in front of him. This edugame reaches the Application Level.

In a problem solving environment like *Betty's Brain* [20] (developed in 2005), researchers believe in the learning-by-teaching paradigm. This game tries to reach the higher levels in Bloom's taxonomy (Analysis and Synthesis), but it fails in helping the players to attain this.

JVM edugame [21] (developed in 2004) to teach the compilation process of Java language with the help of an agent embedded in the game environment. Players are immersed in micro-worlds, not learning any particular domain but becoming part of the environment. This game illustrates that long, traditionally tedious, and difficult tasks can be engaging and fun when they are part of a good story. This game reaches the Analysis Level.

The *Lincoln* edugame [22] (developed in 2006) proves the effectiveness of taking over the role of the virtual character in a game as a good way of involving and engaging the student. Although this game can be considered one of the good games to teach history, it lacks the presence of a user model that targets individual preferences. This edugame reaches the Analysis Level.

Some attempts to teach computer programming concepts include *RoboCode* [23], *ToonTalk* [24] and *CeeBot-4* [25]. *RoboCode* is a Java-based virtual robot game intended to teach Java programming techniques. The programmers implement their robots in the Java programming language, and test their creations either: using a graphical environment in which battles are held, or by submitting them to a central web site where online tournaments regularly take place. *ToonTalk* is a game to teach programming concepts but without the writing of source code. *CeeBot-4* is a game to learn programming, or to teach programming at middle school, high school and university. It uses a language close to Java and C# to program robots that will solve various tasks ranging from finding the way out of a labyrinth over car racing to playing soccer.

RoboCode and *CeeBot-4* lack a pedagogical agent while *ToonTalk* uses agents to provide hints and help but without making use of any user model. These games are examples of using entertaining goals to motivate students to practice perceived dreary activities like programming. *ToonTalk* reaches the Synthesis Level, while *RoboCode* and *Ceebot-4* reach the Application Level.

During a recent literature survey/review many issues were noticed. Among these issues, adaptation has not been achieved through adapting the game environment itself to be contingent with the educational needs of the player as dictated by the user model and its state in the game environment. For example in the edugame [8] the objects and obstacles on the same level were fixed for all users unless changed by externally by someone (say, the teacher); otherwise the ability to adaptive in this edugame is only through the style of help and hints provided to the user by the pedagogical agent. Likewise, in the edugame [11] adaptation is acquired through the idea of the presence of various sub-games that are assigned to different users according to their profile.

Another issue noted was that none of the existing edugames that contains a user model has dealt with the *mental state bandwidth* where bandwidth is a parameter for categorizing student models. User input gives an indication of both the knowledge and intentions underlying a user action. Making use of these indicators can help in the adaptation process. Lastly, it was noted that that the highest learning levels in Bloom's taxonomy have not yet been reached any of these edugames surveyed.

6 PROPOSED RESEARCH

6.1 The proposed model

As mentioned above the idea of adapting the edugame environment according to the users needs in a dynamical fashion during the playing of a game not been investigated. Therefore, we argue that tackling this issue can be achieved through our proposed research. Figure 1 shows the proposed model where the interactions between the story engine, the educational material and the user model are identified.

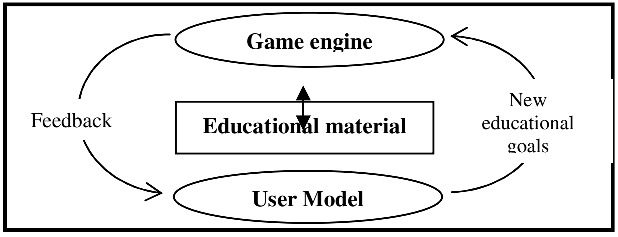


Figure1. The proposed model

By understanding the relationship between the educational needs and the game elements can allow development of edugames to include visualization and problem solving skills [7]. This idea can be extended using the model presented in figure 1 to incorporate the dynamic generation of the game elements that are associated with educational goals.

The proposed model incorporates the notion of direct interaction between the story engine and the educational material; while the game engine provides feedback to the user model which in turn provides new educational goals to the engine. The task of the game engine is to generate game objects associated with specific educational task dynamically during the playing of the game. Such generation is in accordance to the information dictated from the user model and the educational material, and it was not specified by the edugame designer beforehand but it is achieved according to some generally coded association rules. The representation of such rules is an area that itself be independently researched.

The proposed model considers two important issues: the first issue is the contraction/expansion of the user knowledge over time and the second issue is the perseverance of engagement and fun during play time (or learning time). The first issue is considered important since game objects are always generated according to constantly updating user model. This means that if a contraction/improvement is noticed in the level of the player's (student) knowledge then the appropriate object associated with the appropriate educational material will be presented or retracted. In this way the level of difficulty of the game is adjusted to the player preventing him from being frustrated by finding the game too difficult or getting bored by finding it too easy.

The second issue is also considered important due to the fact that the educational material is integrated as a part of the game story itself and the success of learning this material leads ultimately to the success completion of the game. In turn this helps in maintaining a fairly constant level of engagement with the edugame. We believe that the outcome of this research is a model that can lead to a deeper understanding of the adaptation process which then in turn leads a better design of edugames with higher educational outcomes.

6.2 Proposed design methodology

The previous subsection discussed some of the shortcomings in the field and proposed ideas to rectify them. In this section a design methodology that incorporates these ideas is presented.

This design methodology has the following characteristics:

- The design must be based on a learning theory.
- The educational aim must be considered within the game design from the very beginning and in every step through the design process.
- The educational material has to be integrated with the story line and be part of the edugame environment.
- Enrich the learning opportunities for users by offering intellectual exploration through individualized user guidance and support to resolve the user's misconceptions within the learning environment.
- Reaching the higher learning levels of Bloom's taxonomy must be achieved as an outcome.
- Educational material, student and tutoring models should be incorporated in the game. The student model should incorporate student goals as well maintaining an idea of the student's knowledge.

The proposed methodology gives the user/player/student the chance to be exposed to higher learning levels. While this can be achieved through the drill and practice puzzles embedded in the edugame environment, pace of game play can be reduced/increase through dynamically varying the difficulty of puzzles, reducing the number of tasks to be performed if the concept has already been mastered; de/increasing the number of interactive characters, or even simply changing the player/characters inventory [26]. In addition, the proposed design recommends dealing with the mental state bandwidth in the student model, where the student model has to incorporate the student goals along with his educational knowledge. We believe this can also help in guiding the adaptation process so leading to better educational outcomes. Finally, a battery of hints and feedback should be designed within the edugame environment as necessary components of tutoring [27, 28].

6.3 Proposed scenario

The proposed methodology/design will be demonstrated through the implementation of an edugame to practice Prolog language programming. Given this short scenario it can be seen how the proposed model can work in an edugame environment.

Assume that the player/student is situated in the hallway of a house and is presented with a problem to solve. The system can capture the level of knowledge and the student intentions from the answer(s) he will give. The student feedback provides information about his knowledge level and how he provides his answer provides information about his intentions. For example, if the system now believes that the student executes certain rules, *rule1* and *rule2*, in a certain order to entail the goal *g*. This can be added to the student model as an indicator of what the student believes and what are his intentions are during the solving this kind of problem.

Now assume that the next task presented to the player is to write a program to deduce a secret number. It is now the job of the game story engine to decide what is the next appropriate object to present to the user. A method to reason about this can be as follows: As the user is indoors, it will not be suitable to present a tree object to introduce the new task, while as the task is to deduce a secret number, an object like a safe is more suitable than a ball. Hence reasoning about the environment together with the educational material plus the user knowledge state is main task of the edugame engine. The engine also has to consider all these issues in order to present the player with the suitable object that better serves the educational task and keeps the fun and engagement in the edugame.

7 EVALUATION OF PROPOSED WORK

The evaluation of the prolog programming edugame prototype that demonstrates the proposed design /methodology will be done in two stages. The first stage will assess the design methodology through an internal evaluation whereby a clear picture of the architecture of the intelligent tutoring facility and how this kind architecture provides the edugame environment is shown.

In the second stage, an external evaluation will take place in which the educational impact of the edugame on the player and how the edugame helps the player to improve his knowledge and skills will be measured. The measurement suggested is a cognitive walkthrough and a heuristic evaluation of what has been learnt. The first stage gives the chance for researchers to take on the role of the users and so identifies potential usability problems. The second stage evaluates the user interface and indicates potential problems that violate the general principles of good design interface. Further to this is the logging of game play which is helpful in understanding how the edugame is played. Finally evaluation through focus groups and pre/post tests will give a measurement of what has/has-not be learnt.

8 CONCLUSIONS

Educational games must be at least as effective as the teaching methods they replace. Therefore the fundamental goal of educational games must be: the player must master the content of the educational material in order to master the game. In other words, success in the game must be dependent on learning skills and/or concepts. In addition there is natural tension in game design between the complexity of rules and the simplicity of interfaces. Player choices and feedback from these choices should be transparent enough to foster freedom, immersion, and flow of movement in virtual worlds without overwhelming the player with information and/or commands. To this we argue that it is important to consider the learning theories during the edugame design and evaluation.

Further, we believe that the proposed ideas in this paper can help in overcoming some of the shortcomings and drawbacks that currently exists in the edugames research field, by noting that the proposed methodology /design/model offers a kind of equilibrium between achieving the desired educational needs and a constant level of fun and engagement during the learning process associated with game play. In addition the capability of the proposed edugame manages not only the player/student knowledge but also his intentions leading to a deeper understanding of the adaptation process which we argue leads to better educational outcomes.

ACKNOWLEDGEMENTS

I would like to thank the AISB'07 Symposium: AI and Narrative Games for Education reviewers for their comments, which helped improving this paper. Also, I am gratefully acknowledges Nim Shah's and Lillian Clark's comments that helped in improving the language of this paper.

REFERENCES

[1] J. M. Randel, B.A. Morris, C.D. Wetzel, and B.V. Whitehill. *The effectiveness of games for educational purposes: A review of recent research.* Simulation & Gaming 23, 3, p. 261-276, 1992.

[2] D. R. Cruickshank, and R. Teller. *Classroom games and simulations*, Theory into Practice, 19(1), p. 75-80, 1980.

[3] J. L. McGrenere, *Design: Educational Electronic Multi-Player Games ALiterature Review*, Dept. Of computer Science, the University of British Columbia, 1996.

[4] D. W. Shaffer, K. D. Squire, R. Halverson, & J. P. Gee. *Video Games and the Future of Learning. Phi Delta Kappan,* 87(2), p. 104-111, 2005. Available at: *http://coweb.wcer.wisc.edu/cv/papers/videogamesfuturelearning_pdk_2005.pdf*

[5] *Intelligent Tutoring Systems: An Historic Review in the Context of the Development of Artificial Intelligence and Educational Psychology.* Available at: http://www.cse.msu.edu/rgroups/cse101/ITS/its.htm#_Toc355707493

[6] J. Tan, C. Beers, R. Gupta, and G. Biswas, *Computer Games as Intelligent Learning Environments: A River Ecosystem Adventure*. Artificial Intelligence in Education, C.K. Looi et al. (Eds.), IOS Press, 2005.

[7] A. Amory, K. Naicker, J. Vincent and C. Adams, *The use of Computer Games as an Educational Tool: 1. Identification of Appropriate Game Types and Game Elements*, British Journal of Educational Technology 30(4) p. 311-322, 1999.

[8] K. Stacey, E. Sonenberg, A. Nicholson, T. Boneh, V. Steinle. *Modelling teaching for conceptual change using a Bayesian Network.* Sumitted to Int. Journal of AI in Education, 2001.

[9] M. Klawe, *When Does the Use Of Computer Games And Other Interactive Multimedia Software Help Students Learn Mathematics?*, Department of Computer Science, the University of British Columbia, 1998.

[10] C. Conati, *Probalistic assesment of user's emotions in educational games*, International Journal of Human-Computer Studies, 59, p. 213-225, 2002.

[11] G. A. Gunter, R. F. Kenny, E. H. Vick, *A case for a formal design paradigm for serious games.* Available at: *http://www.units.muohio.edu/codeconference/papers/papers/ Gunter%20Kenny%20Vick%20paper.pdf*

[12] C. M. Reigeluth, M. D. Merrill, B. G. Wilson & R. T. Spiller. *The elaboration theory of instruction: A model for sequencing and synthesizing instruction.* Instructional Science, 9(3), p. 195-219, 1980.

[13] K. Kruse, *Gagne's Nine Events of Instruction: An Introduction*, 2006. Available at: *http://www.e-learningguru.com/articles/art3_3.htm*

[14] M. Prensky, *Digital Game-Based Learning.* Blacklick, OH, USA: McGraw-Hill Professional, 2000. p. 106. Available at: *http://site.ebrary.com/lib/york/Doc?id=10043861&ppg=124*

[15] http://www.officeport.com/edu/blooms.htm

[16] R. B. Richard, J. S. Brown., *A tutoring and Student Modelling Paradigm for Gaming Environments*. ACM SIGCUE Outlook, ACM SIGCUE Bulletin, Proceedings of the ACM SIGCSE-SIGCUE technical symposium on computer science and education , vol.10, 8 Issue SI, 1, February 1976.

[17] I. Goldstein, B. Carr., *The Computer as Coach:As Asthletic paradigm for intellectual education*. Proceedings of 1977 annual conference, Pub: ACM Presss, January, 1977.

[18] M. Virou, V. Tsiriga, *Involving Effectively teachers and Students in the life cycle of an Intelligent Tutoring System*. Educational Technology & Society 3(3). ISSN 1436-4522, 2000.

[19] J. Elliott, L.Adams, and A. Bruckman. *No Magic Bullet: 3D Video Games in Education*. Proceedings of ICLS 2002, International Conference of the Learning Sciences, Seattle, WA, October 23-26, 2002.

[20] J. Tan, C. Beers, R. Gupta, and G. Biswas, *Computer Games as Intelligent Learning Environments: A River Ecosystem Adventure*. Artificial Intelligence in Education, C.-K. Looi et al. (Eds.), IOS Press, 2005

[21] M. A. Gómez-Martín, P. P. Gómez-Martín, and P. A. González-Calero, *Game-Driven Intelligent Tutoring Systems*, M. Rauterberg (Ed.): ICEC 2004, LNCS 3166, pp. 108–113, 2004.

[22] J. Leon, M. Fisher, *Interactive Educational Storytelling: The Use of virtual Characters to Generate Teachable Moments*, Night Kitchen Interactive, Albuquerque, New Mexico, March 22-25, 2006.

[23] RoboCode Central. Available at: *http://robocode.sourceforge.net/*

[24] K. Khan, *A Computer game to teach programming*. In Proc. National Educational Computing Conf., 1999. Available at: *http://www.toontalk.com/*

[25] CeeBot4: Teaching programming software. Available at: *http://www.ceebot.com/ceebot/4/4-e.php*

[26] L. Sheldon, *Character Development and Storytelling for Games*. Published by Thomson Course Technology, 2004.

[27] A. S. Gertner, C. Conati, & K. Van Lehn. *Procedural help in ANDES: Generating hints using a Baysian network student model*. In Proceedings of the 15th National Conference on Artificial Intelligence, 106-111, 1998.

[28] T. Jackson, E. Mathews, K. Lin, & A. Graesser. *Modeling Student Performance to Enhance the Pedagogy of Auto Tutor*. Proceedings of 9th International Conference, UM 2003, Johnstown, PA, USA, June 2003.

Intelligent Mobile Tour Guide

MeiYii Lim and **Ruth Aylett** [1]

Abstract. ‘Agents’ research has been going on for more than two decades now. The idea of improving assistance by agents capable of responding to the needs and emotional states of the users is not new, yet not much has been achieved so far. The main aim of this paper is to describe an intelligent context-aware mobile tour guide, having a biologically inspired architecture of emotion that allows the guide to adapt to the user’s needs and feelings. The resulting agent guides visitors touring an outdoor attraction as well as personalises the story presentation. A review of related work is presented, followed by a technical description of the system focusing on the core element - the guide’s emotional architecture and concluded with the current state of the project.

1 Introduction

Many research projects have explored the new possibilies of context-aware tour guide systems (eg. [1, 34, 24, 26, 14, 20, 4, 2, 33]) for augmenting the environment to provide guidance to users during a tour visit. This is part of the effort of ubiquitous computing to integrate computation into environment to enable people to interact with information in an inherently social manner. However, in interaction with current virtual guides, users tend to lose interest rapidly due to lack of ‘life’ and unmet expectations. This problem should be solved in order to prolong and produce a more engaging and natural interaction between the guide and user, also, to increase appreciation of a heritage site.

The better computational agents can meet our human cognitive and social needs, the more familiar and natural they are and the more effectively they can be used as tools [8]. Hence, intelligence and emotions are necessary for an effective computer system. Picard argues that “a machine, even limited to text communication, will be a more effective communicator if given the ability to perceive and express emotions” [27].

Supporting these arguments, the Intelligent Mobile Tour Guide is a guide with personality and beliefs, to provide guidance and engaging interaction during a tour visit. It addresses the frustration that usually occurs in the interaction with an emotionless computerised system that does not react sensitively to user’s feelings. The guide applies its beliefs, interests, user’s interests and its current memory activation to narrate stories. Decisions on story generation and updating of beliefs about user’s interests are affected by its internal processing controlled by an emotional model which receives input from the user.

The guide not only tells stories based on its own experiences and point of view, but attempts to evoke empathy in the user [19]. It attempts to persuade the user to think in the way it thinks, that is, to put the user in its own shoes. By seeing things from the guide’s perspective coupled with his/her own knowledge and understanding, a user will be able to analyse, enquire, reflect, evaluate and use the source of information critically to reach and support conclusions. In short, it makes the user envisage an event in a deeper sense and fosters learning, the attainment target of the UK National History Curiculum [23].

2 Technical description

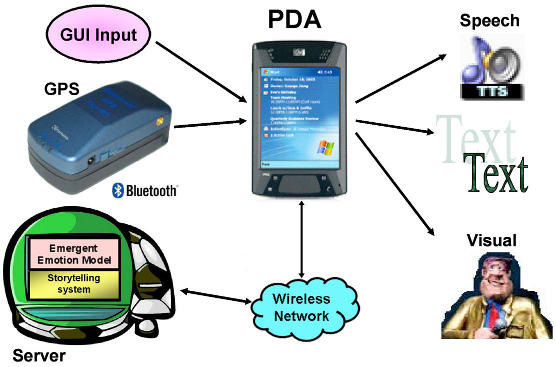


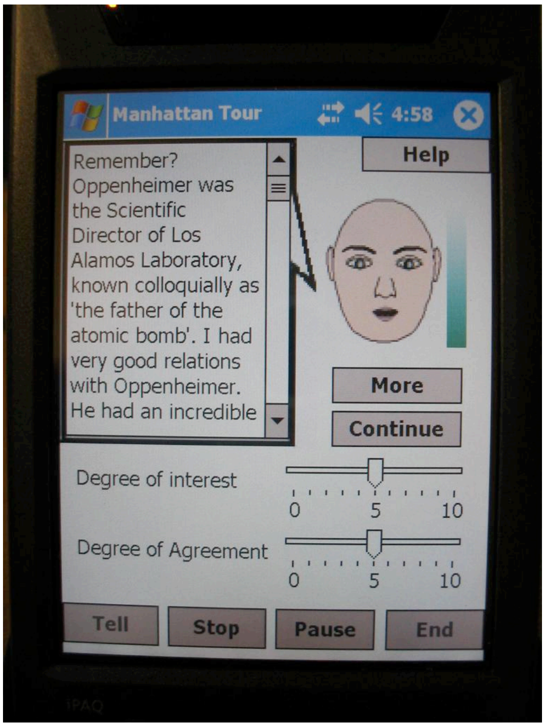
Figure 1. System Architecture

The Affective Guide is implemented on a PDA, taking advantage of the expanding mobile technologies such as Wi-Fi wireless hotspots and bluetooth access points. Multiple modalities are used to complement each other and focus the user’s attention on the information presentation. User’s position is determined by a Global Positioning System while user’s orientation is calculated based on previous and current location information.

Prior to a tour, there is an ice-breaking session where the guide extracts information about the user’s name and interests. It then chooses attractions that match the user’s interests and plans a route to the destinations in such a way that it is the shortest route possible. It then navigates the user to the chosen locations by giving directional instructions as well as presenting the user with an animated directional arrow. The guide will notify the user upon arrival at a destination and start the storytelling process. Since tourist information is location-dependent by nature, the system links electronic data to actual physical locations, thereby augmenting the real world with an additional layer of virtual information. A server performs the processing and holds the guide’s memories, both long-term and current, and sends the results of processing to the PDA on demand.

A ‘Head up’ approach is adopted where stories are presented using speech allowing the user to have full appreciation of the attraction visited. The text is also displayed on the screen allowing the user to read any information missed in the speech. After each storytelling cycle, the user can choose to have ‘More’ stories about the current

[1] Heriot-Watt University, Scotland, email: {myl, ruth}@macs.hw.ac.uk



Remember? Oppenheimer was the Scientific Director of Los Alamos Laboratory, known colloquially as 'the father of the atomic bomb'. I had very good relations with Oppenheimer. He had an incredible

Figure 2. The Graphical User Interface

location or 'Continue' to the next location. This is the only time the user is required to look at the screen to provide feedback on the degree of interest of the story and his/her degree of agreement to the guide's argument. The user inputs reflect his/her current feeling and opinion about the story content, useful for personalising further stories.

As can be seen in Figure 2, a simple graphical user interface has been designed due to the limited space on the PDA and to reduce cognitive load for the user. The user is given the flexibility to stop the speech if they are not interested in the currently presented story. User can also ask the guide to pause, resume or repeat the current story. Furthermore, acknowledgement and notification using both speech and message boxes are provided periodically to reduce long idle states and as an assurance to the user that the system is operating as intended. If the user is attracted to a site which is not in the preplanned route, they can activate the storytelling process manually by clicking the 'Tell Story' button. If information is available, the guide will start the narration, else the user will be notified about the unavailability of information. At any time, the user can activate the 'Help' menu if he/she is unsure about the function of a particular button on the interface.

3 Body-Mind Architecture

3.1 Related Work

As mention in the Introduction, emotional systems are essential part of an intelligent computer agent. Thus, researchers on character development are paying attention to the design of motivational structures, emotional and personality traits and behavior controls systems for characters to perform in context-specific environments with well-defined goals and social tasks [10, 16]. They have long wished to build creatures with whom you'd want to share some of your life whether as a companion or a social pet.

Cañamero [5] proposed an architecture that relies on both motivations and emotions to perform behavior selection. Under normal circumstances, behavior selection is driven by the motivational state of the robot. Emotions constitute a 'second order' control mechanism running in parallel with the motivational control system to continuously monitor the external and internal environment for significant events. However, the main problem with this architecture is that it was totally hand-coded.

On the other hand, Velásquez's work [36] is inspired by findings in neuropsychology that relies on the use of computational frameworks for Emotion-Based Control. The model integrates perception, motivation, behavior and motor control with particular emphasis on emotions as building blocks for the acquisition of emotional memories. Velásquez's robot, *Yuppy*, utilized feed backward operation of emotion where previous emotional experiences are fed back to the behavior system forming an emotional memory, which affects action selection strategy when it re-encounters similar situations. However, *Yuppy* capabilities are prespecified and it does not show emotional responses to a novel object or situation.

Next, the OCC model [25] is one of the most used appraisal models in current emotion synthesis systems although the theory was not intended to be used for emotion synthesis by the authors. OCC model works at the level of emotional clusters, called emotion types, where the emotions within each cluster share similar causes. This model proposes that emotions are the results of three types of subjective appraisals: the appraisal of the pleasingness of events with respect to the agents goal, the appraisal of the approval of the actions of the agent or another agent with respect to a set of standard for behavior and the appraisal of the liking of objects with respect to the attitudes of the agent. Numerous implementations of this model were seen, for example, the Affective Reasoner architecture [11] and the Em component of the Hap archtecture [3].

Klaus Scherer [32] explicitly proposes treating emotion as a psychological construct consisting of five components: cognition appraisal, physiological activation, motivation tendencies, motor expression and subjective feeling state. He proposed the 'component process model of emotion' and suggested that emotion can be defined as an episode of temporary synchronization of all major subsystems of organismic functioning represented by these components. Furthermore, he suggest that there may be as many emotions as there are different appraisal outcomes.

The Oz project [3, 21] attempted to build a small simulated world, containing several real-time, interactive, self-animating creatures. It aimed at producing agents with a broad set of capabilities, including goal-directed and reactive behavior, emotional state, social knowledge and some natural language abilities where individual *Woggles* had specific habits and interests which were shown as different personalities. Social relations between the agents directly influenced their emotional system and vice versa. Oz focused on building specific, unique believable characters, where the goal is an artistic abstraction of reality, not biologically plausible behavior.

AlphaWolf's [35] emotional model is based on the Pleasure-Arousal-Dominance model presented by Mehrabian and Russell [22]. It captures a subset of the social behavior of wild wolves, involving models of learning, emotion and development. The wolves' emotions lead to formation of context-specific emotional memories based on the "somatic marker hypothesis" presented by Damasio [7], which affects how they will interact in the future. This research emphasises social learning and offers initial steps toward a computational system with social abilities.

All the above projects involves explicit labelling of emotions and focus either on the neurophysiological aspect of emotion, or on the cognitive aspect, adopting the notion of appraisal. Very few attempts have been carried out to bridge the gap between these two aspects where models such as perception, motivation, learning, action-selection, planning and memory access are integrated. Two effort in this direction are [9], the emotional model adopted by the Intelligent Mobile Tour Guide described further in Section 3.2 and [28].

[28] aims to investigate improved realism in generating complex human-like behavior by integrating behavior moderators with higher cognitive processes. It integrates a connectionist cognitive model of emotional processing called SESAME [6] with a synthetic force model, SOF-Soar architecture [13] for training in a battlefield simulation. The response system accepts information from, while appraisal system provides information to, the connectionist emotions model. Emotional states can be viewed as arising from a combination of pleasure/pain, arousal, clarity/confusion components and by changing these connection strengths, different personalities result.

3.2 Emergent Emotion Model

The emotional architecture of the guide is based on the ‘PSI’ model [9]. It is biologically inspired where the interest lies in modelling the conditions to the emergence of emotions to avoid rigidness in behavior and provide more colors to the resulting emotions. In this architecture, emotions emerge from the modulation of information processing, action selection, planning and memory access. The guide continuously forms memories, expectations and immediate evaluations, resulting in behavior that can be termed emotional.

The guide has two built-in motivators to maintain. It needs to preserve its level of competence and adjust its behavior appropriately to the level of uncertainty. The level of competence refers to its capability to cope with differing perspectives about an issue or event whereas the level of uncertaintly is the degree of predictability of the environment and the user interests. For example, if user disagrees with the guide’s opinion, its level of competence decreases. Furthermore, if the user finds the stories uninteresting, its level of uncertainty increases as its prediction about user’s interests is incorrect.

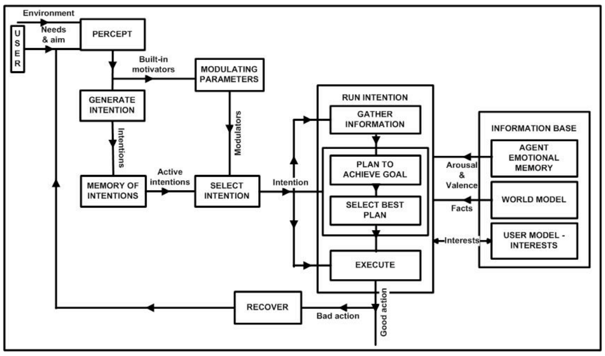


Figure 3. The Emergent Emotion Model

Functionally, the guide reads the user inputs, system feedback and the GPS information continuously, then, generates an intention, let’s say to tell a story. The intention together with its built-in motivators are stored in a memory of intentions. The guide makes assumption about the user’s interest based on the initial information extracted through the ice-breaking session. Basically, the guide has three possible intentions that it can select - update its belief about the user’s interests, adjust the story topic and presentation or perform storytelling.

More than one intention can be active at the same time. Depending on the importance of the need and the urgency for its realization, one of the active intentions is selected. For intention execution, the guide decides autonomously whether to explore for more information, to design a plan using the available information or to run an existing plan. The decision on how to perform the intention is made based on the value of its built-in motivators and modulators such as arousal level, resolution level and selection threshold, or in in other words, the agent’s current emotional state. Arousal level refers to the speed of processing or the agent’s readiness to act. Resolution level determines the carefulness and attentiveness of the guide’s behavior. Lastly, selection threshold is the limit competing motives have to cross in order to take over.

Besides emotions, personality plays an important role in the guide. Results from our survey of human tour guides show that factors like role, interest, experience, guide’s belief, guide’s personality, type of tour and visitor group affect the information presentation. Different guides have different styles and most guides tend to incorporate belief and past experiences whether his/her own or others while narrating a story. Similarly, the intelligent mobile guide’s personality is reflected through its perspective about a particular historical event. Furthermore, in our model, personality emerges from varying the weight of each modulator as discussed in [18]. Like emotions, personality is not defined explicitly but results from overall activity of the guide and by its patterns of interaction.

3.3 The Guide’s Memory

The guide possesses a long-term memory that is made up of declarative memories, both semantic and emotional memories [18]. Semantic memory is its memory for facts, including location-related information, definition of concepts, the user’s profile, etc. Each piece of the guide’s semantic memory contains the following features:

name	:	as an identification of the memory piece
type	:	the type of event
subjects	:	the subjects involved in the event
objects	:	the objects involved in the event
effects	:	the effects of the event
concepts	:	basic elements in the piece of memories that has a more detailed definition
attributes	:	describes the nature of the story element, for example, science, military, social
location	:	the associated physical location where the event occur
text	:	the text encoding the event

While the semantic memory contain facts, emotional memory is a memory for events that have emotional impact on the guide. The emotional memory is tagged with ‘arousal’ and ‘valence’ [17] tags analogous to the *Emotional Tagging* concept [30], which recorded the guide’s emotional states for an event. The guide’s emotional memory holds a certain ideology, defined simply as beliefs held by the guide, that reflects its perspective about an issue or event. It is a manifestation of the guide’s past experiences. The guide’s emotional memory pieces have a similar structure to the semantic memory pieces with the addition of the following:

arousal : the arousal value when an event took place
valence : the emotional valence value an the event took place

3.4 Storytelling System

When interacting with the visitor, the guide will be engaged in meaningful reconstruction of its own past, at the same time presenting facts about the site of attraction. The guide adopts the storytelling technique proposed by [15], however, with some modifications.

In every step, the guide decides what to tell dynamically. It constructs stories by improvising taking into account factors such as the already told story at the current moment and the affinity between story element and the guide's interests as well as the user's profile. Three scores corresponding to these factors are calculated each time, which are then combined to calculated an overall score for each candidate pair of story element and location. It selects a memory spot, that is a memory element with the highest overall score. This spot will lead to further extension of facts as well as emotional memory elements depending on its current resolution level. The retrieval of memory pieces continues until the combined memory pieces is large enough to generate a story as illustrated in Figure 4. All these extension processes are performed by Jess [2], a Java-based rule engine.

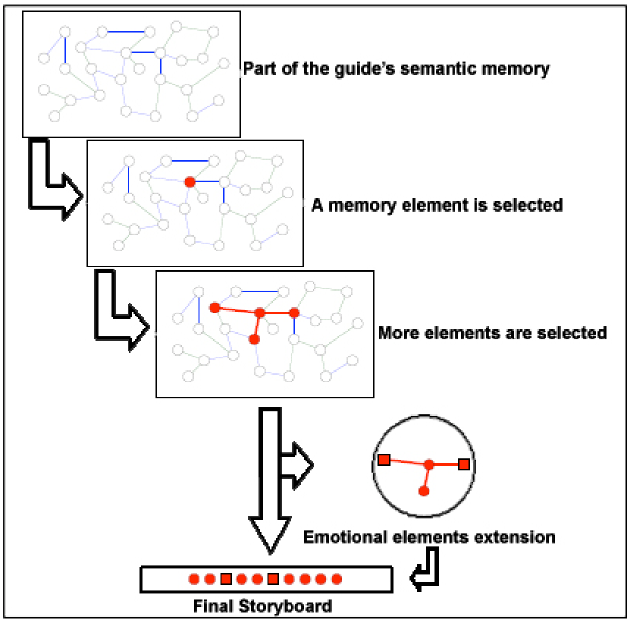

Part of the guide's semantic memory
A memory element is selected
More elements are selected
Emotional elements extension
Final Storyboard

Figure 4. The Story Extension Process

After each story presentation, the guide will update its current memory so that the next retrieval will be based on the current active memory elements resulting in a reminding process. Reminding is a crucial aspect of human memory and it can take place across situations. The memory elements of the guide are retrieved based on reminding methods: processing-based reminding and dictionary-based reminding as discussed by [31].

Processing-based reminding occurs during the normal course of understanding or processing of new information. A *scene or location* is a kind of structure that provides a physical setting serving as the basis for reconstruction. Besides that, the guide's memory elements are activated based on *subject-object* links as one person can remind us of another, one object can remind us of another object or one event can remind us of another. *Cause-effect* links also act as a reminding criterion where a particular event leads to the activation of another element of memory which is the cause or the effect of the current memory element.

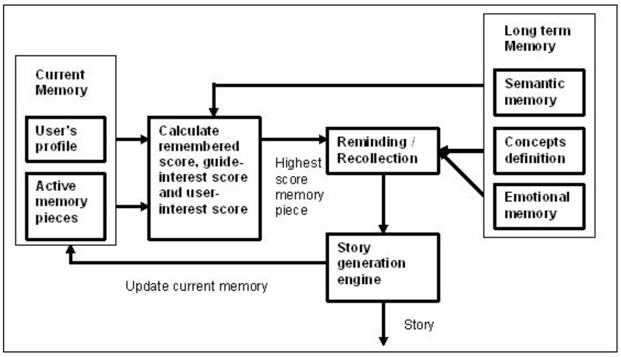


Figure 5. The Storytelling Process

The changeability of dynamic memory makes people's memory acts differently in apparently similar situations. We are usually being reminded of similar events or the closes previously experienced phenomenon. In this sense, *attributes* are use to link and retrieve the guide's memories on similar events or circumstances. On the other hand, dictionary-based reminding occurs when the guide searches for the definition of an infrequent word, concept or object. It uses the *concept* element of the memory piece to further elaborate the stories by retrieving the definition when a concept occurs for the first time. Analogous to human memory, a concept strength in the guide's memory increases when it is activated frequently and will be forgotten if not used after a few iterations.

3.5 Overall Process

The recollective experience of the guide is related to the evocation of previously experienced emotions through the activation of the emotion tags. These values combine with the built-in motivators values to trigger the resolution level and selection threshold, resulting in re-experiencing of emotions, though there might be a slight variation due to the input from the user. The user's response, contributes to the guide's certainty level by confirming or disconfirming the guide's prediction. On the other hand, the degree to which he or she agrees with the guide's argument, contributes to the guide's level of competence.

Let's take a look at some examples. If the guide's prediction about the user's interests is correct (high level of certainty) and the user perspective is consistent with that of the guide (high level of competence), the guide may experience low to medium level of arousal and selection threshold with a medium resolution level. In this case, the guide may be said to experience pride because it could master the situation. It is not so easy for another goal to take over. The agent will perform some planning and provide a more elaborated story on the current subject based on its active ideology. The guide's belief about the user's interests is srengthened. This is consistent with the

[2] http://herzberg.ca.sandia.gov/jess/

argument of Fiedler and Bless [12] in which agent experiencing positive affective states fosters assimilation that supports reliance and the elaboration of its existing belief system.

However, if the guide's prediction about the user's interests is right (high level of certainty) but the user's perspective is in conflict with the guide's ideology (lower level of competence), then the arousal level of the guide may be higher than the previous case. The resolution level decreases while the selection threshold increases. In this case, the guide may have some difficulties coping with the differing perspective, but since it has anticipated the situation, it is motivated to concentrate on the specific goal and adjusts the presentation of story appropriately by giving a more general view on the issues instead of presenting them from its own ideological standpoint.

Next, in the case that the guide's prediction about the user's interests is wrong (low level of certainty) but the user's perspective is consistent with the guide's ideology (high level of competence), the arousal level of the guide may be equal to or lower than the second case. The guide is still in control of the situation making the uncertain environment look less threatening. Nevertheless, the guide may be disappointed or sad in relation to its wrong prediction. The selection threshold decreases and the resolution level increases. Now, the guide will perform more detailed and substantive processing to elaborate its perspective and overcome the discrepancy by changing its beliefs about the user's viewpoint. This is again supported by the discussion of Fiedler and Bless that negative states trigger accommodation processes, allowing beliefs to be updated.

On the other hand, if the guide's prediction about the user's interests is wrong (low level of certainty) and the user's perspective is in conflict with the guide's ideology (low competence level), the arousal level of the guide will be very high. It is reasonable to react quickly, concentrate on the respective task and refrain from time consuming memory search. Therefore, the selection threshold should be high while its resolution level should be low in which case, we may diagnose that the guide is experiencing anxiety. In this situation, a biasing effect occurs and the guide tends to give a more general story of the current site without details. The current situation will be fedback to the system so that the guide can adjust its beliefs appropriately to better cope with the situation in future.

By doing so, it adapts its behavior according to its internal states and the environmental circumstances. Each execution of intention will produce a feedback into the system and recovery or the guide's belief is updated as necessary.

3.6 Example Stories

The 'Los Alamos' site of the Manhattan Project [3] has been choosen for the prototype implementation of the Intelligent Mobile Guide System, where the buildings are mapped onto Heriot-Watt Edinburgh campus buildings. Hence, all the stories are related to the 'Making of the atomic bomb' [29]. Below is an extract of a story from the non-emotional and emotional guide presentation.

Non-emotional guide's presentation:

> *The first Japanese bombing target, Hiroshima was of such size that the damage would be confined within it, so that definite power of the bomb could be determined. Little Boy exploded at 8:16:02, August 6, 1945, Hiroshima time, one thousand nine hundreds feet above the courtyard of Shima Hospital, with a yield equivalent to twelve thousands five hundred tons of TNT. Trinitrotoleune or TNT is a pale yellow crytalline aromatic hydrocarbon compound that melts at eighty one degree Celcius. It is an explosive chemical used on its own or in many explosive mixtures such as Torpex, Tritonal, Composition B or Amatol. It is difficult to dissolve TNT in water; it is more soluble in ether, acetone, benzene and pyridine. The explosive yield of TNT is considered the standard measure of strength of bombs and other explosives.*

Emotional guide's presentation (medium level of resolution):

> *The first Japanese bombing target, Hiroshima was of such size that the damage would be confined within it, so that definite power of the bomb could be determined. Little Boy exploded at 8:16:02, August 6, 1945, Hiroshima time, one thousand nine hundreds feet above the courtyard of Shima Hospital, with a yield equivalent to twelve thousands five hundred tons of TNT. Trinitrotoleune or TNT is a pale yellow crytalline aromatic hydrocarbon compound. Its explosive yield is considered the standard measure of strength of bombs and other explosives. The important result of Hiroshima bombing and the one that we sought, was that it brought home to the Japanese leaders the utter hopelessness of their position. When this fact was reemphasized by the Nagasaki bombing, they were convinced that they must surrender at once. The Air Force is operating primarily to laying waste all the main Japanese cities. Their procedure had been to bomb the hell out of Tokyo, bomb the manufacturing and assembly plants, and in general paralyze the aircraft industry so as to eliminate opposition to its operations.*

Emotional guide's presentation (high level of resolution):

> *The first Japanese bombing target, Hiroshima was of such size that the damage would be confined within it, so that definite power of the bomb could be determined. Little Boy exploded at 8:16:02, August 6, 1945, Hiroshima time, one thousand nine hundreds feet above the courtyard of Shima Hospital, with a yield equivalent to twelve thousands five hundred tons of TNT. Trinitrotoleune or TNT is a pale yellow crytalline aromatic hydrocarbon compound. Its explosive yield is considered the standard measure of strength of bombs and other explosives. The important result of Hiroshima bombing and the one that we sought, was that it brought home to the Japanese leaders the utter hopelessness of their position. When this fact was reemphasized by the Nagasaki bombing, they were convinced that they must surrender at once. The Air Force is operating primarily to laying waste all the main Japanese cities. Their procedure had been to bomb the hell out of Tokyo, bomb the manufacturing and assembly plants, and in general paralyze the aircraft industry so as to eliminate opposition to its operations. With the success of the Hiroshima weapon, the pressure to be ready with the much more complex implosion device became excruciating. We felt that the sooner we could get off another mission, the more likely it was that the Japanese would feel that we had large quantities of the devices and would surrender sooner.*

4 CONCLUSION

The focus of this research is the development of the body-mind model for the guide. A prototype of the system has been completed and is currently being evaluated. It is hoped that the evaluation will finish in a few weeks time and a detailed analysis can be performed to

[3] http://www.lanl.gov/

verify the usefulness and adaptive capability of the system. In future, it will be desirable if morphing technique can be utilised to reflect the guides emotional states, providing an infinite range of expressions.

ACKNOWLEDGEMENTS

Work supported by the European Union's Sixth Framework Programme, in the IST (Information Society Technologies) Thematic Priotity IST-2002-2.3.1.6 Multimodal Interfaces, HUMAINE (Human-Machine Interaction Network on Emotion) [4] (Contract no. 507422). The authors are solely responsible for the content of this publication. It does not represent the opinion of the European Community and the European Community is not responsible for any use that might be made of data appearing therein.

REFERENCES

[1] G. D. Abowd, C. G. Atkeson, H. Hong, S. Long, R. Kooper, and M. Pinkerton, 'Cyberguide: A mobile context-aware tour guide', *Wireless Networks*, **3**(5), 421–433, (1997).

[2] P. Almeida and S. Yokoi, 'Interactive character as a virtual tour guide to an online museum exhibition', in *Proceeding of Museum and the Web 2003*, (2003).

[3] J. Bates, A. B. Loyall, and W. S. Reilly, 'An architecture for action, emotion, and social behavior', *Lecture Notes in Computer Science*, **830**, 55–69, (1994).

[4] A. C. Bertolleti, M. C. Moraes, and A. Carlos da Rocha Costa, 'Providing personal assistance in the sagres virtual museum', in *Proceeding of Museum and the Web 2001*, (2001).

[5] D. Cañamero, 'Modeling motivations and emotions as a basis for intelligent behavior', in *Proceedings of the 1st International Conference on Autonomous Agents*, eds., W. Lewis Johnson and Barbara Hayes-Roth, pp. 148–155, New York, (February 5–8 1997). ACM Press.

[6] E. Chown, *Consolidation and Learning: A Connectionist Model of Human Credit Assignment*, PhD thesis, University of Michigan, 1993.

[7] Antonio R Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*, G.P. Putnam, New York, 1994.

[8] K. Dautenhahn. The art of designing socially intelligent agents – science, fiction and the human in the loop, jul, 7 1998.

[9] D. Dörner, 'The mathematics of emotions', in *Proceedings of the Fifth International Conference on Cognitive Modeling*, eds., Dietrich Dörner Frank Detje and Harald Schaub, pp. 75–79, Bamberg, Germany, (apr, 10–12 2003).

[10] P. Doyle and K. Isbister. Touring machines: Guide agents for sharing stories about digital places, 1999.

[11] C. D. Elliot, *The Affective Reasoner: A process model of emotions in an multi-agent system*, Ph.D. dissertation, Northwestern University, Illinois, 1992.

[12] K. Fiedler and H. Bless, 'The formation fo beliefs at the interface of affective and cognitive processes', in *Emotions and Beliefs*, eds., Nico H. Frijda, Anthony S. R. Manstead, and Sacha Bem, 144–170, Cambridge University Press, Cambridge, UK, (2000).

[13] F. Koss G. Taylor and P. Nielsen, 'Special operations forces ifors', in *Proceeding of the 10th Conference on Computer Generated Forces and Behavioral Representation*, pp. 301–306, Norfolk, VA, (May 15–17 2001).

[14] T. Höllerer, S. Feiner, T. Terauchi, G. Rashid, and D. Hallaway, 'Exploring mars: developing indoor and outdoor user interfaces to a mobile augmented reality system', *Computers and Graphics*, **23**(6), 779–785, (dec 1999).

[15] J. Ibanez, *An Intelligent Guide for Virtual Environments with Fuzzy Queries and Flexible Management of Stories*, Ph.D. dissertation, Departamento de Ingenieria de la Informacion y las Communicaciones, Universidad de Murcia, Murcia, Spain, 2004.

[16] W. Lewis Johnson, Jeff W. Rickel, and James C. Lester, 'Animated pedagogical agents: Face-to-face interaction in interactive learning environments', *The International Journal of Artificial Intelligence in Education*, **11**, 47–78, (2000).

[17] Elisabeth A. Kensinger and Suzanne Corkin, 'Two routes to emotional memory: Distinct neural processes for valence and arousal', *PNAS*, **101**(9), 3310–3315, (March 2 2004).

[18] Mei Yii Lim, Ruth Aylett, and Christian Martyn Jones, 'Emergent affective and personality model', in *The 5th International Working Conference on Intelligent Virtual Agents*, Kos, Greece, (September 12–14 2005).

[19] Mei Yii Lim, Ruth Aylett, and Christian Martyn Jones, 'Empathic interaction with a virtual guide', in *Proceeding of the Joint Symposium on Virtual Social Agents, AISB'05:Social Intelligence and Interaction in Animals, Robots and Agents*, pp. 122–129, Hatfield, UK, (April 12–15 2005).

[20] R. Malaka and A. Zipf, 'Deep map challenging it research in the framework of a tourist information system', in *Information and Communication technologies in tourism*, eds., D. Buhalis, D. R. Fesenmaier, and S. Klein, Springer-Verlag, New York, (2000).

[21] Michael Mateas. An oz-centric review of interactive drama and believable agents, February 25 1997.

[22] A. Merahbian and J. Russell, *An Approach to Environmental Psychology*, MIT Press, Cambridge, MA, 1974.

[23] NHC. History: The level descriptions, 2006. http://www.ncaction.org.uk/subjects/history/levels.htm, Accessed Oct 15, 2006.

[24] M. J. O'Grady, R. P. O'Rafferty, and G. M. P. O'Hare, 'A tourist-centric mechanism for interacting with the environment', in *Proceedings of the First International Workshop on Managing Interactions in Smart Environments*, pp. 56–67, Dublin, Ireland, (dec 1999). Springer.

[25] A. Ortony, G. Clore, and A. Collins, *The cognitive structure of emotions*, Cambridge University Press, Cambridge, UK, 1998.

[26] D. Petrelli, E. Not, and M. Zancarano, 'Getting engaged and getting tired: What is in a museum experience', in *Proceedings of the Worshop on Attitude, Personality and Emotions in User-Adapted Interaction held in conjunction with UM '99*, Banff, (June, 23 1999).

[27] R. W. Picard, *Affective Computing*, MIT Press, 1997.

[28] Eric Chown Randolph M. Jones, Amy E. Henninger, 'Interfacing emotional behavior moderators with intelligent synthetic forces', in *Proceeding of the 11th CGF-BR Conference*, Orlando, FL, (May 7 2002).

[29] Richard Rhodes, *The Making of the Atomic Bomb*, Simon & Schuster, New York, 1986.

[30] Gal Richter-Levin and Irit Akirav, 'Emotional tagging of memory formation - in the search for neural mechanisms', *Brain Research Reviews*, **43**, 247–256, (2003).

[31] Roger C. Schank, *Dynamic memory: A theory of reminding and learning in computers and people*, Cambridge University Press, United States, 1982.

[32] K. R. Scherer, 'Appraisal considered as a process of multi-level sequential checking', in *Appraisal processes in emotion: Theory, Methods, Research*, eds., A. Schorr K. R. Scherer and T. Johnstone, pp. 92–120, New York and Oxford, (2001). Oxford University Press.

[33] O. Stock and M. Zancarano. Intelligent interactive information presentation for cultural tourism. Invited talk at the International Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, Copenhagen, Denmark, Jun 2002.

[34] Y. Sumi, T. Etani, S. Fels, N. Simone, K. Kobayashi, and K. Mase, 'C-map: Building a context-aware mobile assistant for exhibition tours', *The First Kyoto Meeting on Social Interaction and Communityware*, (jun 1998).

[35] Bill Tomlinson and Bruce Blumberg. *AlphaWolf*: Social learning, emotion and development in autonomous virtual agents, October 04 2002.

[36] J. Velásquez, 'A computational framework for emotion-based control', in *Proceeding of the Grounding Emotions in Adaptive Systems Workshop, SAB '98*, Zurich, Switzerland, (1998).

[4] http://emotion-research.net

Evaluating synthetic Actors

Sandy Louchart and **Ruth Aylett**[1]

Abstract

We discuss the extension of an emotionally-driven agent architecture already applied to the creation of emergent narratives. Synthetic characters are enhanced to perform as actors by carrying out a second cognitive appraisal, based on the OCC model, of the emotional impact of their projected actions before execution. We present the evaluation of this approach and some initial results on whether it produces more 'interesting' narratives.

1 INTRODUCTION

Narrative has become a topic of great interest in video and computer games development as a way of drawing the player into the game play [16], and is seen as a focus for the development of mobile and Augmented Reality-based gaming [21]. Much active research addresses the generic use of interactive graphical environments and intelligent synthetic characters to extend the power of narrative in new ways [16]. Specifically it has played a central role in a number of interactive graphics-based e-learning systems both for adults [24] and children [9, 18]. Narrative is also used as a generic method for adding intelligence to virtual environments, for example, through the development of virtual guides [4].

The key characteristic of all these environments is interactivity: users expect to move freely and interact at will with objects and synthetic characters. Yet this interactional freedom clashes badly with the conventional narrative requirement for a definite structure, creating a narrative paradox [13]. A plot-based narrative structure supposes the right actions at the right places and times but these may not be those the user chooses to carry out [19]. More generally, an authorial plot-based view of narrative where particular actions must execute in a particular order conflicts with a character-based view where characters autonomously select their actions in response to their sensing of the state of the virtual world – strong autonomy [15].

Merging the roles of spectator and author evades rather than reconciles the contradiction since authoring merely allows a plot-based approach to be maintained; this approach has been exploited in a number of systems [9, 18, 20]. The God-like perspective of games such as 'The Sims' gives the privileged user overall responsibility for the activity within the virtual world in a similar fashion. Creating a branching narrative is another solution [24, 14], though either the user is constrained into a few key choices, breaking their immersion in the narrative world, or characters must be supplied with "universal plans" [23] covering every possible response to whatever the user does. Façade [15] is an impressive example of the result of doing this, using the concept of 'beats', based on an adaptation of Aristotelian theory, but required substantial authoring effort for a short (20 minute) narrative, with clear implications for scalability. Limiting the interactive stance of the user is a third solution: one may apply concepts such as Boal's [3] spect-actors, in which participation and spectating are episodically interleaved [2]. In [5] characters have universal plans expressed as AND/OR trees but the role of the user is confined to manipulation of key objects, forcing character re-planning.

Strong autonomy for characters offers a potential solution to the problem of interactivity since if synthetic characters are allowed to autonomously select actions, then a participating user can also be allowed to do so on the same terms. Given that in general, structure can emerge from interaction between simpler elements, it seems possible that interaction between strongly autonomous characters could under specific circumstances produce narrative structure, or an emergent narrative (EN) [1].

The main objection to character-based narrative based on strong autonomy is that there is no guarantee that interesting narrative structure will result precisely because characters are responding to their internal state and individual goals in choosing actions and not to the overall story structure. However, an existential proof of the EN approach can be found in interactive forms such as improvisational drama and human RPGs: in the former actors start from a well-defined initial state and strong roles and select 'dramatically-interesting' actions, while in the second, a game-master dynamically manages the experience of the autonomous participants [13]. In this work we discuss the application of both these ideas within the additional framework of affective appraisal theory.

The hypothesis being explored is that an autonomous agent that explicitly assesses the emotional impact of its actions on other agents around it, much as an actor would, will produce a more engaging emergent narrative than one that only uses its own 'in-role' emotional state to select its next action. Other virtual actors [22] have not tried to assess the differential emotional impact of a set of possible 'in-role' actions, making this a novel approach. Because it uses emotional impact, it is also different from assessing the goals or plans of other agents [11].

2 NARRATIVE AND EMOTION

If narrative is to emerge from interaction between characters, then the character architecture is fundamentally important. It is the contextual relevance and richness of the actions selected by each character that will or will not produce sequences with the post-hoc structure of a story: that is a coherent compound of external interest and surprise (causal chains of actions) with internal perceived intentionality and emotional impact (motivation and expressive behaviour). Displaying role-specific emotional reactions to the actions of other characters and the emotion behind their own actions is an important component of successful human acting.

[1] MACS Heriot-Watt University EH14 4 AS e-mail: {sandy, ruth}@macs.hw.ac.uk

For this reason a number of researchers in synthetic characters, starting with Elliot's Affective Reasoner [7] have integrated affect into their agent architectures [8, 2], usually drawing on cognitive appraisal theory. Appraisal is the human perceptual process through which objects, other characters and events are related to the needs and goals of an individual, generating a resulting emotional response and thus linking emotion to cognition. The most widely implemented system is the taxonomy of Ortony, Clore and Collins (OCC) [17], used by the FatiMA agent architecture which formed the basis for the work described here. The OCC model is an approach based on a valenced (good or bad) reaction to an event and the structure of emotions it defines can be seen as a hierarchical taxonomy organising 22 emotion types.

3 AFFECTIVE AGENT ARCHITECTURE

The FatiMA (Fearnot Affective Mind Architecture) [6] agent architecture is shown in **[Figure 1]** (with the additions of the work reported here added in red) and is that used in FearNot!, an application that generates episodes of emergent virtual drama relating to bullying for educational purposes [2]. In this architecture, an agent's emotional status affects its drives, motivations, priorities and relationships, with an OCC-based appraisal system and resulting coping behaviour [12] - those internal emotional adjustments made or external actions taken in order to deal with negative emotions. Characters may also have different thresholds and decay rates for each of the 22 OCC emotions, implicitly defining a large set of different personalities.

As shown in *Figure 1*, the appraisal mechanism consists of both a reactive and deliberative layer [2,6]. The former is handled by a set of emotional reaction rules consisting of an event that triggers the rule and values for the OCC appraisal variables affected by the event (desirability, desirability-for-other, praiseworthiness etc).

The deliberative layer is responsible for appraising events according to the character's goals, thus generating prospect-based emotions like hope and fear. These emotions relate to future events: those congruent with the IVA's goals (hope) or those threatening them (fear). They thus connect the affective system to the planning component of coping behaviour [8].

The action selection process is also composed of reactive and deliberative levels. Reactions consist of a set of action rules: each contains a set of preconditions that must be true in order to execute the action and an eliciting emotion that triggers this particular action, for example sadness may trigger weeping. The action set is matched against all the emotions present in the character's emotional state (arising from appraisal) and the set of rules with positive matches is activated. The action rule triggered by the most intense emotion is selected for execution. If more than one action rule is selected, the most specific one is preferred.

The deliberative coping process - deeply connected to the deliberative appraisal process - is more complex. More than one goal can be active at the same time, so the first stage of the deliberative reasoning process is to determine which goal to attend to. In the original architecture, the intentions generating the strongest emotions are the ones that require the most attention from the agent, and thus are the ones selected by the planner to continue deliberation.

The next step is to choose the most appropriate existing plan to execute or to continue planning. An evaluation metric is used that: weights plans that achieve the same conditions but use fewer steps more highly; weights plans with more instantiated pre-conditions more highly; and plans with fewer inter-goal threats more highly. For example, within the bullying scenarios to which FatiMA has so far been applied, a plan by a victim to hit the bully threatens the victim's own goal of not getting hurt. At this point, the best plan is brought into focus for reasoning, as if in the forefront of the agent mind, and at this point it generates/updates the corresponding emotions [6]. It is here that there is an opportunity to have the agent consider what the emotional impact of plans on other characters might be.

The planner removes only one flaw or starts to execute one action in each cycle of coping, so that an agent does not 'freeze' in prolonged thought. Building up a plan takes several coping cycles, so that an appraisal may change from an initially strong hope to a strong fear as the character realizes that no feasible plan exists. This type of appraisal is called *Reappraisal* since it is not based on external events or stimuli, but is driven by the agent's internal processing. However it is an entirely self-centred reappraisal which does not in the original architecture take into account the impact of plans on other agents.

3.1 Double appraisal

The design of an agent action-selection mechanism that selects dramatically interesting actions is a technical and conceptual challenge. In particular, the subjective nature of drama and its perception makes the development of a reliable and quantifiable assessment measure very difficult. The idea explored here is to take emotional impact (EI) as a surrogate for dramatic interest, hypothesising that the EI of a specific action relates to its dramatic impact and could thus substitute for dramatic value. A character would therefore take an action not solely on the basis of its emotions, goals and motivations but also on the EI of these actions for both itself and other characters. This approach would allow the characters to conjointly assume in a distributive manner the dramatic weight of an unfolding story without relying on a pre-determined plot.

3.2 Architectures

We argue that the implementation of such a concept requires a novel agent action-selection mechanism whose function is not only to make action decisions but also to project the possible impact of these decisions. The mechanism described in this section features a double appraisal cycle as opposed to the single approach discussed above. This allows the agent to appraise events as in any conventional appraisal-based system but then carry out conflict resolution over a set of possible actions by running another appraisal cycle (in parallel), assessing each member of the feasible in-role action set according to its potential emotional impact. Thus the selection of an action is made not just on the inherent value of a particular action but on its ability to generate EI. The mechanism has been implemented within the already existing FAtiMA architecture, at the coping level, and features two related approaches for evaluation purposes.

In the first implementation, [Double Appraisal (DA)], the agent generates a set of possible actions using its emotions and goals and then assesses the emotional impact each action would have *if directed at itself*. An extra loop is added into the appraisal process by recasting each possible action into an event and feeding it back into the agent's own appraisal system. This corresponds to a "Theory of Mind" approach [25] in which the agent assumes that everyone else would react as they would: "how would I feel if someone did this action to me?" In order not to affect the actual current emotional state of the agent, this re-appraisal cycle is executed in parallel with the agent "appraisal-coping" cycle and takes place within an instance of the agent's mind that is not connected to its running emotional state.

The second application [Double Appraisal with Modelling (DAM)] [Figure 1] draws on the same principles but conducts the re-appraisal with respect to the emotional reactions sets of all the agents present in the scenario. It aims at selecting the action that would have the highest emotional impact of that on all the characters within a scenario. This corresponds to "how would the most-affected of the people around me feel if I did this action?" A significant parameter in either approach is the size of the set of possible actions. Each of the implementations DA and DAM has been evaluated with a low value for the number of actions in the possible set (3) and with a higher number (9). The aim here is to establish whether the number of actions presented to the re-appraisal cycle significantly impacts the decisions made by the agent.

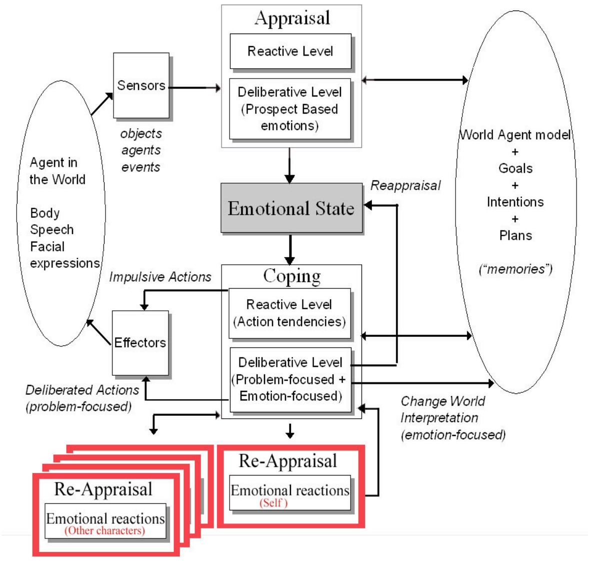


Figure 1. DAM architecture

4 – EVALUATING DOUBLE APPRAISAL

Evaluation of generative narrative is known to be very difficult and there is no agreed approach to doing so [20]. The subjective nature of storytelling is a major issue for the design of efficient and reliable evaluation procedures. Evaluating applications based on satisfaction and user experience is very different from the usual task oriented evaluation designs and is therefore still very much an open research question [10].

Another issue arises from the emergent nature of the storytelling form. Depending on the agents' minds, moods and emotions, a story might not unfold in the same way twice making a direct comparative analysis difficult. The EN approach is character-based and is aimed at participation rather than spectating. It is therefore necessary to devise an evaluation framework that focuses on the characters' decisions and behaviour, rather than 'the' story displayed. However combining a participant/spectator perspective in evaluation supports a direct comparison of data from both participant and spectator users.

4.1 Evaluation set

In this evaluation, the original FearNot! agent framework without any double appraisal has been used as a benchmark against which the implementations DA (DA.1/DA.2[1]) and DAM (DAM.1/DAM.2*) have been compared. The scenarios are composed of interacting agents who act a role and have their own personalities and goals and a Game-Master (GM) whose aim is to provide narrative events and make decisions about the world environment (outcome of physical actions, entry of new characters, removal of characters etc). In this implementation, the role of the Game-Master is played by a disembodied agent dedicated to story management. Like the actors, the Game-Master agent has been extended by DA and then by DAM. The combinations of different types of agents and Game-Masters resulted in 25 simulations. These simulations were all run with identical configuration setups and resulted in 5 different story-variations of the same scenario with identical configuration set ups.

The simulation plan [Table 1] reflects the narrative elements necessary for the development of an EN scenario (i.e. characters and game-master) and shows the appearance of story variations across the different simulations. It also includes different versions of the GM. For the purpose of this evaluation, different versions of the GM (i.e. DA, DAM) were also implemented, just as for characters, in order to test the validity of both DA and DAM for an agent playing the GM role.

	GM v1.0	GM DA.1	GM DA.2[1]	GM DAM.1	GM DAM.2[1]
FAtiMA v1.0	S1	S2 Story 1	S3	S4 Story 2	S5
FAtiMA DA.1	S6	S7	S8	S9	S10
FAtiMA DA.2	S11	S12 Story 3	S13	S14 Story 4	S15
FAtiMA DAM.1	S16	S17	S18	S19	S20
FAtiMA DAM.2	S21	S22	S23	S24 Story 5	S25

[Table 1] Simulation plan and story repartition

4.2 Evaluation methodology

For this evaluation, we reduced the output of the stories created by the software to a text form (actions and speech actions) to avoid graphic quality or specific user interaction modalities influencing the outcome. Stories record the interactions between characters and were generated by the software itself. **[Table 2]** shows an example. The stories were presented to a test-audience whose reactions, dramatic perceptions and judgment of dramatic intensity were documented with respect to character-based actions and plot events.

The evaluation plan designed for this application was composed of 5 different tests that aimed at assessing the dramatic values of the stories generated by the system. The first two tests (T1, T2) assess stories from a spectator perspective by presenting the user with a set of stories and asking them to mark and rank them by order of preference. Although T1 and T2 display the same stories to their test audience, these are slightly modified in T2 so all are of the same length. This is to establish whether the length of stories plays a role in the marking or ranking by the user. The final three tests (T3, T4 and T5) aimed at assessing stories from a participative perspective and presented the users with the possible game-master (T3) and character (T4, T5) decisions at every cycle allowing them to choose for themselves what would happen. These stories, like their counterparts in T1 and T2 are then marked by the user. When the marking/ranking has been executed, the users are

[1] Note that both implementations have two entries in [Table 1] since they present two slightly different versions (i.e. small and high ranges of pre-selected eligible actions (cf. section 3.2.1)). The same versioning design applies to the different implementations of the game-master (i.e. GM v1.0, GM DA.1/DA.2, GM DAM.1/DAM.2).

given further indications on the character's motivations and are asked about their decisions. This part of the evaluation could be related to the de-briefing session common to Role-Playing Games (RPGs).

Agent	Action
Colonel	Let's be clear about what we are all about to do! No one has ever been down there! Our intelligence reported this site has a potential threat to our land security! We all know why we are here today! The technology possibly hidden in there is all that matters to us. In the unlikely event of an encounter of any type, we are to wipe this place down and make sure no one or nothing ever come out of this temple! Dead or alive!
Colonel	God bless you all. Military personnel in formation, others behind me, keep an eye for traps, and loose sight of each other. All right, let's go!
The party	Following your order Sir!
Sergeant	Colonel! Here! Here come here. I have something odd here; it looks like a metal door with strange writings on top of it!
Colonel	Professor! Are these hieroglyphs there above the door say anything of what might be behind it?
Professor	Hum Yes Colonel!, Well, this is strange, these do not appear to be conventional hieroglyphs! There are actually two sets of text there. One that can be interpreted as a death threat to any mortal disturbing the lizard gods, no idea whose these can be! The other one although looks like Egyptian hieroglyphs contains many symbols I have never encountered and does not make any sense to me I am afraid!
Colonel	Ok, Everybody step back! We are going to blow this one up and see what it is hiding. Bellini, MCLean hold assault position!
Professor	Colonel, this temple is thousands of years old, this door is magnificent and such artefact has never been discovered before! Surely we can't just blow it up, we need to find a way to open it or leave it as it is. This is an archaeological wonder!
Colonel	I am not sure you are getting the whole picture there Professor! Right here and right now I am in charge! You do what I tell you to do when I tell you to do it!
Colonel	Destroys the door and the door opens

Table 2 An example of story generated (Story 1)

The evaluation methodology has been designed in order to achieve the aims summarized in **[Table 3]**.

Aim	Description
1	Determine which story is judged most interesting by the test audience (spectators)
2	Determine if the length of the story is a factor in determining its dramatic factor and general level of interest
3	Rate the meaningfulness/interest of agents and game-master actions/decisions from a spectator perspective
4	Determine whether a better understanding of the characters and roles would influence the ranking and marking of stories
5	Determine which story would be generated by the user if given authorial powers
6	Determine which story is judged most interesting by the test audience (interactive users)

Table 3 Evaluation aims and objectives

5 RESULTS AND CONCLUSIONS

In so far, the evaluation has been carried out on a total of 47 subjects with a 68 – 32 ratio between males (68.1%) and females (31.9%). The results presented herein should be interpreted as early results as the full data analysis for the entire scope of the evaluation was not yet available at the time of this article's submission. The results have however all been subjected to an analysis of variance (ANOVA) and are statistically significant within the evaluation test batches. The probability of insignificance (p) and degree of significance (%R) are indicated for each result.

5.1 Evaluation pointers

As with every evaluation process, it is essential to identify pointers that would indicate whether or not a given hypothesis possesses some tangible truth. In the case of this evaluation, we have identified a series of questions [Table 4] that require to be answered positively in order to demonstrate the validity of our approach. This list is not exhaustive by all means and focuses on the main aspects of the double appraisal theory (i.e. Dramatic efficiency, and comparison of the two implementations). It covers the basis for a more complete data analysis.

Evaluation question	Analysis pointer
(Q1) Does a double appraisal mechanism contribute in generating stories dramatically more interesting than if generated by a simple appraisal mechanism?	**(P1)** Story 1 (original FearNot!) should ranked and score lower than stories 2,3,4,5 (generated via double appraisal)
(Q2) Does an implementation considering the Emotions of all characters better at generating interesting stories than one only considering one character (self)?	**(P2)** Based on our assumption that DAM is potentially more complete than DA, Story 4 should score lower than Story 5.
(Q3) Is the consideration of all characters in a double appraisal contributes in generating overall more interesting stories?	**(P3)** Story 5 should score high on dramatic marking since it incorporate a double appraisal mechanism that takes into consideration all the characters of the scenario for both agents and game-master.

Table 4. Evaluation pointers and questions

5.2 Results

Q1

The overall story ranking (before debriefing) shown below in [Figure 2] provides elements of answers to Q1. The results have been provided by the test T1 and T2 and reflect a spectator's perspective on the ranking of our 5 stories. Whilst it shows a high ranking for story 3 (to be acknowledged in section 5.2.3), it also shows a poor ranking for Story 1.

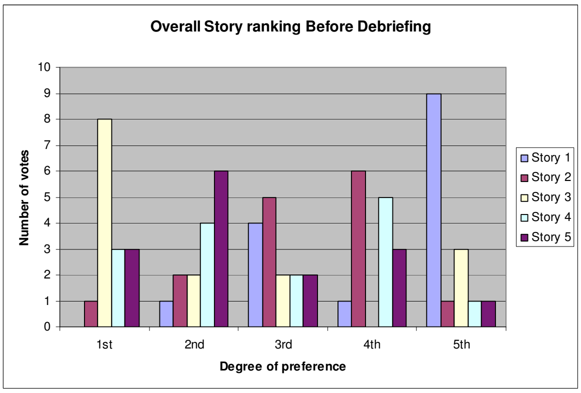


Figure 2. Overall Story ranking before debriefing

The story generated by FAtiMA did not perform well in the spectator ranking and has been perceived as the worst story of the test batch. This trend is also confirmed in [Figure 3] (p = 0.00061/ 99.39 %R) where individual story rankings have been translated into values in order to get a clearer picture of a story performance (averaging). This diagram shows to which extent Story 1 has been negatively perceived by spectator/reader users.

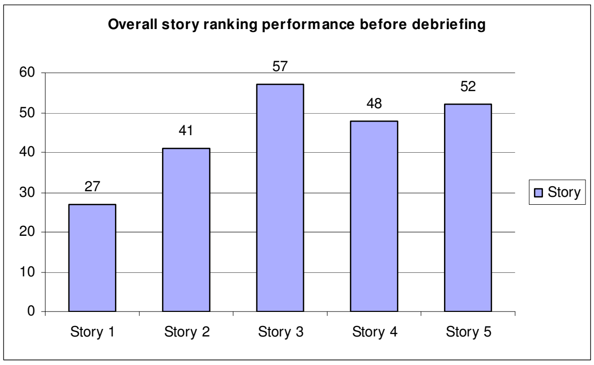


Figure 3. Overall Story ranking (points table)

Note that there are no significant differences in performance for story 1 between pre and post debriefing markings by users.

The results presented in this section indicate clearly that the single appraisal-based implementation (SA) scored lower than its double appraisal-based counterparts (DA/DAM).

On another hand, whilst the DAM.2 implementation of the game-master generated a different story (Story 2) than the original SA-based approach (Story 1), its counterpart in DA did not make any difference on the outcome of the scenario and still resulted in Story 1. The two stories using the SA-based agents (Story 1 and Story 2) score also sensitively lower than agents fitted with either DA (Story 3, Story 4) or DAM (Story 3, Story 4 and Story 5).

Q2

The results presented in this paper also show that agents or game-masters conforming to DAM tend to score higher than the ones conforming to DA. [Figure 3] demonstrates this by showing that Story 2 (game-master DAM) scores better than Story 1 (game-master DA). On another hand, the results detailed in [Table 1] indicate that they are no major changes in the actions of the agents unless they are interacting with a game-master of type DAM. The distinction between the two implementations discussed herein can however still be highlighted by the performances of stories 4 and 5. Both stories whilst, they feature the same version of the game-master DAM, present agents of the two different implementation types (DA = Story 4 and DAM = Story 5]).

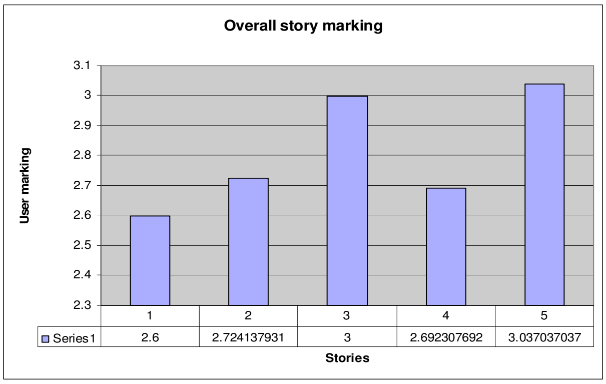


Figure 4. Overall Story marking

Both [Figure 2] and [Figure 3] show that overall, Story 5 outperformed Story 4 in the spectator/reader user ranking. This is further confirmed in [Figure 4] (p = 0.0917/ 90.83 %R) where the overall marking by all users (i.e. spectator/reader and interactive user) shows a net difference of appreciation between Story 4 and 5 in favor of the latter.

Q3

The results calculated for Q3 are interesting in the sense that two opposing claims could be regarded as significant in answering this particular question.

> Claim 1: [Figure 3] seems to indicate a better performance and appreciation of Story 3 over Story 5.
> Claim 2: [Figure 4] shows that Story 5 is the preferred story from a marking perspective.

The interpretation of these results alone is not sufficient for us to claim that the consideration of all characters in a double appraisal contributes in generating overall more interesting stories (Q3). It is necessary at this point of our analysis to focus on the nature of the tests performed in order to get a clearer idea of the validity of each claim. Claim 1 is based on spectator/reader user types whilst Claim 2 relies on interactive users. It is important to regard the marking for both perspectives (i.e. spectator/reader and interactive user) in order to make an educated decision on the validity of each claim.

[Figure 5] (p = 0.0068/ 99.32 %R) shows the overall story marking for non-participant users (Spectator/reader). It confirms, to a certain extent the results observed in [Figure 3] (Story 3 ranked better than Story 5) and shows that Story 5 is not the story

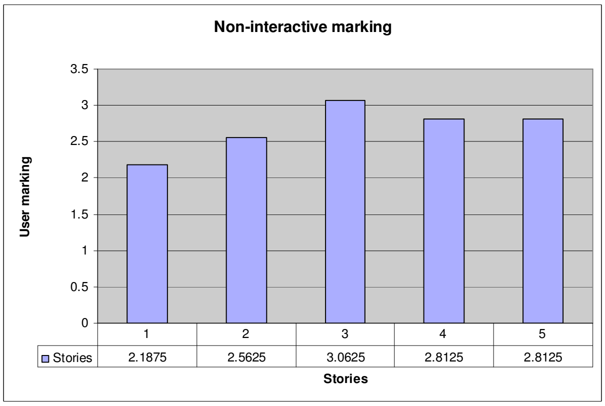


Figure 5. Non-interactive story marking

receiving the better marks. It therefore contributes negatively to the hypothesis developed in this paper that a double-appraisal mechanism considering all the characters of a given scenario performs better than both its self-centered counterpart and a single appraisal mechanism.

On another hand, [Figure 6] ($p = 0.0185/ 98.15$ %R) presents another picture by showing a net marking advantage for Story 5 over the rest of the stories by interactive users.

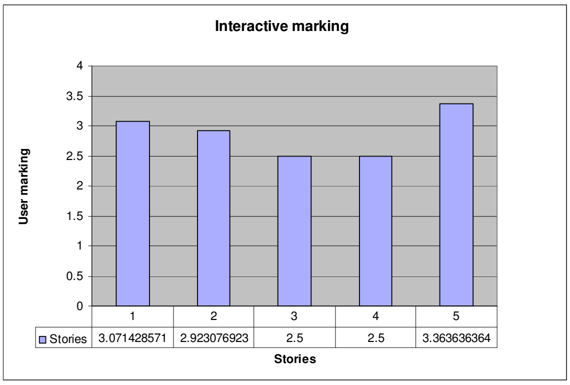


Figure 6. Interactive story marking

It is also interesting to notice in **[Figure 6]** the high marking performance of Story 1. This reinforce some of the claims made in [1] that an emergent narrative might not be perceived as interesting from a spectator/reader perspective as it would be from an interactive perspective.

In consideration to Q3, since the aim of this work is to produce interactive emergent narrative, we could understandably consider Claim 2 rather than Claim 1 as being the most significant for our results in the scope of this evaluation.

6 CONCLUSION

In this paper, we have demonstrated that synthetic characters can be enhanced to perform as actors by carrying out a second appraisal of their projected actions. The results presented herein show that the implementations proposed to extend an emotionally-driven agent architecture applied to the creation of emergent narratives (FearNot!) have positive impacts on the perceived dramatic values of the generated stories. Whilst these implementations were not equally as good in generating dramatic interest for the user (i.e. both spectator/reader and interactive user), they still produced simulations that scored higher than the original single appraisal-based architecture. On the basis of a direct comparison between the two different implementations carried out, DAM which considered the emotions of all of other characters in a scenario in order to make dramatic choices scored consistently higher than the more self-focused DA. This leads us to consider that DAM possesses a stronger dramatic potential than DA.

Finally, when comparing user marking for all stories, Story 5, which features DAM in both its agents and game-master architectures, scored the highest overall mark and was considered as the most interesting story to experience by interactive users. The results presented in the previous section show the validity of our approach and establish firmly our belief that narrative control can be exercised at character level in a distributive manner with satisfying results as long as the agents (i.e. characters) are provided with a mechanism that allows them to assess the emotional consequences of their actions on others.

This work is part of a larger theoretical work that has been investigating the emergent narrative concept for several years. Whilst significant, the results presented in this paper should be regarded as an early insight of what the overall evaluation process should come to deliver once the analysis of the data collected completed. Further work will consist in measuring the reactions, decisions and motivations of the participants in both marking and ranking the stories (spectator/readers and interactive users). Data will also be analyzed with regard to the dramatic weight associated to particular actions of the scenario and their potential impact on the user rating/marking. Finally, further theoretical work will investigate the areas of real-time narrative control, character-based narrative authoring and emergent narrative user interaction interfacing.

This work could also be extended to look at emotional trajectories rather than one-shot double-appraisal by considering sequences of planned actions rather than the goal-achieving action as at present. This would allow actors to explicitly consider the issue of dramatic climaxes.

REFERENCES

[1] Aylett, R.S & Louchart,S (2003) Towards a Narrative Theory for Virtual Reality. Virtual Reality 7: 2-9 Springer

[2] Aylett, R.S; Louchart, S; Dias, J; Paiva, A; Vala, M; Woods, S. & Hall, L. (2006) Unscripted Narrative for affectively driven characters. IEEE Journal of Graphics and Animation. May/June 2006 (Vol. 26, No. 3) pp. 42-52

[3] Boal, A. (1999) Legislative theatre: Using performance to make politics. Routledge.London 1999

[4] Braun, N. (2002) Automated narration – the path to interactive story-telling. Proceedings, NILE02, Edinburgh

[5] Cavazza M, Charles F, & Mead S. (2001) Agent's interaction in virtual storytelling. International Conference on Virtual Storytelling 2001: 145-154

[6] Dias, J and Paiva, A. (2005) Feeling and Reasoning: a Computational Model. 12th Portuguese Conference on Artificial Intelligence, EPIA 2005. Springer. pp 127 – 140.

[7] Elliot C.:"The Affective Reasoner". (1992) - A process model of emotions in a multi-agent system". PhD Thesis, Illinois

[8] Gratch,, J & Marsella,S (2005) Lessons From Emotion Psychology For The Design Of Lifelike Characters. Applied Artificial Intelligence 19(3-4): 215-233

[9] Hayes-Roth, Barbara, R. van Gent, and D. Huber (1997) 'Acting in Character' in R. Trappl and P. Petta (ed.) *Creating Personalities for Synthetic Actors*. Springer

[10] Knickmeyer, R and Mateas, M (2005*)* Preliminary evaluation of the interactive drama facade. Conference on Human Factors in Computing Systems.2005

[11] Laird, J. (2001) It knows what you are going to do: adding anticipation to a Quakebot. Autonomous Agents, 385-392, ACM, Montreal

[12] Lazarus, R.S& Folkman, S. (1984). Stress, appraisal and coping. New York: Springer

[13] Louchart, S, & Aylett, R.S: (2003) Solving the Narrative Paradox in VEs - Lessons from RPGs. IVA2003, LNAI 2792 Springer 2003 pp244-248

[14] Marsella, S. C., Johnson, W.L, LaBore, C, (2000) Interactive Pedagogical Drama, Proceedings of the fourth international conference on Autonomous agents, 2000.

[15] Mateas, M & Stern, A (2000) Towards Integrating Plot and Character for Interactive Drama - Working notes of the Social

Intelligent Agents: The Human in the loop, AAAI Fall Symposia 2000

[16] Murray, J. (1998). Hamlet on the Holodeck: The Future of Narrative in Cyberspace. MIT Press, Cambridge, MA.

[17] Ortony, A; Clore, G. L. and Collins, A. (1988) The Cognitive Structure of Emotions, Camb. Univ. Press 1988

[18] Prada, R; Isabel Machado, Ana Paiva (2000) TEATRIX: Virtual Environment for Story Creation. Intelligent Tutoring Systems 2000: 464-473

[19] Riedl, M & R. Michael Young (2004) An intent-driven planner for multi-agent story generation. 3rd Int. Conf. on Autonomous Agents and Multi Agent Systems, July 2004.

[20] Riedl, M & R. Michael Young (2005) An objective character believability evaluation procedure for multi-agent story generation systems IVA 2005, Springer LNCS 3661 pp278 – 291, 2005

[21] Rogers, Y., Scaife, M., Gabrielli, S., Smith, H. and Harris, E. (2002) A conceptual framework for mixed reality environments: Designing novel learning activities for young children. Presence, 11 (6), 677-686.

[22] Rousseau.D & Hayes-Roth,B (1997) A Social-Psychological Model of Synthetic Characters, Knowledge Systems Lab, Report No. KSL 97-07, Stanford University 1997

[23] Schoppers, M.J (1987) Universal Plans for Reactive Robots in Unpredictable Environments - Proceedings of the Tenth International Joint Conference on Artificial Intelligence (IJCAI-87)

[24] Swartout, W and Gratch, J and Hill, R and Hovy, E and Lindheim, R and Marsella, S and Rickel, J and Traum, D. (2005) - Simulation meets hollywood. Multimodal Intelligent Information Presentation, Eds. Oliviero Stock and Massimo Zancanaro.

[25] Whiten Andrew (1991), editor. Natural Theories of Mind. Basil Blackwell, Oxford, UK

FearNot! An Anti-Bullying Intervention: Evaluation of an Interactive Virtual Learning Environment

Scott Watson[1]**, Natalie Vannini**[2]**, Megan Davis**[1] **, Sarah Woods**[3]**, Marc Hall**[4]**, Lynne Hall**[4] and **Kerstin Dautenhahn**[1]

Abstract. The eCIRCUS (Education through Characters with Interactive Role-playing Capabilities that Understand Social interaction) project aims to develop an anti-bullying software, FearNot!, and evaluate its effectiveness in the classroom. This paper presents findings from two evaluations conducted during the 2006 National I-Power-I Anti-bullying Conference for Young People. Participants interacted with FearNot! v.1 (*scripted* version) and then either completed a short questionnaire (in Study 1) or took part in focus groups (in Study 2) evaluating the difference between two versions of FearNot! (*scripted* versus *unscripted*). Overall the results suggest that perfect graphics are not necessary for users to engage empathically with autonomous agents, and that the virtual characters did evoke emotional reactions. It is concluded that development of the FearNot! demonstrator is progressing well and that FearNot! will be a useful and engaging intervention against bullying in primary schools.

1 INTRODUCTION

1.1 Bullying in Primary Schools

Defining bullying and victimisation behaviour is difficult due to its complicated nature. However, a common definition states that "a student is being bullied or victimised when he or she is exposed repeatedly and over time to negative action on the part of one or more other students" [1]. Furthermore, most bullying behaviour can be grouped into one of three categories [2]:

- direct physical bullying - e.g. pushing, hitting, kicking, and stealing belongings.
- direct verbal bullying - e.g. name calling, teasing, and threatening.
- indirect (or relational) bullying - e.g. social exclusion, rumour spreading, withdrawal of friendships.

In the same way that bullying styles can be categorised, the roles taken on by children involved in acts of victimisation can also be categorised. The most significant roles are: the 'pure' bully, the 'pure' victim, the bully-victim (someone who bullies others and is bullied themselves), the bully-assistant, the bystander/neutral, and the defender (of the victim) [3], [4].

While studies report varying prevalence rates, bullying is acknowledged as a cross-cultural problem which can affect between 8% to 46% of primary age school children [5]. Bullying is a serious issue as victims can continue to show psychological problems (e.g. anxiety, depression) even after the bullying has ceased. In extreme cases victimisation can lead to psychiatric referral [6] or even suicide [7].

1.2 Current Bullying Interventions

Having examined the extent of bullying, many studies have attempted to demonstrate effective interventions against victimisation. Due to the complex interaction between bullying styles, coupled with the different roles that children may take, there is a large number of interventions that have been proposed. These include approaches which emphasise the role of the bully individually, the role of bully and victim together, and even whole schools [3].

Smith & Madsen (1997)[8] found that one third of schools in the UK have a specific anti-bullying policy, but Woods & Wolke (2003)[9] have shown that these measures are often ineffective against direct bullying, and can even lead to an increase in relational victimisation. As a result, Woods & Wolke (2003)[9] suggest that "individualised strategies may help to take the differential needs of bullying roles into account". Unfortunately, there currently appears to be few or no interventions which provide such individual education about anti-bullying coping strategies directly to children involved.

1.3 FearNot! as an Innovative Intervention

One potential medium for providing cheap, safe, and individual advice on coping with bullying could be a Virtual Learning Environment (VLE) which is populated by Intelligent Virtual Agents (IVAs).

FearNot! (Fun with Empathic Agents Reaching Novel Outcomes in Teaching) is such an application. FearNot! provides 8-11 year old children with the opportunity to visit a virtual school environment complete with characters representing the most significant roles in bullying (bullies, victims, assistants, bystanders, and defenders), locales (playground, classroom, library, and local streets), and scenarios (direct and indirect victimisation) that are commonplace in real-life bullying incidences. Characters in FearNot! are autonomous agents capable of making their own decisions and acting out their

[1] School of Computer Science, Adaptive Systems Research Group, University of Hertfordshire, College Lane, Hatfield, Hertfordshire, AL10 9AB, UK. email: {s.e.j.watson, m.davis, k.dautenhahn}@herts.ac.uk

[2] Department of Pedagogical Psychology, Psychology Institute, Julius-Maximilians-University of Würzburg, Germany. email: natalie.vannini@psychologie.uni-wuerzburg.de

[3] School of Psychology, University of Hertfordshire, College Lane, Hatfield, Hertfordshire, AL10 9AB, UK. email: s.n.woods@herts.ac.uk

[4] AMAP, Edinburgh Building, University of Sunderland, Sunderland, SR1 3SD, UK. email: marc.hall@sunderland.ac.uk

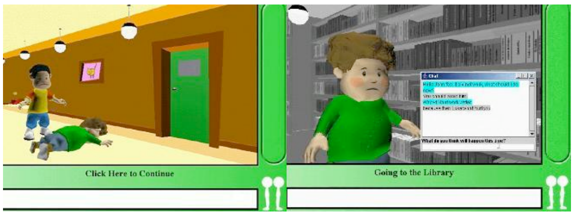


Figure 1. FearNot! v.1 Bullying and Interaction Episodes

own behaviours, thus leading to an emergent narrative as the storyline progresses [10]. Children interact with FearNot! on an individual basis by witnessing an emergent bullying episode, and then ‘talking to’ the victim character in order to advise them how best to cope in the following episode. The fundamental idea behind the FearNot! application is to allow children to try out various coping strategies without being directly involved themselves - the usefulness of a coping strategy can be learned safely and vicariously through the victim character’s experiences. In this way the user takes on the role of an invisible ‘peer buddy’, or friend, to the victim character. Support for this kind of approach - learning through activity and play in virtual environments is privided by Roussou (2004) [11].

The eventual aim is for FearNot! to be voluntarily adopted by primary schools as an addition to the UK’s existing Key Stage 2 Personal and Social Health Education (PSHE) curriculum. A German language version of FearNot! is also in development. The FearNot! prototype designed and evaluated during a preceeding EU Framework 5 project, VICTEC (Virtual ICT with Empathic Characters), was well accepted and reported e.g. [12]. Under the eCIRCUS project, though, FearNot! continues to be developed further - with an updated version made available for initial testing in 2006.

1.4 FearNot! Versions and Specifications

1.4.1 FearNot! v.1

FearNot! v.1 is an applet which runs within a webpage with the WildTangent(WT) Plugin(R). As a showcase demonstration, this version comprises three consecutive, *scripted* male bullying episodes with an interaction episode between each. During interaction, coping strategies can only be suggested to the victim character by means of a drop down menu. Follow-up questions are answered through free text (typed) input. The suggested coping strategy has no impact on events in a following episode. Once the three male episodes are completed three female episodes are also available.

1.4.2 FearNot! v.1.5

FearNot! v.1.5 is an intermediary version of FearNot! which improves on v.1, but is still in final development. This version is also an applet which runs within a webpage with the WildTangent(WT) Plugin(R), but boasts a number of improvements including new graphical and language specifications. The graphical design of the characters was changed so that they all wear the same school uniform instead of their own clothes, which improves validity for the UK where most primary schools require their students to wear a uniform. The language was also updated to include more colloquialisms and more valid dialect that is used by children within the target age group. A drop-down menu has been replaced by free text input during interactions, which now allows children to input their own ideas instead of forcing them to select from pre-set options. Open dialogue is a valuable research tool for understanding what children know about how to cope with bullying. Finally, the virtual characters are now able to act upon advice given by the user during an interaction episode, giving rise to an *unscripted* and *emergent* nature for the bullying episode. This version allows for a greater range of different user experiences. Only male episodes are available in this version.

1.5 The Current Study

While FearNot! v.1 was extensively investigated during the VICTEC project, the development to v.1.5 has not yet been evaluated. With the eCIRCUS project aiming to place FearNot! into schools for longitudinal investigation in 2007, it is imperative to ensure that the final version is ecologically valid - that the characters are believable and engaging, that the episode storylines are understandable and true-to-life, and that the overall user experience is fun and educational. This study aims to seek initial feedback about improvements to FearNot! made since the VICTEC project, and serves to demonstrate that FearNot! is still an innovative approach to a continuing problem.

In this paper we present findings from two studies conducted during the National I-Power-I Anti-bullying Conference for Young People held during November 2006 in Weston-Super-Mare, UK. While this setting may seem uncontrolled at first, one advantage of this approach is that it yields greater ecological validity since FearNot! is designed to be used in an unconstrained classroom environment. It also allows for an excellent cross-section of participants from schools across the UK which can differ in terms of achievement and socio-economic status. Study 1 evaluates user’s perception of FearNot! v.1, while Study 2 investigates user’s preference of the similarities and differences between FearNot! v.1 and v.1.5. Sections 2 and 3 of this paper present the methods and results of these studies respectively, while Section 4 provides an overall discussion of both studies and describes future directions for FearNot! and the eCIRCUS project.

2 Study 1

2.1 Method

In total 54 participants returned questionnaires. Of these 35 were male, and 18 were female (1 missing data point) with 14 respondents in primary school, 33 in secondary school, and 5 adults (2 data points missing). While the majority of participants stated that they were in secondary school, the investigators observed that these children were young enough to be comparable to FearNot!’s target age group.

Throughout the conference, laptops were used to simultaneously run four different instances of FearNot! v.1 at a stand accessible to all conference delegates. Respondents interacted freely and individually with FearNot!, but investigators were on-hand to answer questions and offer advice if necessary. Each interaction lasted approximately 15 minutes - long enough for participants to play fully through 3 related episodes. Once their interaction had ended, participants were asked to complete a short questionnaire and return it to one of the investigators.

The questionnaire used was adapted from the VICTEC project’s Character Evaluation Questionnaire (CEQ). This questionnaire asked about six items of interest:

- Most likeable character

- Least likeable character
- Character graphic design (5-point Likert scale from ‘Strange’ to ‘Good’)
- Which character looked best/which character looked strangest
- Storyline believability (5-point Likert scale from ‘Unbelievable’ to ‘Believable’)
- Estimated usefulness of FearNot! in Primary Schools (5-point Likert scale from ‘Not Useful’ to ‘Useful’)

2.2 Results

2.2.1 Likeability of FearNot! Characters

The most likeable character was John - the male victim, while the least likeable character was Luke - the male bully. This pattern is also repeated for the female characters where Frances (the victim) is the most likeable character, and Sarah and Janet (the bullies) are liked the least (Figures 2 and 3). This suggests that the characters are evoking the kind of empathic reactions that they were designed to evoke.

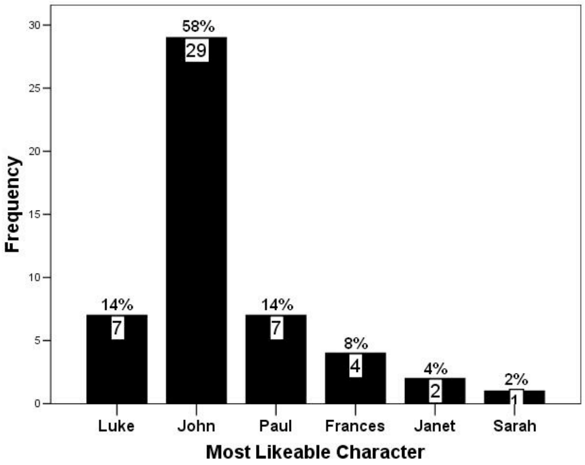


Figure 2. Most Liked FearNot! v.1 Characters (n=50)

Although it appears that the male characters are generally more well liked than the female characters this may be due to the simple explanation that more participants interacted with, and therefore gave more ratings of, the male episodes than female episodes. This explanation is upheld by the fact that the male characters receive more ratings on both the most likeable *and* least likeable scales.

2.2.2 Graphical Design of FearNot! Characters

With regards to the graphical presentation of the characters, Luke and John were jointly rated as the best looking designs, while John was also rated as the strangest character in appearance. From the female characters Frances and Janet were rated as the best looking designs, with Frances also rated as the strangest (Figures 4 and 5). This pattern (that the same characters were chosen as demonstrating both the best and strangest design) could be explained by the fact that these characters are the main protagonists in the story, and so have the greatest on-screen time. Another cause, however, could be due to the phraseology of the questionnaire which asked participants to nominate the ‘best looking’ and ‘strangest looking’ characters. It is possible that characters which ranked highly on both questions were thought to have been drawn well, but that the actual design was disliked - e.g. John is portrayed as slighly over-weight, and Frances wears glasses; both of which can be used to tease victims of bullying.

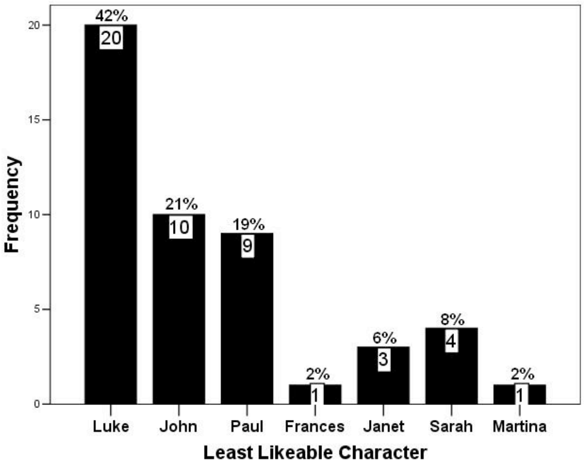


Figure 3. Least Liked FearNot! v.1 Characters (n=48)

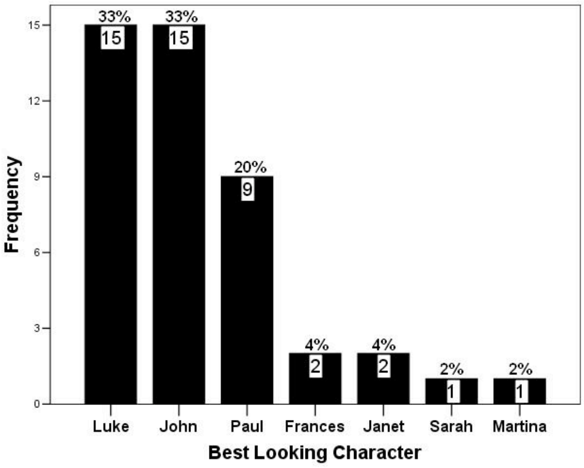


Figure 4. Best Looking FearNot! v.1 Characters (n=45)

2.2.3 Overall Impressions of FearNot!

While it is necessary to look at the characters in isolation, it is also of the utmost importance to evaluate the user’s general impression of FearNot! The current sample rated the overall graphical presentation as above average, with high ratings for storyline believability and usefulness in primary schools (Figure 6). Taken together, these findings are positively in favour of the validity and realism of the FearNot! episodes, and also show that the application has great educational potential . Given that the target age group comprised only a small proportion of the overall sample, the final analysis was re-run using data from just the primary school age participants. The results from this sub-set are quite similar to those of the whole sample. The

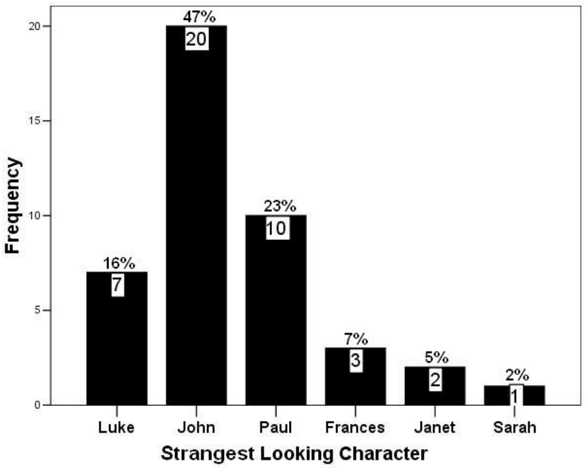


Figure 5. Strangest Looking FearNot! v.1 Characters (n=43)

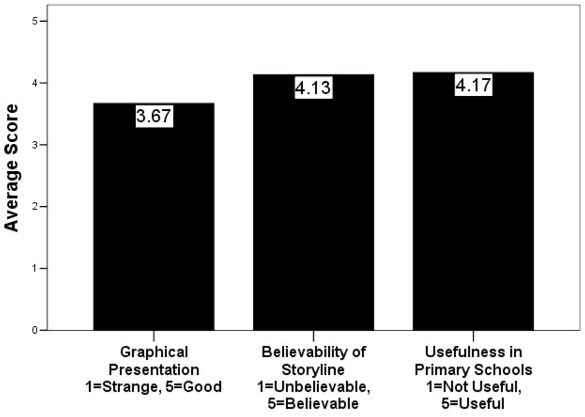


Figure 6. Overall Impression of FearNot! v.1 (n=54)

graphics were again rated as above average in quality, with storyline believability and usefulness in primary schools both scoring highly (Figure 7). These findings are especially useful as they provide great support for the FearNot! application directly from the user group it is aimed at.

Gender differences show that females liked the graphical presentation more than males, while males found the storyline more believable and rated FearNot!'s classroom usefulness as higher than females (Figure 8). These results can be explained by the observation that males are more likely to interact with video games in everyday life, and so will expect higher standards for graphical presentation and will be more open to using such an application at school. That girls found the storyline less believable could be due to the fact that most participants interacted with the male episodes as opposed to the female episodes - naturally these episodes are less relevant to females. Unfortunately the small size of invidual groups did not allow for deeper inferential analysis.

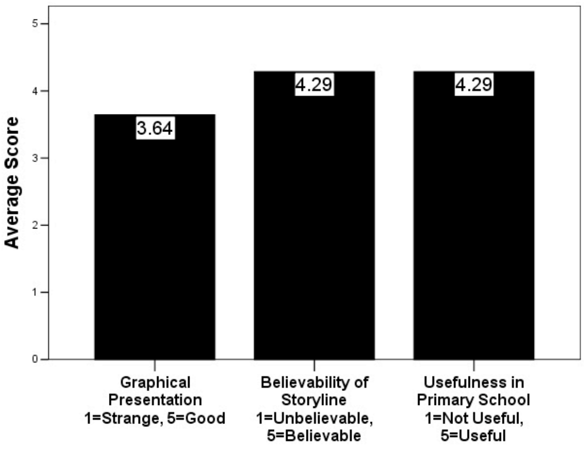


Figure 7. Primary School Children's Overall Impression of FearNot! v.1 (n=14)

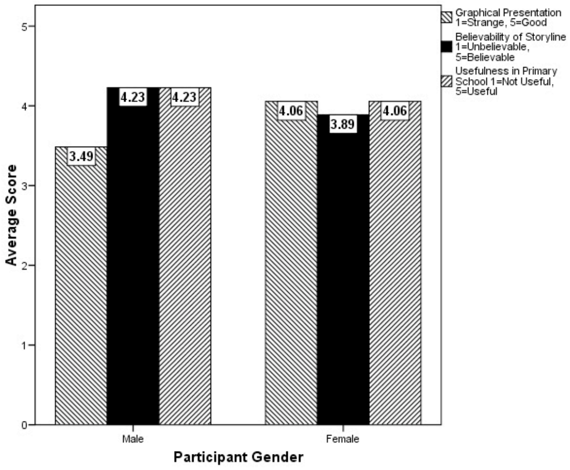


Figure 8. Gender-Split Overall Impression of FearNot! v.1 (n=54)

3 Study 2

3.1 Method

45 participants attended a FearNot! workshop run as part of the anti-bullying conference. This sample's demographics were similar to those from Study 1. Participants interacted with FearNot! v.1 in groups of around 6 people to each laptop. This interaction lasted long enough to allow each group to experience both male and female episodes. After this interaction, participants were shown a prerecorded video of FearNot! v.1.5 which lasted approximately 5 minutes. Participants were then organised into four small focus groups, each led by an investigator, to discuss the two different versions of FearNot!. Topics of discussion were similar to those from Study 1's questionnaire, but preferences of the different versions of FearNot! were also drawn out.

3.2 Results

The results from the workshop's four focus groups are descriptive/qualitative in nature and give a first impression of FearNot! v.1.5 as well as serving to expand on the quantitative data obtained in Study 1.

The most liked characters were John and Paul (the male victim and defender) with Luke (the male bully) liked the least. John's and Frances' (the female victim) graphic design were considered to need the most improvement. The characters were able to elicit the kind of empathic engagement that they had been designed for - participants reported that they felt sorry for John and were angry at Luke, Janet and Sarah (the bully characters).

The storylines were generally well accepted with Frances' situation considered to be worse than that of John - presumably because of the relational nature of the bullying that Frances suffers, compared to the direct physical aggression that John is subjected to. This finding could be due to the sample. Because the participants were mostly of senior school age (12 years old and above), and slightly older than the target age group, it is possible that their more advanced cognitive development meant they were able to understand the relational bullying more easily than the target age group. In addition to this, the relational episodes were also considered more believable and realistic (when speaking to secondary school age girls) than the physical scenarios.

While the storylines were enjoyable and believable there was concern that the pacing was too slow and most participants agreed that longer, quicker-paced episodes would be more enjoyable. In keeping with the findings from Study 1, there was consensus that imperfect graphical design did not affect engagement.

FearNot! v.1.5 was greatly preferred to FearNot! v.1 in terms of graphic design (especially that characters now wore a school uniform which is appropriate for a UK setting), language used by the characters (though even more colloquialism/slang would be preferred by the target age group), storyline enjoyability, and interaction style. However, most participants reported that they would like even more interaction - specifically the ability to control their own personal avatar within the virtual environment. Many of the younger participants thought that FearNot! would be "better than normal" curriculum, that children "could learn from it" and that FearNot! "will make people think".

4 Discussion

In Study 1, victim characters were generally the best liked and the bully characters were liked the least. This shows that not only are human users willing to engage with virtual agents, but that the FearNot! characters are successful in eliciting the right kind of empathic and emotional reactions that are necessary for the user to experience a meaningful and educational interaction. While some of the graphical designs were considered to be strange, the overall quality of the graphical presentation was consistently rated as above average. In addition to this, the storylines presented were considered believable by both the whole sample, and the target age group in particular. The FearNot! application was thought to have great potential if included as part of existing primary school curriculum.

Interestingly, the graphical design of the characters seemed to have little impact on the user's rating of their believability or on the elicitation of empathy. For example, while the male victim was rated more often as the strangest looking character than the best looking character, he was also rated as the most likeable character. Taken with Study 2's findings that refined graphic design is preferred, this pattern of results suggests that excellent graphical design is not necessary to create an engaging experience as long as characters act in a believeable manner. However, graphical presentation can provide the 'icing on the cake' for an engaging VLE.

Study 2 corroborated these findings and provided further depth. Participants felt sorry for the victim character, and were angry at the bully characters. The relational episodes were seen as more serious than the physical episodes. This was thought to be due to the cognitive development of the sample, which would be in keeping with the suggestion that the understanding and use of relational bullying requires more advanced social cognition [3]. It would be interesting to investigate this further with specific reference to age differences in understanding of different bullying styles. The most positive finding to emerge from Study 2 was the consensus that FearNot! v.1.5 was preferred over v.1. This shows that the changes made to graphics, character language, and interaction style all affect the user's experience in a positive manner and improve engagement and enjoyability.

This study's methodology could be criticised for being too informal in nature. However, it is argued that the informal methodology of this study does show a number of advantages. While FearNot! is not designed to be used in the conference environment that this study took place in, the method does not lack ecological validity entirely. FearNot! is to be used in primary school classrooms with little teacher input. In this sense, the current study closely fitted this setting in terms of amount of adult supervision, background noise, and equipment (many primary schools in the UK prefer the flexibility that laptops offer over a rigid suite of desktop machines).

Given that the setting was not fully controlled, the results are strong and robust enough to demonstrate that FearNot! is successful in creating engagement and eliciting empathy even in less-than-ideal settings - this can only be a positive sign given that FearNot! will eventually be used in a quieter and more controlled school environment.

In addition to this, while there were many exhibitors at the conference, the FearNot! stand was consistently among the busiest and most popular with primary aged children and generated a great deal of interest in children and their guardians alike. Many children returned to the stand a number of times over and again - demonstrating that children actively *choose* to play FearNot! It must be acknowledged, however, that such positive outcomes could be due to a social desirability effect. Since the participants were all delegates of an anti-bullying conference it is safe to assume that they will already have a vested interest in this area, and will react positively to any potential intervention.

While mainly positive comments have come out of these studies, it was also shown that certain areas would benefit from some improvement. Most notably among these are the graphic design and language used by the characters. While the graphics have improved from FearNot! v.1 to v.1.5 there is thought to be still more room for improvement, especially when compared to commercial video games.

The findings taken from studies which utilise an informal and qualitative methodology are especially useful in the design of VLEs and IVAs as they allow developers to gain a more detailed understanding of their user's attitudes and needs than statistical approaches allow for. A number of recommendations about the development of FearNot! are also of relevance to the development of virtual environments in general.

Firstly, agent and environment believability can be improved by ensuring cultural similarity with target users. Study 2 also shows that, with regards to language issues, local and temporally relevant phraseology/colloquialisms can improve believability, as can accents for any audio output.

For virtual environments that also include a cohesive storyline, the issue of pacing must be taken into consideration. While it is beyond the scope of this study to demonstrate the effect of pacing on engagement, it is suggested that quicker paced but longer lasting episodes are more engaging than shorter and slower episodes - at least for a younger audience.

Many respondents stated that they would like to have 'more control' over a character within FearNot! It is thought that such interaction could lead to deeper immersion within a virtual environment, and even superficial interaction - such as selecting physcial characteristics of an otherwise unplayable agent - could lead to users identifying more with a given character. Some support for this claim could be found in the popularity of commercially available role-playing computer games. Because one of the fundamental ideas behind the FearNot! application is to allow children to try out various coping strategies without being directly involved themselves (the usefulness of a coping strategy can be learned safely and vicariously through the victim character's experiences), the inclusion of personal avatars is not possible in FearNot! However, it is an interesting issue which should be taken into consideration when designing a VLE, and is currently being investigated as part of the eCIRCUS project in the development of ORIENT - a VLE aimed at aiding refugee/immigrant integration into the host nation's school system.

A central aspect of the eCIRCUS ethos is 'user-centered design', in which target users are consulted iteratively on all aspects of a VLE's design. A further advantage of using an open methodology similar to that employed in this study is that it allows for a more varied sample to participate and become involved in the design of a VLE. While the VICTEC project allowed children to become involved in the design of FearNot! this study has now also given teachers and adults the opportunity to contribute toward FearNot!'s implementation. Furthermore, teachers and educational experts will play a larger future role with regards to the development of educational materials which will support the use of FearNot! as a classroom tool.

The final version of FearNot! is currently undergoing technical development. This version runs under the .net framework, and makes use of the Ogre3D graphical environment. Some major developments will include improved graphical design (such as fully motion-captured animation) [13], and more natural speech/audio output between characters (voices will be recorded by professional voice-artists, and the language and grammar will be generated and checked by a team including native English speakers who are familiar with the accents and linguistic nuances in the geographical areas in which FearNot! will be evaluated). A sophisticated text-recognition engine will be trained for use with younger users to allow full-text (typed) interactions. More characters, locations, and bullying incidences will be included to ensure a more believable and engaging experience. Finally, the characters will be much more responsive to the user's input.

The characters themselves are also undergoing development: More believable character actions and behaviour will be achieved by integrating an affective appraisal system which includes flexible management of goals [14]. This system will be further bolstered by a simplified version of the model of autobiographic memory devised by Ho and Watson (2006)[15].

This version of FearNot! will be piloted in schools during early 2007, along with a number of psychological evaluations. These include measurements of participant roles, children's knowledge about bullying and coping strategies, their empathic abilities, and moral disengagement. Once any necessary changes are made to either FearNot!, the psychological measurements, or the accompanying curriculum, a large-scale (900 children) longitudinal (6 week) intervention will be evaluated in primary schools in the UK and Germany to assess the impact of FearNot! on incidences of bullying and the children involved.

5 Conclusion

The final conclusions that can be taken from the current studies are positive for FearNot!. Although certain aspects, such as graphical design, still require further refinement, this does not interfere with storyline believability or the user's ability to empathise with the characters. The FearNot! application is well received by children and adults alike as an innovative, engaging and educational intervention against bullying. This conclusion will be fully investigated during 2007, when the final version of FearNot! is placed into primary schools in the UK and Germany for a large-scale longitudinal evaluation. Recommendations for the success of other VLEs include ensuring cultural relevance, appropriate pacing of a storyline, and allowing users greater control in the environment. Finally, agents who behave in a believable manner are more engaging than attractive graphical presentation.

ACKNOWLEDGEMENTS

The work described in this paper was condcuted within the EU Framework 6 project eCIRCUS ("Education through Characters with Interactive Role-playing Capabilities that Understand Social interaction") (www.e-circus.org) and was funded by the European Commission under contract number IST-4-027656-STP.

REFERENCES

[1] Olweus, D. (1999). Norway. In P.K. Smith, Y. Morita, J. Junger-Tas, D. Olweus, R. Catalano, & P. Slee (Eds.), *The nature of school bullying: A cross-national perspective (7-27)*. London: Routledge.

[2] Wolke, D., Woods, S., Schulz, H., & Stanford, K. (2001). Bullying and victimisation or primary school children in South England and South Germany: Prevalence and school factors. *British Journal of Psychology*, **92**, 73-696.

[3] Wolke, D., & Stanford, K. (1999). Bullying in school children. In D. Messer and S. Millar (Eds.). *Developmental Psychology (341-360)*. London: Arnold.

[4] Salmivalli, C., Lagerspetz, K., Björkqvist, K., Österman, K., & Kaukiainen, A. (1996). Bullying as a group process: participant roles and their relations to social status within the group. *Aggressive Behavior*, **22**, 1-15.

[5] Wolke, D., Woods, S., Bloomfield, L., & Karstadt, L. (2001). Bullying involvement in primary school and common health problems. *Archives of Disease in Childhood*, **85**, 197-2001

[6] Kumpulainen, K., Rasanen, E., Henttonen, I., Almqvist, F., Kresanov, K., Linna, S.L., Moilanen, I., Piha, J., Puura, K., & Tamminen, T. (1998). Bullying and psychiatric symptoms among elementary school-age children. *Child Abuse & Neglect*, **22**, 705-717.

[7] Carney, J.V. (2000). Bullied to death. Perceptions of peer abuse and suicidal behaviour during adolescence. *School Psychology International*, **21**, 213-223.

[8] Smith, P.K., & Madsen, M. (1997). *A follow-up survey of the DFE anti-bullying pack for schools. Its use and the development of anti-bullying work in schools*. London: Department for Education and Employment.

[9] Woods, S., & Wolke, D. (2003). Does the content of anti-bullying policies inform us about the prevalence of direct and relational bullying behaviour in primary schools? *Educational Psychology*, **23**, 381-401.

[10] Aylett, R.S., Louchart, S., Dias, J., Paiva, A., & Vala, M. (2005). FearNot!: An experiment in emergent narrative. *In proceedings of IVA 2005: Intelligent Virtual Agents*, 305-316.

[11] Roussou, M. (2004). Learning by doing and learning through play: An exploration of interactivity in virtual Environments for Children. *ACM Computers in Entertainment*, **2**, 1-23.

[12] Hall, L., Woods, S., Aylett, R., Newall, L., & Paiva, A. (2005). Achieving empathic engagament through affective interaction with synthetic characters. *In proceedings of ACII 2005: Affective Computing and Intelligent Interaction.*, 731-738.

[13] Hall, L., Vala, M., Hall, M., Webster, M., Woods, S., Gordon, A., & Aylett, R.S. (2006). FearNot!'s appearance: Reflecting on children's expectations and perspectives. *In proceedings of IVA 2006: Intelligent Virtual Agents.*, 407-419.

[14] Aylett, R.S., Dias, J., and Paiva, A.(2006). An affectively-driven planner for synthetic characters. *In proceedings of ICAPS 2006: International Conference on Automated Planning and Scheduling.*, 2-10.

[15] Ho, W.C., & Watson, S. (2006). Autobiographic knowledge for believeable virtual characters. *In proceedings of IVA 2006: Intelligent Virtual Agents.*, 383-394.

A Mixed Initiative Authoring Environment For Emergent Narrative Planning Domains

M. Kriegel and **R. S. Aylett** [1]

Abstract. In this paper we present a novel interactive method of authoring planning domains for emergent narrative applications. We explain how the emergent narrative concept focuses on the interaction between autonomous agents and point out that one of the main tasks of an emergent narrative author is to design a planning domain for those agents. By reviewing existing authoring tools for interactive storytelling, we show that so far none of them has been applied to this particular task. We then describe the design of an authoring software that might be suitable to support a non technical minded author in creating planning domains in an intuitive manner. In the authoring process the author is stepping through a hypothetical storyline that is created both by the planner and by the author. The software extends and grows the planning domain by taking into account the way the author shapes the storyline and more importantly, the reasons the author gives for shaping it that way.

Introduction

Digital interactive narrative is a research field that has received growing attention during recent years. Various storytelling systems have been created that use a variety of approaches to create electronic narrative environments, in which the user can influence the unfolding of the story. However, there is clearly a perceivable mismatch between the great amount of academic, theoretical ideas and the very small amount of actual full-scale implementations of the interactive narrative concept that go beyond a small proof of concept. To put it in other words, there are lots of good ideas of how to build systems to tell interactive stories but almost no stories that are actually told. The problem, however, is that any interactive storytelling system can only be put into good use with a lot of story content. Facade[6], at the current date is the only implementation of interactive narratives, that has really striven to break through this content barrier. One reason for this lack of stories of course is that the interactive narrative community to this date consists mainly of computer science academics and resources for the implementation of complex stories are just not available in academia. This, however, is only an issue, because the content development for interactive narrative systems is both time consuming and complex, often requiring some programming skills. This complexity prevents traditional story authors with a non-technical background from creating interactive story content. Those problems might be tackled with authoring software to support the story content creation process. Ideally an authoring software is both accessible (i.e. easy to use) and productive (i.e. even speeds up the authoring process for expert users). In this paper we will introduce the emergent narrative approach to interactive storytelling and describe the tasks of an author in the emergent narrative framework. We will make an argument for the need of intelligent authoring tools and review existing authoring systems for interactive storytelling. Finally an intelligent authoring environment for emergent narratives will be suggested, in which simulation and authoring are intervened.

Authoring Emergent Narrative

As pointed out by [9], amongst the existing theories of how interactive storytelling should be approached, a distinction can be made between two main approaches: character-centered and plot-centered. While the former approach provides strong character believability, the latter one can guarantee more plot coherence. The holy grail of interactive storytelling seems to be a solution that guarantees both character believability and plot coherence. Emergent Narrative[5] can be assigned to the class of character-centered approaches. The idea behind it is that a story emerges from the interaction of believable autonomous virtual characters. Unlike plot-centered top-down approaches, where the course of a story is planned according to a narrative model of plot and where characters are merely puppets whose actions contribute to plot-level goals, in emergent narrative there is a planner for each character that plans the actions that the character is taking. This way, character believability is maximised, because characters are never forced to act out of character in order to achieve a plot goal. Thus, the authors main task is to configure the planners that drive the characters. Stories are created in bottom-up fashion by specifying the character's behavior.

The specific details of how the planner inside the characters artificial minds work can vary. We have implemented the emergent narrative concept in the educational interactive Drama FearNot, so we will assume an agent architecture similar to that of FearNot[1]. Configuring a planner means specifying a planning domain. A basic planning domain consists of actions and goals. Actions have preconditions and effects, both of which are logical descriptions of a world state. It is the planner's main task to assemble a sequence of actions that reaches a certain goal. Goals have preconditions that need to be fulfilled before the character can try to achieve that goal and success-conditions that indicate the world state, in which the goals is considered to be fulfilled. In FearNot the planner is also coupled with a simulated emotional system that helps the planner to prioritize goals and plans, depending on how the character feels about certain events, characters or objects.

While creating content for FearNot we noticed that it requires a long rethinking for people to specify story content in this way. It seems inevitable that authors think about interactive stories in terms

[1] Heriot-Watt University, UK, email: {michael,ruth}@macs.hw.ac.uk

of a variety of possible linear stories, instead of concentrating only on the characters. Being able to let go of the control of the story as an author is one of the key concepts that emergent narrative authors need to learn. We will investigate how authoring software can facilitate this process.

Related Work

Many researchers working on interactive storytelling have identified the need for authoring software and several tools have been developed usually specifically for a certain storytelling engine. A good overview of the tools available can be found in [8]. All authoring tools have in common that they ease data entry significantly for the designer/author compared to hand-coding. Regarding their appearance and user interface many interactive storytelling authoring tools [8, 14, 10] are similar to the editors of video games (e.g. Unreal Tournament, Neverwinter Nights or Warcraft 3) or storyboarding tools like Kar2ouche or Mediastage[4]: The author uses those tools mainly to create the 3d-environment and to place objects and characters in the environment. Additionally to just arranging the environment, most of these tools also include some storytelling features to allow the creation of branching story lines, triggers, plot segments, etc. While those functions are both helpful and necessary if the interactive narrative is visualised graphically, they do not facilitate the configuration of intelligent characters, which is the main task of an author of emergent narrative. However a smaller number of tools are a bit more unconventional and contain some ideas that an emergent narrative authoring tool might benefit from:
DraMachina [3] supports authors in annotating a linear story with meta-information to identify important story elements (entities like characters, scenes or objects, actions, etc.). In Thespian [11], a character based storytelling system, authoring can also be done by feeding linear stories into the system, but here the system automatically extracts information from those stories, whereas in DraMachina, the author has to participate in the process of extracting data from the linear input stories. Thespian uses a fitting algorithm to adjust parameters that define a character's behavior. The target function of this fitting algorithm is the degree of similarity between the simulated stories and the linear stories that are fed into the system as training data. In other words, the author provides the system with examples of how characters behave in certain situations (stories), and the system tries to generalize the character's behavior from those examples. Unfortunately this approach cannot be directly applied to solve the emergent narrative authoring task, because although it helps to create a cast of characters with distinct personalities it does not help in creating the planning domain that is necessary for those characters to act at all. Jim Thomas and Michael Young[13] describe another interesting approach to authoring interactive stories. In their idea of an *author in the loop*, the author participates in the planning process (mixed initiative planning). With this system, while the author is testing and adjusting the story world, they would have a number of sliders at their disposal to modify their story preferences while the planner is constructing a story. This is similar to a sound engineer mixing several sound sources in real time. Unfortunately also this method requires a complete planning domain and does not facilitate the construction of a planning domain in the first place. Finally, it is worth noting that since we are essentially talking about authoring planning domains, a lot of relevant work has been carried out by the planning and knowledge engineering research community, although not necessarily with a narrative background in mind. GIPO[12] is a knowledge engineering tool that allows the creation of planning domains through a graphical user interface. An authoring tool for emergent narratives will have very similar design requirements as a tool like GIPO. The same group that designed GIPO has also worked on the induction of operator descriptions from examples[7], which is a very similar concept to the Thespian approach, only that in this case the deduced information is used to grow the planning domain.

A suggested authoring environment

The emergent narrative authoring environment that we are going to suggest in the following differs from the tools introduced in the last section in one main aspect: Simulation is directly integrated with the authoring process. This idea is remotely similar to that of a debugger as it can be found in some authoring tools like Scribe[8] or Storytron[2]. However, in those tools just like in traditional programming environments, debugging and development are seen as different stages, whereas in our proposed architecture both processes are inseparable. DraMachina[3] and Thespian[11] prove that it is possible to author interactive stories by specifying linear stories, if an authoring tool extracts information from those stories. The kind of information we want to deduce is planning domain data, so there are some parallels with the work described in [7]. Finally, the idea of mixed-initiative planning as suggested by [13] is also part of our suggested authoring system design.

Story Worlds and Planning Domains

As we pointed out earlier, an authors main task in authoring emergent narrative, is to configure the planners that drive the characters. In contrast to other plot-centered interactive storytelling systems, in emergent narratives there is not only one planner that plans the course of a story, but one planner for each character that plans only the actions of that character. We ultimately want the author to construct a planning domain for those characters[2], without being an expert in planning. The elements of this planning domain (actions and goals) are the main driving force behind the events that will occur in the story. Since the user's actions also contribute to the story line, one single emergent narrative application can tell many different stories, depending on the users choices. We thus do not refer to one such application as a story but as a story world. For the use of the suggested authoring environment, we assume the following situation: The author has the intention of creating a certain story world. They might use an already existing story world as a starting point or start with a completely new one. If the planning domain is empty (i.e. the author started a new story world) the characters will do nothing in a test run of the system and if it is not empty some action might emerge but the story is probably not leading into the anticipated direction. With the help of the authoring tool the author can now incrementally shape the planning domain and as a result the story world toward their vision.

Authoring Method

In our suggested authoring method, the main interaction with the authoring tool will take place in a mixed initiative planning / debug

[2] All characters can share one planning domain, a personlisation of those domains and thus individual behavior can be reached by referring to character properties. For example an action fly can have the precondition that the character needs to have wings and thus will not be available for characters without wings resulting in different beahviour for the same planning domain.

mode. Before entering that mode, the author takes some characters and objects and places them in an environment. They also assign goals to the characters. Starting from this initial situation, a story will develop that is both created by the planner and the author. The main purpose of running through that story however is not the story itself but the development of the participating characters by adding data to the characters planning domain. The author can control the time line, pause or rewind at any time and will usually go through the story step by step. Initially none of the characters might perform any action, because their planning domain is empty or incomplete. In this case the author can control the characters and direct them to perform certain actions. The author is acting out a story like a puppeteer. However, they have to justify every action they are suggesting by specifying the reason for this action. For example a certain action A might be a necessary step before being able to carry out an action B. If the author provides this information the software can create a causal link between the two actions and add it to the planning domain, by adding preconditions and effects to the actions. In a similar way the software can also automatically generate new subgoals or specific instantiations of actions. Once the planning domain is not empty anymore the characters might start making decisions on their own. In this case the author can just step through the story until a point is reached where the author either wants to order a character to do something or a character performs an action on their own that the author does not approve. In this case the author can discard the action but just like specifying a reason for performing an action they will also have to specify the reason for not performing the action. This again will result in a more elaborated planning domain, because those restrictions lead to more detailed pre-conditions or effects. The authoring method is illustrated in figure 1.

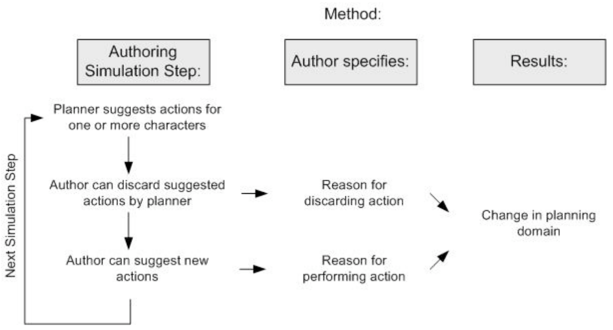

Method:
Authoring Simulation Step:
Author specifies:
Results:
Planner suggests actions for one or more characters
Author can discard suggested actions by planner
Author can suggest new actions
Reason for discarding action
Reason for performing action
Change in planning domain
Next Simulation Step

Figure 1. main authoring method

Example

To clarify the authoring method we will describe a very simplified example that illustrates a common situation within the authoring process. A lot more research has to be conducted in order to specify how exactly the author will communicate their reasons for performing or discarding an action. The challenge here is to allow the user/author to be very specific in communicating their intentions but at the same time to provide a very simple user interface for doing that. For the following example we will assume that the author specifies their motivations via a natural language interface in a pseudo-dialogue with the respective agent.

Imagine a story situation with two agents A and B. Agent A is a pedestrian in the street, Agent B is the bartender in a pub in that street. In that situation both agents are idle and the planner does not generate any action sequences for them to carry out. Even if the author fast forwards in the time line the characters will still stand there and do nothing. In order to change that situation the author has to intervene and take control of one of the characters. We assume, the author orders Agent A to enter the Pub. Now before the simulation can go on from there, the author has to specify their reasons for ordering this action. The following pseudo dialogue between Agent and Author represents the authors specification of their intent:

Agent A: Why do I enter the Pub?
Author: Because you want to buy a drink.
Agent A: Why do I want to buy a drink?
Author: Because you're thirsty.

From this dialogue, the software can deduce at least two facts and add them to the planning domain: The knowledge that you can buy drinks in Pubs (could be expressed as a pre-condition of the buy drink action) and the knowledge that the goal of getting a drink gets activated when the agent is thirsty (precondition of a goal). Now the author can step further through the simulation. After agent A has entered the Pub he will order a drink on his own, without the author having to order that action. If this is the storyline the author anticipates, they can just step forward in the story. In the next step Agent B is selling a drink to Agent A. In this example the author wants to create some conflict and cancels the bartender's action. Again this decision will have to be justified by the author:

Agent B: Why don't I give him a drink?
Author: Because he looks too young.

This time the software can deduces a new pre-condition for the sell drink action. The story could now go on with the bartender asking for an ID, agent A becoming aggressive or whatever the author anticipates. We have to point out that those stories that the author plays through during the authoring process are not necessarily replicable when an end-user is experiencing the story world, because the behavior of the software is determined by the planning domain, which is constantly reshaped by the author. However, ideally the planning domain will incrementally improve and the more stories the author plays through during authoring, the more elaborated the characters will be.

Conclusion

We envision a lot of advantages in using an authoring tool as described in this paper. First of all it forces an author to think about the effects and pre-conditions of actions and thus helps him understand the philosophy of emergent narrative. By allowing the authors to act out linear example stories, the software would facilitate the transition from traditional writing. We also believe that authoring in this environment will be intuitive and also accessible to storytellers without a strong technical background. Because debugging is integrated directly in the authoring process, the author is less likely to produce long time errors. The options of canceling actions and rewinding time make it easy for the author to correct mistakes or wrong conclusions that the software might have drawn.

We will have to refine the suggested authoring method by reviewing work on knowledge engineering and plan authoring. Especially the way the author communicates their intent to the software still requires a lot of attention. Another question that we have not focused on in this paper yet is concerned with the integration of character's individual simulated emotions into the authoring process. Ultimately our long term goal is the implementation of such an authoring tool within our Emergent Narrative Storytelling System.

REFERENCES

[1] R.S. Aylett, J. Dias, and A. Paiva, 'An affectively driven planner for synthetic characters', in *International Conference on Automated Planning and Scheduling (ICAPS)*, pp. 2–10. AAAI press, (2006).

[2] C. Crawford, *Chris Crawford on interactive storytelling*, New Riders, 2005.

[3] S. Donikian and J. N. Portugal, 'Writing interactive fiction scenarii with dramachina', in *Technologies for Interactive Digital Storytelling and Entertainment (TIDSE)*, pp. 101–112. Springer, (2004).

[4] http://www.immersiveeducation.com. Immersive education.

[5] S. Louchart and R.S. Aylett, 'Narrative theory and emergent interactive narrative', *Int. J. of Continuing Engineering Education and Life-long Learning*, **14**(6), 506–518, (2004).

[6] M. Mateas and A. Stern, 'Structuring content in the facade interactive drama architecture', in *Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*. AAAI press, (2005).

[7] T. L. McCluskey, N. E. Richardson, and R. M. Simpson, 'An interactive method for inducing operator descriptions', in *Proceedings of the 6th International Conference on AI Planning and Scheduling (AIPS-2002), Toulouse, France*, (2002).

[8] B. Medler and B. Magerko, 'Scribe: A tool for authoring event driven interactive drama', in *Technologies for Interactive Digital Storytelling and Entertainment (TIDSE)*, pp. 139–150. Springer, (2006).

[9] M. Riedl, *Narrative Generation: Balancing Plot and Character*, Ph.D. dissertation, Department of Computer Science, North Carolina State University, 2004.

[10] S. Sauer, K. Osswald, X. Wielemans, and M. Stifter, 'U-create: Creative authoring tools for edutainment applications', in *Technologies for Interactive Digital Storytelling and Entertainment (TIDSE)*, pp. 163–168. Springer, (2006).

[11] M. Si, S. C. Marsella, and D. V. Pynadath, 'Thespian: Using multiagent fitting to craft interactive drama', in *Autonomous Agents and Multi Agent Systems (AAMAS)*, pp. 21–28. IEEE Computer Society, (2005).

[12] R. M. Simpson, T. L. McCluskey, W. Zhao, R.S. Aylett, and C. Doniat, 'An integrated graphical tool to support knowledge engineering in ai planning', in *European Conference on Planning, Toledo, Spain*, (2001).

[13] J. Thomas and R. M. Young, 'Author in the loop: Using mixed-initiative planning to improve interactive narrative', *ICAPS 2006 Workshop on AI Planning for Computer Games and Synthetic Characters*, (2006).

[14] N. Zagalo, S. Goebel, A. Torres, R. Malkewitz, and V. Branco, 'Inscape: Emotion expression and experience in an authoring environment', in *Technologies for Interactive Digital Storytelling and Entertainment (TIDSE)*, pp. 219–230. Springer, (2006).

Spatial Reasoning and Communication

Spatial cognition has a significant role in our everyday lives. When commuting from our home to our work place, we need a spatial map that enables us to find a reasonable route through the city's road network. When looking for a folder or a textbook in our office, it helps if we know the spatial location at which the item is to be found. When constructing a building, it is essential to understand the spatial-functional relations between the parts of the building: ceilings have to be supported by walls, windows should be inside walls, etc.

Humans interacting with spatial environments typically do so without major conscious efforts; also, communication about spatial relations mainly proceeds smoothly. In spite of the fact that spatial language is highly ambiguous and context-sensitive in many respects, humans generally manage to agree on a suitable interpretation. Space has become such an integral part of our lives that it is used even outside a concrete spatial framework, in metaphorical ways, as in phrases like "on top of the world". It can therefore be successfully argued that any ambient intelligence must have the capability of some form of spatial cognition, which needs to be successfully integrated with, and communicated to, the humans interacting with the environment.

This symposium brings together recent research developments in spatial cognition in relation to ambient intelligence, addressing in particular the relationship between humans and intelligent technology interacting and communicating in spatial environments. Contribution address a number of aspects of spatial cognition concerning communication and computation, including:

- Formal analyses of spatial calculi and models
- Integration of spatial calculi with other reasoning formalisms (e.g., temporal calculi)
- Spatial database queries
- Context-sensitive interpretation and formalization of spatial language, and its mediation towards system-relevant aspects, for example via spatial ontologies
- Spatial human-machine communication via language and/or other modalities
- Computational treatment of functional-spatial relationships in natural environments
- Handling of different spatial granularities
- Dealing with uncertainty in spatial cognition

Hans W. Guesgen, Reinhard Moratz & Thora Tenbrink (Symposium Chairs)

Programme committee: Hans W. Guesgen, Reinhard Moratz, Thora Tenbrink John A. Bateman, Brandon Bennett, Thomas Bittner, Laura Carlson, M. Teresa Escrig, Kathleen Stewart Hornsby, Lars Kulik, Stefan Wölfl

Symposium: Spatial Reasoning and Communication

Spatial cognition has a significant role in our everyday lives. When commuting from our home to our work place, we need a spatial map that enables us to find a reasonable route through the city's road network. When looking for a folder or a textbook in our office, it helps if we know the spatial location at which the item is to be found. When constructing a building, it is essential to understand the spatial-functional relations between the parts of the building: ceilings have to be supported by walls, windows should be inside walls, etc.

Humans interacting with spatial environments typically do so without major conscious efforts; also, communication about spatial relations mainly proceeds smoothly. In spite of the fact that spatial language is highly ambiguous and context-sensitive in many respects, humans generally manage to agree on a suitable interpretation. Space has become such an integral part of our lives that it is used even outside a concrete spatial framework, in metaphorical ways, as in phrases like "on top of the world". It can therefore be successfully argued that any ambient intelligence must have the capability of some form of spatial cognition, which needs to be successfully integrated with, and communicated to, the humans interacting with the environment.

This symposium brings together recent research developments in spatial cognition in relation to ambient intelligence, addressing in particular the relationship between humans and intelligent technology interacting and communicating in spatial environments.
Several symposium contributions deal with robotics applications. Stolzenburg discusses qualitative methods based on angle information and qualitative navigation operators for robot control. Melchert et al. show how a mobile robot can use spatial information of objects to improve communication with humans and other devices located in an intelligent environment. Moratz investigates ambiguous landmark recognition methods using qualitative spatial reasoning.

Other contributions focus specifically on linguistic spatial expressions. Schwering uses Gärdenfors' conceptual spaces as framework to implement Shariff et al.'s model and to determine the similarity of spatial relations. Tenbrink et al. investigate the use of projective reference systems in (simulated) human-robot interaction.

Finally, Hudelot et al. propose an ontology of spatial relations enriched by fuzzy representations of concepts. Dylla et al. show how to model spatial aspects of legal rules by qualitative methods.

Altogether, the contributions of this symposium present a diversified account of current issues and research endeavours in the field spatial reasoning and communication.

Spatial Relations for Perceptual Anchoring

Jonas Melchert and **Silvia Coradeschi** and **Amy Loutfi**[1]

Abstract. In this work we show how a mobile robot can use spatial information of objects to improve communication with humans and other devices located in an intelligent environment. In particular, this work focuses on using spatial relations to facilitate the creation of a connection between symbolic and perceptual representation that refer to the same physical object (anchoring). We extend an anchoring framework to include a set of binary spatial relations which can then be used to exchange information about objects with a human user. To illustrate the performance of the framework, a number of scenarios are presented using a mobile robot. These scenarios are a first step towards the goal of having mobile robots integrated in an intelligent environment and communicating with human users.

1 INTRODUCTION

An emerging trend in the field of robotics is the notion of *symbiotic robotic systems* which consists of a robot, human and (smart) environment cooperating together in performing different tasks [4]. By assisting the robot with information provided by the human or smart objects, some of the current challenges in robotics can be circumvented. For instance, localisation of the robot can be done with a system of surveillance cameras and object recognition tasks can be assisted by passive technologies like RFID. Human assistance and cooperation can also be used to provide instructions to the robot and to assist the robot in case of failure or ambiguous situations when several choices are possible. The motivation behind the symbiotic system is the integration of robotics into everyday life. Therefore, it is essential to allow a range of different users to be able to communicate to the system, this range should include both expert users and even bystanders.

A natural form of communication between humans and the robots is natural language dialogue. In a system where a human provides assistance to the robot it is most convenient for the human to communicate to the robot using dialogue, particularly in the case of a non-expert interacting with the robot. Among the many challenges that this task presents, in this paper, we concentrate on the correspondence that must necessarily exist between the linguistic symbols used by a human and the sensor data perceived by the robot. We call *anchoring* the process of creating and maintaining over time the connection between the symbols and the corresponding perceptual representation that refer to the same physical objects. Already in the field of robotics, anchoring has been explored in systems that use planning and a variety of sensing modalities (e.g. vision and olfaction) [2, 10]. In this paper we examine the possibility to integrate the anchoring framework in a symbiotic robotic system. In particular, we focus on the inclusion of spatial relations in the anchoring framework for the purpose of human-robot communication via language.

To accomplish this task, we extend our existing framework [3] to include a set of binary spatial relations; “at”, “near”, “left”, “right”, “in front”, and “behind” for 2D space. As spatial prepositions are inherently rather vague, a technique using fuzzy sets is applied to define graded spatial relations. The proposed method computes a spatial relations-network for anchored symbols and stores that in the anchors. The relations are then used to assist the robot in resolving ambiguities, identifying objects and improving general task performance of the anchoring framework.

This paper is organised as follows: Section 2 summarises related work on spatial relations and perceptual anchoring. In Sections 3 and 4 we detail the perceptual anchoring and the designed spatial relations used in this work. Section 5 describes some initial experimental scenarios and future work. Section 6 gives a conclusion.

2 RELATED WORK

Most of the work on spatial relations is concerned with connecting the visual domain with the verbal domain of humans. Gapp [6] describes a computational model to compute and evaluate graded spatial relations in 3D space for a visual scene description generator. Objects are approximated by their centre of gravity and bounding rectangle, since only the object’s location is required for the applicability of the spatial relation. The semantics of the relations are defined by evaluation functions depending on the proximal distance and orientation angle between a reference object and the object to be located. Abella and Kender [1] present a system that qualitatively describes the spatial layout of objects with binary relations, from a birds eye view. To account for the vagueness of spatial prepositions, they apply a fuzzyfication technique and use a threshold to decide if two objects are no longer describable by a given preposition. Our computational model for the evaluation of spatial relations is mainly based on the one presented by Gapp, and we apply a thresholding function to select relevant relations.

Work that deals with the abstraction of spatial information from sensory data on robotic platforms is e.g. the one by Skubic et al. [14]. They use a more complex computational model based on the “histogram of forces”. Their system generates linguistic expressions that describe spatial relations between a mobile robot and its environment, based on range readings from a ring of SONAR sensors. Luke at al. [11] present a stereo vision system that can generate linguistic spatial relations for 3D scenes, adopting a fuzzy-set approach and the above mentioned histogram of forces. Hois et al. [9] describe an object recognition system based on 3D LASER scans. The recognition process is supported by interaction with the user and ontological deduction. Unidentified objects can be labelled by the user, using a speech interface, or are classified through the designed domain ontology. In a subsequent phase, the user can query the system for scene descriptions, involving spatial relations to specify object locations.

[1] AASS Mobile Robotics Lab, Örebro University, Sweden
e-mail: jonas.melchert@aass.oru.se, web: www.aass.oru.se

We would like to investigate a similar approach, the exploitation of semantic knowledge, in our future work.

In the above examples the human-robot interaction (HRI) is limited to a (conventional) "master-slave" mode of communication, but our interest is to enable the robot to make use of humans in order to compensate for perceptual or cognitive deficiencies. A good example in this line of thought is the "Peer-to-Peer Human-Robot Interaction" project [5], that aims to develop a range of HRI techniques so that robots and humans can work together in teams and engage in task-oriented dialogue. One of the key components are computational cognitive models for human space perception and spatial reasoning. Of more practical relevance is the work by Moratz and Tenbrink, e.g. [12], that deals with the use of spatial language in human-robot communication. They describe a computational model for a mobile robot platform with a visual object recognition system. The model is evaluated in a number of experiments with uninformed users instructing the robot in spatial identification tasks. Their results provide hints for possible communication scenarios and the employed communication strategies and spatial reference systems, that we will consider in our dialogue system.

So far, the use of spatial relations for anchoring has not been studied in detail. Earlier work [2], investigating the use of planning techniques to recover from perceptual failures and ambiguous cases in perceptual anchoring, incorporated a simple means to refer to an object by specifying its relations to other anchored objects, but only the relations "at" and "near" where supported, and computed on-the-fly using a simple and crisp computational model.

3 PERCEPTUAL ANCHORING

As described in the introduction, the task of anchoring is to create and maintain in time the correspondence between symbols and percepts that refer to the same physical object. This correspondence is reified in a data structure $\alpha(t)$, called an *anchor*. It is indexed by time as the perceptual system continuously generates new percepts; and the created links are dynamic, since the same symbol may be connected to new percepts every time a new observation of the corresponding object is acquired. So at each time instance t, $\alpha(t)$ contains a symbol identifying that object, a percept generated by the latest observation of the object, and a perceptual signature meant to provide the (best) estimate of the values of the observable properties of the object. See figure 1 for a graphical illustration. Following [3] the main parts of anchoring are:

- A *symbol system*, including a set $\mathcal{X} = \{x_1, x_2, \ldots\}$ of individual symbols (variables and constants), a set $\mathcal{P} = \{p_1, p_2, \ldots\}$ of predicate symbols, and an inference mechanism whose details are not relevant here.
- A *perceptual system*, including a set $\Pi = \{\pi_1, \pi_2, \ldots\}$ of possible percepts, a set $\Phi = \{\phi_1, \phi_2, \ldots\}$ of attributes, and perceptual routines whose details are not relevant here. A percept is a structured collection of measurements assumed to originate from the same physical object; an attribute ϕ_i is a measurable property of percepts with values in the domain $D(\phi_i)$. Let $D(\Phi) = \bigcup_{\phi \in \Phi} D(\phi)$.
- A *predicate grounding relation*, $g \subseteq \mathcal{P} \times \Phi \times D(\Phi)$, which embodies the correspondence between (unary) predicates and values of measurable attributes. The relation g maps a certain predicate to compatible attribute values.

The following definitions allow to characterise objects in terms of their (symbolic and perceptual) properties:

- A *symbolic description* σ is a set of unary predicates from $\mathcal{P}$.
- A *perceptual signature* $\gamma : \Phi \mapsto D(\Phi)$ is a partial mapping from attributes to attribute values.

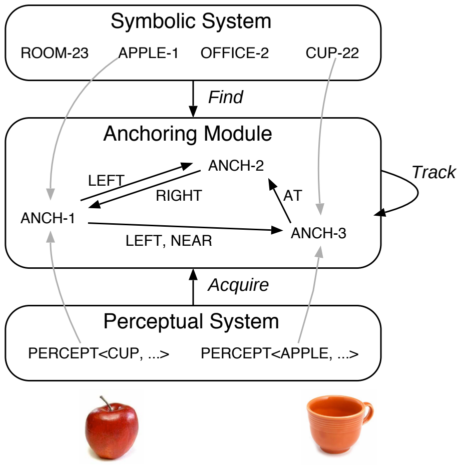


Figure 1. Graphical illustration of the anchoring framework: the anchoring module connects the perceptual and the symbolic systems in a physically embedded intelligent system. Spatial relations between anchored objects are maintained within the anchoring module.

The extension of the framework [3] presented in [10] allows the creation of anchors in both a top-down and a bottom-up fashion: bottom-up acquisition is triggered by recognition events from the sensory system when percepts can not be associated with existing anchors; top-down acquisition occurs when a symbol needs to be anchored to a perceptual description (such a request may come from a top-level planner). These functionalities are realised through:

- *Acquire*: creates a new anchor whenever a percept is received which currently does not match any existing anchor, and inserts symbolic information about the object and its properties into the planner's world model.
- *Find*: takes a symbol x and a symbolic description and returns an anchor α defined at time t (and possibly undefined elsewhere). If an existing anchor, created by *Acquire*, satisfies the symbolic description it selects one; otherwise it searches for matching percepts and, if one is found, creates an anchor for it. Matching of anchor or percept can be either partial or complete: it is partial if all the observed properties in the percept or anchor match the symbolic description, but there are some properties in the description that have not been observed.

At each update cycle of the perceptual system, when new perceptual information is received, it is important to determine if the new information should be associated to an existing anchor (data association problem). The following functionality addresses the problem of tracking objects over time:

- *Track*: takes an anchor α defined at $t-k$ and extends its definition to t. The track assures that the anchor's percept is the most recent and adequate perceptual representation of the object. This facilitates the maintenance of a stable representation of the world on a symbolic level.

By having an anchor structure maintained over time, the challenge is to determine if the association of new percepts is justified or whether certain anchors should be removed. According to [10], this is a difficult problem, because conceptually it is not clear when it is appropriate to remove anchors from the system. The current system adopts a simple solution in which objects that are not perceived when expected decrease in a "life" value. When the anchor has no remaining life, it is removed.

4 SPATIAL RELATIONS

For the computation and evaluation of basic spatial relations' meanings we follow the approach presented in [6] and apply it to 2D space. Two classes of binary spatial relations between a reference object $REFO$ and the object to be located LO (located object) are considered: the topological relations "at" and "near", and the projective relations "front of", "behind", "right", and "left". To model the vagueness of spatial prepositions, the evaluation of a spatial relation results in a degree of applicability in the interval $[0..1]$, representing the range between "not" and "fully" applicable, respectively.

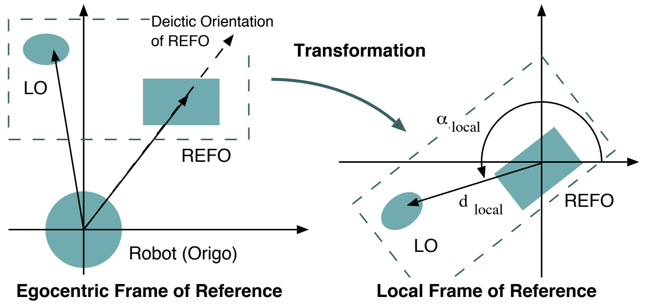


Figure 2. Frame of reference, and computation of distance and orientation angle. Objects are represented by their idealised point location.

4.1 Idealised Object Representation and Frame of Reference

In order to establish spatial relationships between anchored objects we need a geometrical representation of the objects. For the purpose of this work, we assume that the perceptual system provides the relative 2D position of objects with respect to the robot, stored in the perceptual signature. Objects are represented by an idealised point location, derived by projecting the object's centre of gravity (in the video image) onto the floor-plane. See Figure 2 for illustration.

An important aspect is the selection of an appropriate frame of reference [8] for the evaluation of spatial relations. No global frame of reference is used for the robot and therefor also not for the spatial relations. Instead we choose an egocentric frame of reference, as we consider this a more intuitive approach, especially with respect to an intended human-robot interaction (see future work, and [12]).

4.2 Topological Relations

The topological relations "at" and "near" both refer to a region proximal to an object. Following [6] their semantics is defined as: "at" localises an object in the proximal exterior of a $REFO$, and contact is not necessary; for the relation "near" contact between objects is explicitly prohibited.

A local coordinate system at the $REFO$, aligned to its deictic orientation, as shown in Figure 2, is defined, and the local coordinates of the LO w.r.t. the $REFO$ are computed through a transformation $\mathcal{T}_{REFO}$ (rotation and translation). From this the Euclidean distance $d_{\text{local}}(LO) := ||\mathcal{T}_{REFO}(LO)||$ is computed. We use simple trapezoidal membership functions μ_{topo} for the evaluation (others are possible, e.g. spline functions [6]), mapping object distances to the degree of applicability a_{topo}:

$$a_{topo} : (LO, REFO) \mapsto \mu_{topo}(d_{\text{local}}(LO))$$

with $topo \in \{at, near\}$. Figure 3 (top) shows a possible definition for membership functions for the relations "at" and "near".

4.3 Projective Relations

The relations "front of", "behind", "right", and "left" mainly depend on the orientation of the LO w.r.t. the $REFO$, and partition the space in qualitative acceptance areas (as suggested in [8]). But also the distance has to be taken into account: if the distance from the $REFO$ to the LO increases, the degree of applicability a_{proj} decreases. The evaluation function is defined as:

$$a_{proj} : (LO, REFO) \mapsto \mu_{\text{dist}}(d_{\text{local}}(LO)) \cdot \mu_{proj}(\alpha_{\text{local}}(LO))$$

with $proj \in \{front, behind, left, right\}$, mapping the orientation onto the linguistic variables, weighed by the distance. Figure 3 shows a possible definition of the functions μ_{proj} (bottom) and μ_{dist} (top). Although Gapp [7] dropped the distance factor μ_{dist} in an empirically validated revision of the model from [6], we retain it to account for the uncertainty in the visual object localisation, which in our approach weighs heavier than the concern for a cognitively valid model.

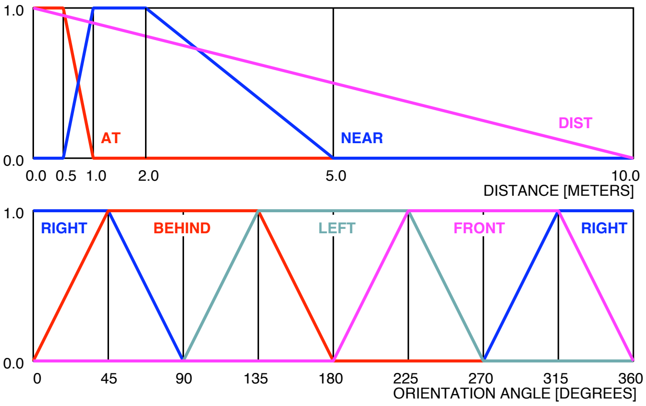


Figure 3. Used membership functions for the evaluation of the spatial relations: μ_{topo} (top), μ_{proj} (bottom), and μ_{dist} (top).

5 ANCHORING WITH SPATIAL RELATIONS

In order to integrate the spatial relations into the existing anchoring framework (see Figure 1), we proceed as follows: At every perceptual update cycle a decision is made for which anchors spatial relations have to be computed. In the current implementation, this is done for all anchored objects. For each anchor, as the located object, all defined spatial relations are computed with respect to all other selected anchors (as reference objects). Only those relations with a degree of applicability greater than a predefined threshold are considered, as in [1], and the others are discarded.

The computed spatial relations, tuples of the form $\langle LO, RO, relation, degree \rangle$, are stored within the anchor of the located object in an additional slot, as we do not consider this information to be part of the anchor's symbolic description. The *find* functionality was extended to include this information for the matching, to allow spatial relations in the symbolic description of a query.

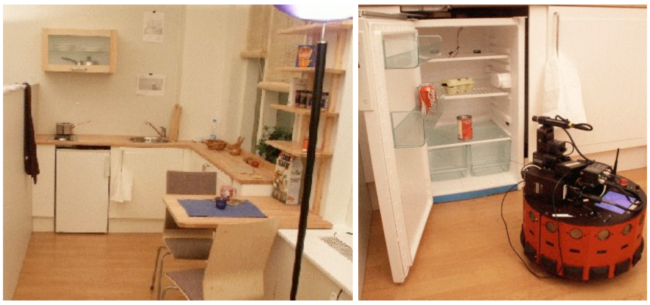


Figure 4. The experimental test-bed, the PEIS-room: view of the kitchen, and robot inspecting the fridge with video camera and electronic nose.

5.1 Example Scenarios

The experimental test-bed for our system is a mobile robot platform that is part of an ambient intelligent environment, called the PEIS Ecology [13]. The robot "shares" a small furnished apartment (the PEIS-room, see Figure 4) with humans and other ambient intelligent devices, and is able to exchange information with these devices.

In the **first example** the robot surveys a static scene with three objects (two green garbage cans and a red ball, see Figure 5) and the anchoring module creates anchors for these objects as soon as they are recognised by the vision system. Then the computation of the spatial relations for these anchors is triggered, resulting in a relation-graph. The list of anchors (in LISP):

```
(ANCHOR ANCH-1 GAR-4
  (SYMBOLIC-DESCRIPTION
    ((SHAPE = GARBAGE) (COLOR = GREEN)))
  (PERCEPTUAL-DESCRIPTION ... )
  (SPATIAL-RELATIONS
    ((GAR-5 ((AT 1.0) (LEFT 0.94)))
     (BALL-2 ((AT 1.0) (BEHIND 0.94)
              (LEFT 0.62)))))
   ... )
(ANCHOR ANCH-2 BALL-2
  (SYMBOLIC-DESCRIPTION
    ((SHAPE = BALL) (COLOR = RED)))
  (PERCEPTUAL-DESCRIPTION ... )
  (SPATIAL-RELATIONS
    ((GAR-5 ((AT 1.0) (FRONT 0.96)
             (LEFT 0.43)))
     (GAR-4 ((AT 1.0) (FRONT 0.96)
             (RIGHT 0.2)))))
   ... )
(ANCHOR ANCH-3 GAR-5
  (SYMBOLIC-DESCRIPTION
    ((SHAPE = GARBAGE) (COLOR = GREEN)))
  (PERCEPTUAL-DESCRIPTION ... )
  (SPATIAL-RELATIONS
    ((GAR-4 ((AT 1.0) (RIGHT 0.94)))
     (BALL-2 ((AT 1.0) (BEHIND 0.96)
              (RIGHT 0.85)))))
   ... )
```

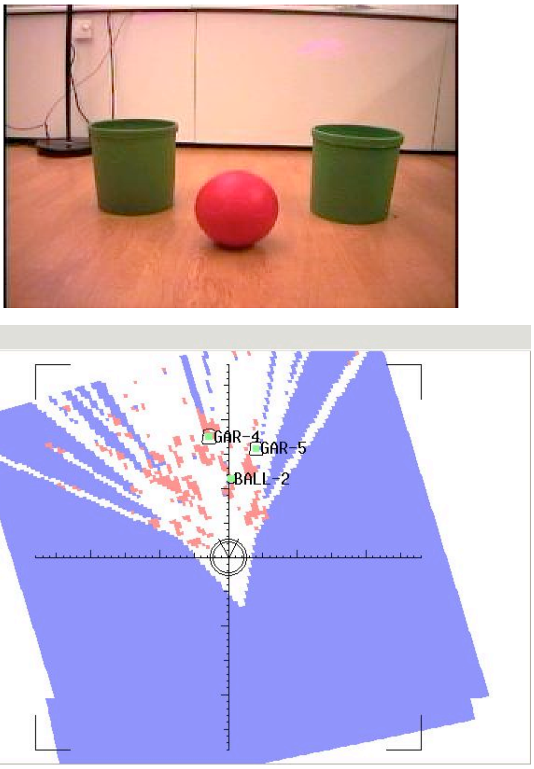


Figure 5. Example scenario: scene from the robot's viewpoint (top) and snapshot of the robot's perceptual space with the created anchors (bottom).

It is now possible to use spatial relations in the *find* functionality (implemented by `(FIND-ANCHOR (NAME SYMBOLIC-DESCRIPTION))`) to search for anchors, for example:

```
(FIND-ANCHOR 'MY-GARBAGE
  '((SHAPE = GARBAGE) (LEFT-TO = BALL-2)))
```

returns `((ANCHOR ANCH-1 MY-GARBAGE ... ))` as result.

In a **second experiment**, a human user is asked to resolve an ambiguity in a *find* request: in the scene from the previous example, the query is "Find the green garbage can". (This experiment is scripted and uses a simple pre-formulated scheme to guide the interaction with the user by text prompts.) As the find request returns more than one anchor (namely ANCH-1 and ANCH-3), the script determines an anchored object that is spatially related to these anchors as reference object and presents the user with a choice, enumerating the returned anchors and their spatial relation(s) to the reference object. Then the query is reformulated using additionally the selected relation(s). For example:

```
? (FIND-ANCHOR 'ANCH
    '((SHAPE = GARBAGE) (COLOR = GREEN)))
- FOUND 2 CANDIDATES: PLEASE CHOOSE
-  1.  GREEN GARBAGE LEFT BEHIND OF RED BALL
-  2.  GREEN GARBAGE RIGHT BEHIND OF RED BALL
? 1
- REFORMULATING:
-  (FIND-ANCHOR 'ANCH '((SHAPE = GARBAGE)
     (COLOR = GREEN) (LEFT-OF = BALL-2)
```

```
      (BEHIND-OF = BALL-2)))
- FOUND: ((ANCHOR ANCH-1 ANCH ...))
```

As outlined in the introduction, the robot should also be able to interact with intelligent devices in the environment, in addition to humans as illustrated in the previous example. Therefore in a possible **third scenario**, the robot could use an external video surveillance system to find an object of interest. In this case, a stationary video surveillance system consisting of several cameras, where each single camera incorporates a private instance of the anchoring module, keeps track of objects in the environment. If the robot is not able to recognise and locate a certain object of interest, but can describe the object in terms of a symbolic description, a request with this description can be sent to the surveillance system. Provided that one of the cameras is able to identify the object, the system, knowing the location of the robot, can qualitatively describe the object's location from the robot's point of view and send a reply.

5.2 Future Work

The current system still lacks a lot of desired functionality and has a number of major shortcomings, e.g.: For now we have not considered ego-motion of the robot; as we use an egocentric frame of reference, spatial relations have to be continuously updated while the robot is moving. The difficulty is to decide when to update the relations (e.g., change of view point), and which anchors are concerned. This demands a more convenient and detailed storage of the relations including view points and reference frames. Furthermore some inference (or reasoning) capability is desirable, to accomplish for example view point-taking (as outlined in [5]).

The linguistic HRI part is still unimplemented and will be one of the next steps. To exploit the anchoring framework in human-robot communication we intend to connect the anchoring module to a symbolic knowledge representation system and a (simple) speech dialogue system, as in [9]. Possible scenarios are semi-autonomous teleoperation of the robot by verbal instructions, like (incremental) navigation instructions, or object identification or localisation tasks involving spatial relations, similar to those described in [12].

6 CONCLUSION

In this work we have extended the anchoring framework to include how objects are spatially related to another in the environment. This is particularly useful for robotic systems working in real environments using real sensor information, as cases of ambiguity may arise where visually identical objects may be present. Furthermore, spatial-relation information facilitates human-robot communication where a human user may find it more intuitive to instruct a robot using spatial communication via language.

The implementation of the anchoring module and the spatial relations-part is still in an early stage and lacking many desired features, so that not all intended scenarios could be tested. This is left for future work.

ACKNOWLEDGEMENTS

This work has been supported through grants from Vetenskapsrådet (the Swedish Research Council) and KK-stiftelsen (the Knowledge Foundation, Sweden). For the first author: This research work was funded in part by CUGS (the National Graduate School in Computer Science, Sweden).

REFERENCES

[1] A. Abella and J. R. Kender, 'Qualitatively describing objects using spatial prepositions', in *Proc. of the 11th Nat. Conf. on Artificial Intelligence (AAAI-93)*, Washington DC, USA, (1993).

[2] M. Broxvall, S. Coradeschi, L. Karlsson, and A. Saffiotti, 'Recovery planning for ambiguous cases in perceptual anchoring', in *Proc. of the 20th Nat. Conf. on Artificial Intelligence (AAAI 2005)*, Pittsburgh, USA, (2005).

[3] S. Coradeschi and A. Saffiotti, 'Anchoring symbols to sensor data: Preliminary report', in *Proc. of the 17th Nat. Conf. on Artificial Intelligence (AAAI-00)*, Austin, USA, (2000).

[4] S. Coradeschi and A. Saffiotti, 'Symbiotic robotic systems: Humans, robots, and smart environments', *IEEE Intelligent Systems*, **21**(3), 82–84, (2006).

[5] T. Fong, I. Nourbakhsh, R. Ambrose, R. Simmons, A. Schultz, and J. Scholtz, 'The peer-to-peer human-robot interaction project', in *Proc. of the AIAA Space 2005 Conf.*, Long Beach, CA, USA, (2005).

[6] K.-P. Gapp, 'Basic meanings of spatial relations: Computation and evaluation in 3D space', in *Proc. of the 12th Nat. Conf. on Artificial Intelligence (AAAI-94)*, Seattle, USA, (1994).

[7] K.-P. Gapp, 'An empirically validated model for computing spatial relations', in *Proc. of the 19th Annual German Conf. on Artificial Intelligence (KI-95)*, Bielefeld, Germany, (1995).

[8] D. Hernández, *Qualitative Representation of Spatial Knowledge*, volume 804 of *LNCS*, Springer-Verlag, 1994.

[9] J. Hois, M. Wünstel, J.A. Bateman, and T. Röfer, 'Dialog-based 3D-image recognition using a domain ontology', in *Proc. of the Int. Conf. Spatial Cognition 2006*, Bremen, Germany, (2006).

[10] A. Loutfi, S. Coradeschi, and A. Saffiotti, 'Maintaining coherent perceptual information using anchoring', in *Proc. of the 19th Int. Joint Conf. on Artificial Intelligence (IJCAI-05)*, Edinburgh, Scotland, (2005).

[11] R. Luke, S. Blisard, J. Keller, and M. Skubic, 'Linguistic spatial relations of three dimensional scenes using SIFT keypoints', in *Proc. of the 14th IEEE Int. Workshop on Robot and Human Interactive Communication (RO-MAN 2005)*, Nashville, TN, USA, (2005).

[12] R. Moratz and T. Tenbrink, 'Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations', *Spatial Cognition and Computation*, **6**, 63–106, (2006).

[13] A. Safiotti and M. Broxvall, 'PEIS Ecologies: Ambient intelligence meets autonomous robotics', in *Proc. of the Joint Conf. on Smart Objects and Ambient Intelligence (sOc-EUSAI '05)*, Grenoble, France, (2005).

[14] M. Skubic, P. Matsakis, G. Chronis, and J. Keller, 'Generating multi-level linguistic spatial descriptions from range sensor readings using the histogram of forces', *Autonomous Robots*, **14**, 51–69, (2003).

Spatial reference in simulated human-robot interaction involving intrinsically oriented objects

Thora Tenbrink and Veronika Maiseyenka and Reinhard Moratz[1]

Abstract. We present the results of a German Wizard-of-Oz (simulated dialogue) study involving referential communication with a (simulated) robot. Users were asked to refer to an object in a complex configuration involving various perspectives on a number of objects together with the robot's position and a chair that could serve as basis for intrinsic reference. Results showed that speakers reliably use the robot's perspective on the scene, that they attend very much to distance when deciding upon a possible relatum, and that the relatum's orientation matters for the decision about a reference system. Furthermore, the data reflect clearly that, given the complexity of the situation, speakers do not reliably account for potential ambiguities. Individual strategies vary considerably. The results are valuable for the development of a robotic system capable of taking speakers' spontaneous spatial reference into account for the identification of otherwise unrecognizable objects.

1 INTRODUCTION

Human-robot interaction in spatial settings has been investigated from a number of perspectives, e.g., [19, 12, 13]. Our aim is to enable robots to identify and categorize objects that are difficult to detect by the robot's visual system, using natural dialogue with human users. Trivially, if only one object is present in a setting, the user may simply name the object for the robot by saying, "This is a bag", for instance. However, in more complex settings the intended object needs to be identified first. In such a situation, human speakers naturally employ either pointing gestures or spatial reference. The former are ruled out in our scenario since gestures are currently still difficult to detect and interpret automatically. Our approach is to identify speakers' spontaneous ways of referring to objects in a complex spatial setting when addressing a robot. For this purpose, one expensive option is to carry out human-robot interaction studies with real systems. In such a scenario, we collected valuable results with respect to configurations that contained up to three objects of the same kind, one further object, a robot, and a human user [17]. Another efficient and well-used method especially in cases in which the system is still under development is to use a so-called Wizard-of-Oz scenario [23]. In such a setting, participants believe that they are talking to an automatic system, while in reality a human is manually triggering the system's responses. Thus, knowledge about how speakers might talk to a system can be gained prior to the actual development of a system that can understand the range of utterances that the speakers will spontaneously produce. In the present study, we wish to compare the efficiency of such an approach with our previous setting, while at the same time considerably widening the scope of configurations and spatial features involved in the scenes. This adds to our knowledge concerning how users spontaneously employ spatial language (in German) in a referential communication task involving an automatic system.

2 SPATIAL REFERENCE

The investigation of spatial language has undergone rapid progress during the past few decades, e.g., [11, 5, 1]. Among the most interesting spatial expressions is a class often labeled *projective terms* [7] to capture the idea that a spatial relationship is *projected* from an origin (position anchoring the view direction) to a relatum (a known object nearby) in order to specify the location of the intended object, here called the *locatum*. This is done using lexical items such as *front, back, left, right*. They allow for the identification of an object in any kind of scenario, without necessitating spatial overlap or inclusion (as with *in*), functional control (as with *on*), or obvious differences in distance (as compared to other objects, as with *close, near, far*). Therefore, and because they are naturally employed in many different kinds of human-human interaction scenarios such as object identification [21], route directions [6], or localization of known objects [2], they are specifically interesting for the scenario envisioned here. They are suitable for the identification of an unknown (or for the robot unrecognizable) object in relation to an already detected object, which can be very useful given the state of current knowledge concerning automatic object recognition.

The employment of projective terms presupposes underlying conceptual reference systems, which were systematically categorized by Levinson [14] as *relative* versus *intrinsic*. Levinson's third major option, *absolute* reference systems, requires different kinds of lexical items, such as *north, south, hill-wards*, etc. In relative reference, a viewer specifies the location of an object relative to a relatum, as in *The chair is in front of the table*. Here, the relatum does not necessarily possess intrinsic sides, and the reference system consists of three different positions (see also [9]). In intrinsic reference systems, the role of the relatum coincides with the role of origin, which therefore needs to possess intrinsic sides, which then serve as basis for reference. In *The table is in front of me*, the speaker serves both as relatum and as origin, and her view direction determines the direction of *front*. In [22], we present evidence that speakers preferably choose a reference system that allows for unambiguous reference, i.e., that produces at least one spatial region in which an unmodified dimensional term can be used unambiguously, if possible at all. This is the case, for example,

- if the goal object is situated at a more extreme position on or near the prototypical axis than any competing objects also situated on the same half-plane; or

[1] University of Bremen, Germany, contact email: tenbrink@informatik.uni-bremen.de

- if the goal object is the only one on a half-plane with respect to the reference system used (regardless of whether or not it is located near the prototypical axis with respect to the relatum).

Generally, spatial descriptions are typically not very precise or detailed, depending - among other aspects - on the discourse task and competing objects within a scenario (e.g., [24]). In [17], we investigated how users employ projective terms in a real human-robot interaction setting. Results showed that speakers are very creative in their instructions, especially in the case of reference failure. Nevertheless, a number of additional systematic patterns emerged. For instance, speakers attended very much to the discourse history in that they consistently re-used syntactic structures that had led to communicative success earlier. Their reference choices when describing the location of the goal object were seldom ambiguous, but rather, they pointed towards the employment of a specific reference system and perspective (which themselves remained underspecified on the linguistic surface), with respect to which the intended goal object could be identified. Furthermore, the robot's exact orientation mattered for the users with respect to the choice of spatial axis, which was interesting in light of the fact that the robot could perceive everything within its front half plane. However, the scenarios used in that previous study were not particularly complex, and they left a number of questions unanswered, mainly for reasons of technological cost: it is simply not possible to carry out a broad range of real interaction studies within a reasonable amount of time. A more efficient method is to use simulated dialogue. This is our approach in the present study. We address the following questions that arose after completion of our previous studies:

1. Do the generalized patterns of usage of projective terms identified previously, for example, in real human-robot interaction studies, hold in simulated dialogue as well?
2. What is the effect of an object within the scene - in addition to robot and locatum - that possesses intrinsic sides?
3. Can we capture the effect of further objects within the scene in a generalizable way?

To answer these questions, we carried out a Wizard-of-Oz study designed to simulate human-robot interaction in spatial settings.

3 METHOD

In order to be able to systematically vary the position of the speaker/observer *in otherwise identical conditions* it is advantageous to work with a two dimensional abstract imaging simulation (such as our example configuration shown in figure 1). In contrast thereto, our earlier efforts [17], used a real robotic system. The application of a real system naturally represents for a test person a direct motivation, a context, which resembles the general purpose of the entire system. The test results thus have an unambiguous validity for real robot systems. However, the relevance for other research groups may be limited, and the results cannot be transferred easily to other languages and systems. In order to build the bridge to real systems for the test persons in a standardized way, this time we present our real system to the participants only in a video. The system is described in [15, 25], and depicted in figure 2 together with a typical configuration of chairs and additional objects on the ground. The instructions of the participants are then based on the abstracted iconified simulation model of an extended version of the actual system (i.e., the simulation can assume more flexible linguistic competence). The comparability of such an approach to applications in real robotic systems can be determined subsequently with fewer participants. Our earlier research [16] indicates, for instance, that typed and spoken instructions directed to a co-present robot resemble each other in the most relevant respects (contrary to earlier expectations), which opens up a nice opportunity to avoid speech recognition problems.

The users were shown a video extract involving the intended robot in action and a possible instruction, so as to direct users towards the intended level of granularity or high-level strategy that the robot will understand. Our previous research [17] showed that speakers are typically completely unsure about how to address an automatic system, and if not informed sufficiently or provided an example utterance, they employ a far broader range of instruction strategies than originally aimed at in the study design. On the other hand, speakers need to remain naive with respect to the robot's exact capabilities, as their utterances could then not be considered natural or spontaneous any more. Our general aim is to develop robots that can interact naturally and efficiently with robots. The collection of suitably natural data which is nevertheless sufficiently controlled to enable informative and generalizable results remains one of the major challenges in human-robot interaction research. The present design is explorative in the sense that the scenarios tested cover a broad range of configurations, and in that the analysis is qualitative rather than quantitative (i.e., it treats the results as a corpus and interprets categories of speakers' choices closely in relation to the situation in which they were produced).

Our simulation system was designed for collecting speakers' references to objects in a spatial situation. It consists of two parts, the user interface and the wizard interface. The participants were seated at the user interface and made to believe that they were communicating directly with the system. They were asked to instruct the robot to show them a specific object in a configuration. The target object was marked (for the user) by a red arrow. The user instructions were forwarded to the wizard interface, where a human operator manually controlled the "dialogue" between the system and the human user. The robot's response was to mark the target object in red. Then the user proceeded to the next configuration and target object.

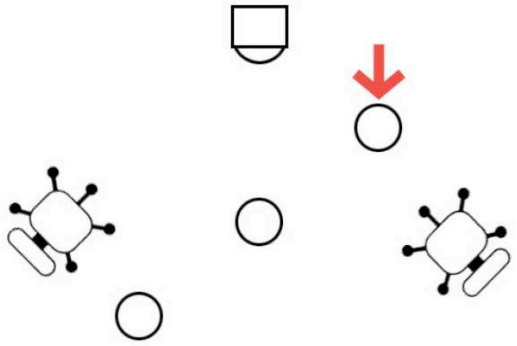


Figure 1. One example configuration

The 24 different configurations include up to five unlabeled objects (circles) together with a chair (and in a few cases a second chair) with clear intrinsic sides, and an indication of the robot's view direction. Most (18) of the configurations contain four of five circles (one

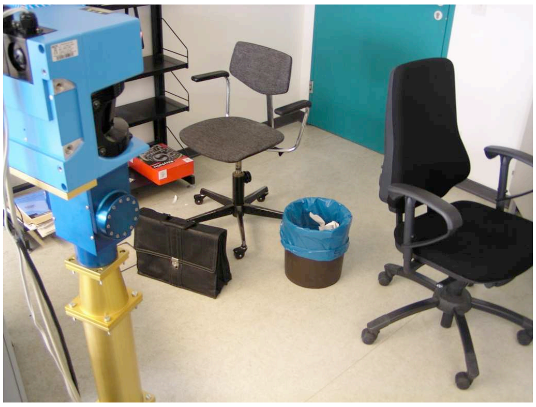


Figure 2. Our robot in action

of which is the target object) in the form of a row that is either identical with or orthogonal to the robot's view direction; in three others, five circles are positioned in the shape of a V, and in the remaining three, three circles are positioned in a diagonal line in front of the robot. The chair is always offset from the circles (at various positions), except in four cases in which it is part of the row. In the three cases containing only three circles in a diagonal line, there are two chairs placed on either side of the line (see figure 1). The user's view upon the scene either does or does not coincide with the robot's view direction on the one hand, and the chair's orientation on the other. In some cases, all three directions collapse, while in others, they differ from each other. The target object may be close to the robot or to the chair, or it is remote from both. This opens up a number of generalizable features of the situation that can be compared with generalized features of the speakers' descriptions. For instance, there was no situation in which the target object was the only one on a half plane with respect to any reference system. This should increase the frequency of linguistic modifications of projective terms, i.e., enhance the linguistic detail given in the instructions. The configurations were always presented to the participants in the same order to avoid confusions with the wizard control.[2]

4 PROCEDURE

9 male and 9 female students from various faculties of the University of Bremen, aged between 20 and 40 years, volunteered for participation in the study. They were not informed about any details concerning the system they were interacting with. They were shown a short introductory video and then asked to type their instructions to the robot in a computer, concerning each of the 24 configurations shown to them. The human operator who triggered the robot's responses was in a different room. The users' instructions as well as the wizard's responses were stored for analysis. Altogether, 411 typed instructions were collected and analyzed qualitatively in relation to systematic features of the spatial configurations.

5 RESULTS

5.1 Linguistic analysis

One third (30.2%) of the utterances contained counting or ordering expressions (the first, second, last, etc.), 33.1% of which did not include a specification of the direction, for example, by a projective term. The probability for using counting expressions was increased (51.8%) in those of the row-like configurations in which the target object was not on an extreme position but rather in the second or third position in the row. On the other hand, the probability was decreased (3.9%) in the three situations in which the objects were situated in the form of a V rather than in a row.

Most (77.1%) of the utterances contained at least one projective term. The other utterances typically relied (only) on counting or ordering terms, or they used distance terms (*the one nearest to the robot*, etc.). The distribution of utterances containing projective terms across participants ranged from 47.8% to 91.7%. Distance terms were used most often in the three situations in which the target object was clearly farther away from the robot than all other objects (32.1%), and nonspatial terms (such as counting and ordering) were used most often (34.0%) in three situations in which four objects were in a row-like configuration identical with the robot's line of sight.

In the remainder of this subsection, we only consider utterances containing projective terms. Altogether, 29.0% of the utterances containing projective terms remained unmodified. The variability between users ranged from no unmodified utterances to 70.6%, reflecting a considerable influence of individual preferences with respect to overall linguistic detail. The frequency according to configurations ranged from 9.1% to 50.0%. The projective terms could be specified in more detail in any of the following ways, or combinations of these:

1. Counting. Of the 317 utterances containing a projective term, 23.0% were specified by a counting or ordering expression. 98.6% of these utterances containing counting together with a direction were not specified further.
2. Combinations of projective terms. Only a low proportion (13.9%) contained more than one projective term. This result was, on the one hand, due to individual preferences: the highest proportion of usage of two projective terms for one user was 39.1%, while others did not employ this option at all. On the other hand, the configurations also clearly influenced the probability. If the goal object was situated at one end of a V-like configuration or at the far end of a diagonal line, the frequency of combined projective terms as in "ganz hinten links" (roughly: at the left back end) increased to 32.1%. But in the 11 (of 24) situations in which the goal object was situated directly on the frontal axis with respect to the robot (whose perspective was typically used, see below), the frequency decreased to 2.6%.
3. Modifications of projective terms by morphemes or by additional lexical items. 25.6% of the projective terms were specified as extreme positions on an axis. This happens linguistically, for example, by the use of modifiers such as "weitest-" (farthest) or "ganz" (all the way), as in "das Objekt ganz links" (the object all the way to the left), or by superlatives, which is not possible in German with the lateral axis (see also [20]). Here, speaker preferences ranged from 34.8% as highest, and no usage as lowest frequency. In the 10 situations in which the object was positioned in an extreme position on the far end of an axis with respect to the robot, the probability increased to 41.2%, while in the 9 situations in which the object was in a mid position within a row, this op-

[2] The order guaranteed that configurations differing only in specific minor (but crucial) aspects were not presented to the participants in direct sequence.

tion was not used at all. Another way of modifying a projective term is to specify not the position *on the axis* (as is the case with extreme positions), but the position *with respect to the axis*. This happens by the usage of terms like *direct, diagonally* and occurred in 15.1% of the utterances containing a projective term. Here, invidual usage varied from no occurrences to 40.9%. This modification was used most often (20.5%) in those situations in which the target object was positioned directly on an axis with respect to a relatum (either the robot or the chair), and no other objects were situated in between them. The highest frequency (41.2%) occurred in two comparatively complex (V-like or diagonal) situations in which the target object was placed directly in the line of vision of the robot.

5.2 Choice of relatum

28.7% of the collected instructions explicitly refer to the chair as a relatum, and 25.8% to the robot. Nearly half (42.3%) remain implicit with respect to the relatum. The distribution between configurations is not equal. We analyze the following systematic features of the configurations:

1. Relative distance. In the 15 (of 24) configurations in which the target object was positioned closer to the chair than to the robot, the probability of using the chair as an explicit relatum increased to 35.7%, and the robot was only referred to as relatum in 15.1% of the cases. In the opposite case (target object clearly closer to robot than to the chair), the chair was referred to in only 7.2%, while the robot served as relatum in 55.1% of the instructions.
2. Orientation. In those configurations in which the robot shares the view direction with the user and the chair's intrinsic front, the chair is referred to in 45.3%, but the robot only in 1.9% of cases.
3. Relation to axis. In those cases in which the target object is situated directly on one of the intrinsic axes of the chair, the chair is referred to in 43.0% of the cases. Where this holds for the intrinsic axes of the robot, the robot is referred to in 34.4% of the cases.

Each of these three criteria contributes to the choice of relatum in its own way. Clearly, however, many factors play together in each individual lexical choice, so that only (more or less clear) tendencies can be detected. Crucially, also individual preferences play a role: Both the chair and the robot are referred to as a relatum by some individuals in up to more than 60% of cases, but are completely neglected by others. One participant leaves the relatum completely unspecified in 82.6% of cases, and otherwise mostly refers to the *room* as a somewhat unusual, outward relatum.

5.3 Choice of reference system

In this subsection, we only consider the 319 utterances containing projective terms, since only these presuppose underlying conceptual reference systems. Due to the frequent underdeterminacy of spatial descriptions, the underlying reference system could not in all cases be determined, which does not necessarily lead to ambiguities: on the contrary, reference systems tend to be less determined especially in those cases in which no ambiguities arise, because reference systems coincide (see also [22]). For example, the chair could be used as a relatum in one of two ways: Users could either rely on an intrinsic reference system, in which the chair's intrinsic sides are used for reference, or they used a relative reference system, employing either their own or the robot's perspective independent of the chair's intrinsic sides. The two options cannot be differentiated by the linguistic surface alone [22]. They could only be differentiated in a subset of the scenarios, namely, where the chair's orientation does not coincide with the robot's or the speaker's point of view, as well as with the usage of the frontal axis where the orientations coincide.

Of the reference systems that could be identified in general (147 utterances, 46.4% of the utterances containing projective terms), intrinsic reference systems were the most frequent (68.7% of the 147 cases). In more than half (61.2%) of the intrinsic cases, the robot was used as relatum, otherwise the chair. A clear identification of relative reference systems was possible in 14.4% of the utterances. The underlying perspective (which was never mentioned explicitly but could be inferred in these cases) was almost always either that of the robot alone, or shared by robot and speaker; only in 8.7% of cases was the speaker's perspective used where it was not shared by the robot. In nearly all cases, the relatum used was the chair (with two exceptions). The chair was used as a relatum in an identifiable intrinsic reference system in 12.6% of cases, and in an identifiable relative reference system in 13.9% of cases.

Some participants never used clear cases of either relative or intrinsic reference systems, while others used one of those options in up to more than half of their utterances. This points to a strong influence of individual preferences.[3] Additionally, configurations can be identified in which either kind of reference system becomes more likely than elsewhere. In one configuration, as many as 71.4% of utterances containing projective terms were intrinsic. In that situation, the robot's front direction was directly oriented towards the object, which was not true for either the chair's orientation nor the speaker's view on the scene. Considering all (four of the 24) situations together that share this feature, the proportion of intrinsic frame use is 67.3% (as opposed to the overall proportion of only 32.3% clear intrinsic frame use). Other aspects of the distribution, which cannot be listed here for reasons of space, further support the conclusion that a direct position on the frontal axis of a relatum (robot or chair) strongly enhances the choice of an intrinsic reference frame with respect to that relatum. Similar effects can be traced, though to a lesser degree, with respect to the other spatial axes. - Clear relative reference frames are used most often in one situation in which the target object is located, again, in the line of vision of the robot, but the chair is situated between the robot and the target object, so that the target object is an obstacle which must be mentioned and employed in either a relative (53.3%) or an intrinsic (40.0%) reference system.

The case just described exemplifies another interesting finding in our data. This configuration, as well as one other, allowed for the meaningful usage of expressions with opposite meaning, namely, *in front of* and *behind*, depending on the underlying reference system. Often, users did not seem to be aware of this possible ambiguity. Most users switched between intrinsic and relative reference systems throughout the interaction. In the cases in point, they often did not provide sufficiently detailed information to avoid the ambiguity. A typical instruction in such a case is, *Show me the object behind the chair*, which in one situation corresponded to the intrinsic frame, and in the other, to the relative frame. Four of the 18 users actually used the same expression twice, obviously unaware (since the scenarios did not follow each other directly in the sequence) that they had to be interpreted differently. This result poses a high potential for communication problems.

[3] These preferences may well have been triggered by the examples given to the participants, which in some cases were intrinsic and in others relative. Unfortunately, the exact distribution in this regard cannot be retraced.

6 DISCUSSION

In this study, we have investigated a small corpus of simulated human-robot interaction data with respect to a range of complex spatial configurations. A range of systematic results emerged from the qualitative analysis.

Our findings show that counting and ordering is a rather natural strategy to use for object reference in row-like configurations. Further specifications are then not considered necessary. The fact that the direction of counting needs to be inferred in as many as one-third of the cases may lead to miscommunication. Also, our results confirm earlier findings in that the speakers often do not provide detailed spatial specifications. Nearly one third of the projective terms were unmodified, which is a fairly high frequency in light of the fact that all of the configurations were complex; they contained at least four distractor objects (including one or two chairs), plus the robot and the target object which was of the same object class as most of the distractor objects. Thus, the mere presence of competing objects in a scenario is not sufficient to guarantee detailed spatial descriptions; rather, the interplay of features of the scenario either does or does not render an unmodified spatial term sufficient. Also, individual preferences played a much higher role than could previously be identified. Interestingly, unmodified projective terms were used at least once in each scenario, and each scenario triggered modifications by at least five of the 18 participants (the highest number was 13).

Our previous findings concerning speakers' principles of using unmodified projective terms could be extended by the following generalizations. The principle that projective terms are more likely to remain unmodified if the target object is positioned at an extreme position on a spatial axis does not seem to hold for more than three objects. In most of our present configurations, the rows contained at least four objects; here, the likelihood for the usage of specifiers concerning the position on the axis increased. Thus, simple spatial descriptions such as *the left one* seem only to be considered sufficient with respect to up to three objects. This principle does not apply in the present scenarios. Neither does a second principle (indicated above) apply, since it presupposes that the target object is the only one on a half plane, which was never the case in the present configurations. Therefore, the fact that the overall frequency of modifications of projective terms is astonishingly low does not seem to be due to shared features of the configuration, but rather, to the participants' individual assessment of the situation at hand. This may well reflect a more general cognitive principle according to which increased complexity does not necessarily lead to increased complexity in speakers' linguistic representations, but rather, to increased variability in speakers' individual solutions to the problem of referential communication. This idea conforms, as do the principles identified for simpler scenarios, with Grice's maxim of quantity [8] as well as with the principle of minimal effort proposed by Clark & Wilkes-Gibbs [4]. Our findings for complex scenarios additionally illustrate that these principles are active even in the case of potential misunderstandings, as the resulting underspecified utterances often result in potential ambiguities: for instance, five of the 18 users instructed the robot to show them *the object left of the chair* in a situation in which there were two objects situated on the left side of the chair. Similar underspecifications could also be detected in other configurations. In many cases, there was no intuitive way of disambiguating these utterances.

With respect to the choice of reference systems, the present scenarios were designed specifically to investigate the impact of the inclusion of an object with intrinsic features (a chair), which adds the option of using another intrinsic reference system in addition to referring only to the addressee (the robot). Results show that speakers choose the chair and the robot approximately with equal frequency as a relatum, depending on an interplay of factors such as individual preferences, (shared) view directions, and relation to the spatial axis. Crucially, the distance of the potential relatum to the target object plays a major role. However, the users clearly preferred to use the robot's perspective, in accord with our own earlier findings in real human-robot interaction [17] as well as with more general principles of partner adaptation [10, 3]. In the case of intrinsic reference systems, the role of origin coincides with the role of relatum, and in the case of (identifiable) relative reference systems, the robot's role of origin led to the use of the chair as relatum. Thus, the chair could be used in two ways, both of which occurred with approximately equal frequency. In some cases, this possibility led to potential ambiguities which, similar to the results concerning linguistic detail, were not always accounted for by the users. Our results suggest that the intrinsic frame is preferably used if the target object is positioned directly on the front axis of the relatum. This can, however, only be judged from the frequency with respect to the clearly identifiable instances. These represent only a subset of cases, due to the underdeterminacy in the utterances and the coincidence of reference systems in many situations. Therefore, further research is needed here to further substantiate these conclusions. They correspond to our own previous results [22] according to which a reference system is chosen that allows for the usage of an unmodified projective term, which is less unproblematic with a position on the focal axis than elsewhere [11]. Also, new (statistically validated) results of a recent (as yet unpublished) study by C. Vorwerg and J. Sudheimer strongly support the idea that the position with respect to a focal axis has an influence on reference frame choice. Apart from that, results concerning the choice of reference systems are at present rather controversial and inconclusive, though there seems to be an influence both of language and of functional factors (see [22] for a systematic review).

There are a number of obvious drawbacks in the design of the current study. First of all, the corpus is too small to warrant statistical validation. Therefore, the analysis was entirely qualitative out of necessity. Second, a number of parameters relevant to the dialogue history were not sufficiently controlled for and could therefore not be accounted for, such as the influence of the example given to the participants to ensure the correct level of granularity. Recent work on alignment in dialogue [18] shows clearly that speakers are influenced by previous utterances not only on such a high level of strategy choice, but also on other levels of interaction which were not analyzed here. Related processes probably also influenced the course of the dialogue for the participants. The fact that the order of configurations could not be randomized is therefore a major drawback. For these reasons, our focus in the present analysis is strictly on the spatial choices in relation to the configurations. Other influencing factors that had to be neglected clearly need to be addressed directly in subsequent research. It would also be interesting to investigate results of cases of miscommunication; in our present design, participants were generally successful, independent of the spatial strategy they used. In natural human-robot interaction, as our own previous research shows, this is not regularly the case, and then speakers can be shown to vary their communicative styles considerably [17]. In the present case, the frequent success (plus, perhaps, to a certain degree the fact that nothing particular was at stake for the volunteers in our experiment) may well have encouraged speakers to produce ambiguous utterances. Clearly, this aspect requires further scrutiny. Nevertheless, there is a parallel here with respect to our previous real human-robot interaction study in that speakers frequently re-use their own previous syntactic or conceptual patterns in the case of success: this was the case only for a narrow range of utterances in the real study - leading to a small range of variety in users' utterances after first success - and for a broad range of utterances in the present simulation. Therefore, speakers here did not have a reason to change their referential strategies; they simply adapted slightly to the demands of the different configurations.

As motivated above, the investigation of naturalistic human-robot interaction data is generally a non-trivial endeavour, not least because of the high technical costs involved in the preparation of real systems and their experimental testing with participants unfamiliar with the system. The present Wizard-of-Oz study, with all its limitations, contributes to the overall range of data currently available as a source of information useful for the development of real systems to be employed in spatial settings. Crucially, as the limitations are not due to the scenario being simulated, and our results do match and extend previous findings in a sensible way, we could show that this approach is valuable in triggering meaningful and systematic user responses. This results in findings that can easily be reproduced, validated, and extended in follow-up studies.

7 CONCLUSION

We have used a simulated human-robot interaction scenario to investigate users' natural spatial reference choices in complex configurations. Results show that speakers vary considerably in their individual solutions to the problem at hand, but they often do not account for potential underlying ambiguities sufficiently. Although a number of systematic patterns and preferences could be identified with respect to conceptual and linguistic user choices, they do not allow for reliable predictions. Robotic systems dealing with this kind of spatial reference therefore need to be equipped, on the one hand, with suitable computational models representing the conceptual options for all kinds of reference systems available in a situation which may lead to different interpretations of the same utterance, and on the other hand, with sophisticated dialogue systems enabling the robot to ask the clarification questions that are necessary to disambiguate underspecified user input.

ACKNOWLEDGEMENTS

We would like to thank Bernd Krieg-Brückner and Michael Wünstel for their support, Benjamin Schröder and Marc-Björn Seidel for designing and implementing the simulation system, and Markus Mörke and Andrej Galkov for their assistance in carrying out the experiments. Also, funding by the DFG (SFB-TR8 Spatial Cognition) is appreciated.

REFERENCES

[1] *Functional features in language and space: Insights from perception, categorization and development*, eds., L. A. Carlson and E. van der Zee, Oxford University Press, Oxford, 2005.

[2] Laura A. Carlson-Radvansky and Gordon D. Logan, 'The Influence of Reference Frame Selection on Spatial Template Construction', *Journal of memory and language*, **37**, 411–437, (1997).

[3] Herbert H. Clark, *Using Language*, Cambridge University Press, Cambridge, 1996.

[4] Herbert H. Clark and Deanna Wilkes-Gibbs, 'Referring as a collaborative process', *Cognition*, **22**, 139, (1986).

[5] Kenny R. Coventry and Simon C. Garrod, *Saying, seeing and acting: The psychological semantics of spatial prepositions*, Psychology Press: Essays in Cognitive Psychology series, 2004.

[6] Michel Denis, Francesca Pazzaglia, Cesare Cornoldi, and Laura Bertolo, 'Spatial discourse and navigation: an analysis of route directions in the city of venice', *Applied Cognitive Psychology*, **13**(2), 145–174, (1999).

[7] Carola Eschenbach, 'Contextual, functional, and geometric components in the semantics of projective terms', in *Functional features in language and space: Insights from perception, categorization and development*, eds., Laura Carlson and Emile van der Zee, 71–91, Oxford University Press, Oxford, (2005).

[8] Herbert P. Grice, 'Logic and conversation', in *Syntax and semantics*, eds., Peter Cole and Jerry Morgan, volume 3, 41–58, Academic Press, New York, San Francisco, London, (1975).

[9] Theo Herrmann, 'Vor, hinter, rechts und links: das 6H-Modell. Psychologische Studien zum sprachlichen Lokalisieren.', *Zeitschrift für Literaturwissenschaft und Linguistik*, **78**, 117–140, (1990).

[10] Theo Herrmann and Werner Deutsch, *Psychologie der Objektbenennung*, Hans Huber Verlag, Bern u.a., 1976.

[11] Annette Herskovits, *Language and Spatial Cognition: an interdisciplinary study of the prepositions in English*, Studies in Natural Language Processing, Cambridge University Press, London, 1986.

[12] A. Knoll, B. Hildebrandt, and J. Zhang, 'Instructing cooperating assembly robots through situated dialogues in natural language', in *Proc. IEEE Conference on Robotics and Automation*, Albuquerque, New Mexico, (1997).

[13] T. Kyriacou, G. Bugmann, and S. Lauria, 'Vision-based urban navigation procedures for verbally instructed robots', *Robotics and Autonomous Systems*, **51**(1), (2005).

[14] Stephen C. Levinson, *Space in language and cognition: explorations in cognitive diversity*, Cambridge University Press, Cambridge, 2003.

[15] Reinhard Moratz, 'Intuitive linguistic joint object reference in human-robot interaction', in *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI)*, (2006).

[16] Reinhard Moratz and Thora Tenbrink, 'Instruction modes for joint spatial reference between naive users and a mobile robot', in *Proc. ICRISSP 2003 IEEE International Conference on Robotics, Intelligent Systems and Signal Processing, Special Session on New Methods in Human Robot Interaction, October 8-13, 2003, Changsha, Hunan, China*, (2003).

[17] Reinhard Moratz and Thora Tenbrink, 'Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations', *Spatial Cognition and Computation*, **6**(1), 63–106, (2006).

[18] Martin J. Pickering and Simon Garrod, 'Towards a mechanistic psychology of dialogue', *Behavioural and Brain Sciences*, **27**(2), 169–190, (2004).

[19] E. Stopp, K. Gapp, G. Herzog, T. Laengle, and T. Lueth, 'Utilizing spatial relations for natural language access to an autonomous mobile robot', in *KI-94: Proceedings of the Eighteenth German Conference on Artificial Intelligence*, Berlin, Heidelberg, (1994). Springer.

[20] Thora Tenbrink, 'Identifying objects in English and German: A contrastive linguistic analysis of spatial reference', in *Proc. WoSLaD Workshop on Spatial Language and Dialogue, October 23-25, 2005*, (2005).

[21] Thora Tenbrink, 'Identifying objects on the basis of spatial contrast: an empirical study', in *Spatial Cognition IV: Reasoning, Action, Interaction. International Conference Spatial Cognition 2004, Proceedings*, eds., Christian Freksa, Markus Knauff, Bernd Krieg-Brückner, Bernhard Nebel, and Thomas Barkowsky, pp. 124–146, Berlin, Heidelberg, (2005). Springer.

[22] Thora Tenbrink, *Space, time, and the use of language: An investigation of relationships*, Mouton de Gruyter, Berlin, in prep.

[23] Zygmunt Vetulani and Jacek Marciniak, 'Corpus Based Methodology in the Study and Design of Systems with Emulated Linguistic Competence', in *NLP 2000, LNCS 1835*, ed., D.N. Christodoulakis, 346–357, Springer, Berlin, Heidelberg, (2000).

[24] Constanze Vorwerg and Thora Tenbrink, 'Discourse factors influencing spatial descriptions in English and German', in *Proc. International Conference Spatial Cognition 2006, September 24-28, Bremen, Germany*, (2006).

[25] Michael Wünstel and Reinhard Moratz, 'Automatic object recognition within an office environment', in *CRV 2004 Canadian Conference on Computer and Robot Vision*, IEEE, (2004).

SailAway: Formalizing Navigation Rules

F. Dylla and **L. Frommberger** and **J.O. Wallgrün** and **D. Wolter**[1] and **B. Nebel** and **S. Wölfl**[2]

Abstract. Agents that have to solve navigational tasks need to consider aspects that go far beyond single-agent goal-directed deliberation: What an agent does in a specific situation often interferes with what other agents do at the same time. In order to avoid conflicts or even collisions, situations in space are governed by laws, rules, and agreements between the involved agents. For this reason, artificial agents interacting with humans must be able to process such rule sets, which are usually formulated in natural language. In this paper we present a case study on how to formalize navigation rules in the domain of sea navigation. We present an approach that uses qualitative representations of navigation rules. Qualitative spatial reasoning methods can be applied to distinguish permissible actions in the set of all possible actions. We argue that an agent's spatial representation can be modeled on a qualitative level in a natural way and that this also empowers sophisticated high-level agent control.

1 Introduction

A considerable part of everyday human activities is guided by regulations, for example, regulations on how to behave in traffic scenarios, recommendations on how to use escalators, rules on how to enter subways and buses, or rules of politeness at bottlenecks. Most of these rules have in common that they are usually formulated in natural language and hence extensively use *qualitative terms* to describe spatial situations and actions. For example, in traffic laws qualitative concepts are used to describe relevant situations and also the "correct" behavior of agents in these situations. Another feature is that most of the rules depend on the agent's *role* in a particular situation. What an agent is allowed to do, may depend on whether he is a pedestrian or on the kind of vehicle she is using.

Representations of rule-compliant behavior, of course, are not limited to navigation. Examples of rule sets guiding the behavior of agents can also be found in sports, in games, in expert recommendation systems, and so on. Rule sets need to be made explicit and be formalized at different stages when artificial agents or multi-agent systems are specified or implemented. First, rules can be used to specify the desired behavior of an artificial agent (for instance a mobile robot or an autonomous vehicle) such that an implemented system can be tested against these specifications. Rules may also be used to actively control an artificial agent, for example, when we wish to restrict possible trajectories of a mobile system. Formal encodings of rules are also crucial for implementing control systems that observe and judge the behavior of other agents. Finally, rule sets need to be formalized in order to evaluate them according to given criteria, to find gaps, inconsistencies, or deadlocks. For instance, if a rule set describes how *two* agents have to behave in specific situations, one could investigate how this rule set would perform in more complex situations involving more than two agents: Is the rule set still sound in the sense that its intentions (e.g., collision avoidance) are met if all agents act in compliance with the rules? And, is the rule set complete in the sense that it covers all possible situations?

In this paper, we investigate how rules in sea navigation can be formalized and discuss the benefits of qualitative spatial representation formalisms. Qualitative representations link metrical information perceivable by the agents to more abstract characterization of situations in which rules can or have to be applied. On the basis of these qualitative representations, we show how spatial reasoning techniques can be used to assign rule-compliant actions to each agent in each concrete situation.

2 Approaches to Formalizing Navigation Rules

Most traffic regulations are written down in natural language texts. For making such rules available to a computer implementation, they need to be formalized or encoded in a suitable language. On the basis of this formalization, concrete situations of objects can be classified and permissible actions can be selected. An appropriate formalization is key to an accurate modeling of the rules and essential for empowering effective reasoning. The formalization serves as a double link: It links continuous real-world scenes to discrete classes described by the representations (scene classification) and it links symbolic rule descriptions to possible actions available to agents in a concrete scenario (navigational reasoning).

Navigation rules in sea navigation generally subsume classes of *configurations* (i.e., spatial constellations of agents) in which they assign permissible or obligatory behavior of agents. For example, in a configuration in which two motor boats are in head-on position, one navigational rule prescribes that both boats need to turn starboard[3]. Besides spatial configurations, rules can also depend on other aspects such as types of vehicles used by the involved agents. Sport vessels, for example, have to give way to commercial shipping vessels. However, since knowledge of this kind can be formalized rather easily, the crucial point for formalizing navigation rules is to formally represent spatial configurations in a suitable way (in terms of the considereded rules) and to formally represent the actions prescribed by these rules.

2.1 Logical Framework

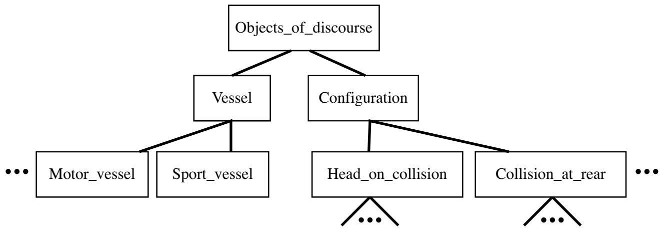
A formalization of navigation rules relates agent types (i.e., classes of vessels) and their spatial constellations as handled by the rules.

[1] Universität Bremen, Germany, email:{dylla, lutz, wallgruen, dwolter}@sfbtr8.uni-bremen.de

[2] University of Freiburg, Germany, email:{nebel, woelfl}@informatik.uni-freiburg.de

[3] *Starboard* is the nautical term that refers to the right side of a vessel with respect to its *bow* (front); *port* refers to the left hand side, *stern* to the back.

This can often be compiled into a small ontology. Using a logical approach representing this information appears most adequate to provide a suitable basis for reasoning. Description logics offer a solid approach to modeling ontological information and provide also the means for formalizing spatial configurations. Agent types and configurations are represented as *concepts*, whereas spatial relations are used as *roles* to interrelate the relative positions of agents. The utilization of qualitative spatial calculi provides us with a suitable set of spatial relations that allows for linking spatial reasoning techniques to the logical framework. Details are discussed in the following sections, at this point, we just assume that a suitable set of spatial relations to model configurations described by rules (e.g. head-on course) exists. We employ one additional role `involves` that relates configurations to agents. For example, if we consider a configuration defined by two agents in head-on course, the role-fillers of `involves` are the specific agents in head-on course. This approach allows us to consider scene classification as ABox-reasoning in description logics: A specific configuration is realized when role fillers for `involves` can be instantiated such that the formula describing the situation is valid. In Fig. 1 we present an overview of the simple ontology employed in this application (a) and give an exemplary logical representation of the exemplary spatial configuration of agents in head-on course (b). It presents the special case of a dangerous configuration of a motor and a sport vessel in head-on collision course.



(a)

$$\begin{aligned}\texttt{Motor\_vessel} &\sqsubseteq \texttt{Vessel}\\ \texttt{Sport\_vessel} &\sqsubseteq \texttt{Vessel}\\ \texttt{Head\_on\_coll\_Type1} &\doteq \exists\,\texttt{involves}.(\texttt{Motor\_vessel} \sqcap\\ &\qquad \exists\,\texttt{headon}.\texttt{Sport\_vessel})\end{aligned}$$

(b)

Figure 1. Overview of the ontology (a) and exemplary configuration (b)

The advantage of embedding rule formalization in a standard logical framework lies in the possibility of exploiting standard logical reasoning techniques. In principle, it is possible to reason about rule systems themselves (meta-level reasoning) as well as reasoning about rule-compliant actions (navigational reasoning). In any case, fundamental prerequisites are that (a) a finite set of (binary) spatial relations can describe configurations in a sufficiently precise way and that (b) the mapping from natural language to formal representations can be performed in an easy-to-use manner.

In summary, typical rule sets can be formalized using the logical framework of description logics to represent the ontology. The logical framework must incorporate a set of spatial relations that is adequate for representing the rules and for navigational reasoning. Thus, we argue for combining ontological knowledge engineering with appropriate qualitative spatial representation techniques.

2.2 Qualitative Spatial Calculi for Formalizing Configurations of Agents

Qualitative spatial calculi are well-suited to bridge between quantitative scene information observable by an agent and linguistic descriptions of object configurations [8]. Technically speaking, qualitative spatial calculi abstract from metrical data by summarizing similar quantitative states into one qualitative characterization. Qualitative calculi reveal the relative nature of spatial information: properties of objects are compared to one another rather than comparing the properties to some external (measuring) scale.

A binary qualitative calculus defines a set of jointly exhaustive and pairwise disjoint (JEPD) binary relations between objects of some domain D. Usually we are interested in calculi that are closed under converse and composition: The converse operation may be considered a shift of perspective, i.e., it allows us to deduce how object P is related to object Q when we know how Q is related to P. The composition operation yields the set of relations that can hold between objects P and Q if the relations between P and some third object R and the relation between Q and R are known. In other words, composition integrates local knowledge to survey knowledge.

Based on these operations, constraint-based reasoning techniques have been developed in the literature (see, e.g., [1]). In our application, we will apply these methods for infering actions that agents are allowed to perform in a given spatial situation (see Section 4).

In the context of sea navigation, position information, i.e., information about direction and distance, is essential. In particular, orientation information is required to differentiate spatial constellations as described by navigation rules. Currently, distance information only plays a subordinate role in our approach: We use such information only to distinguish those boats that are close enough to other boats such that they need to be considered when navigation rules are evaluated. Several calculi for dealing with positional information have been presented in the literature (e.g. [4, 7]). In our context, the $\mathcal{OPRA}_4$ calculus [7] is of particular interest, because this calculus is well-suited for dealing with objects that have an intrinsic front or move in a particular direction.

$\mathcal{OPRA}_4$ is designed for reasoning about relative orientation relations between *oriented points* (points in the plane with an additional direction parameter)[4]. For each pair of oriented points, 4 lines are used to partition the plane into 8 planar and 8 linear regions (see Fig. 2). The orientation of the two points is depicted by the arrows starting at $\vec{A}$ and $\vec{B}$, respectively. The regions are numbered from 0 to 15, where region 0 always coincides with the orientation of the point. An $\mathcal{OPRA}_4$ base relation is a pair (i, j), where i is the number of the region, seen from $\vec{A}$, that contains $\vec{B}$ and j vice versa. These relations are written as $\vec{A}\,{}_4\angle_i^j\,\vec{B}$. Additional base relations describe situations in which both oriented points are at the same position. However, these are not of particular interest in this work, because superpositions of oriented point represent collision situations. It should not go unmentioned that $\mathcal{OPRA}_4$ is not the only calculus rule sets might be modeled with. But since we focus here on translating rules from natural language descriptions to a qualitative formalization for agent control, $\mathcal{OPRA}_4$ is expressive enough for the translation and shows a good run-time behavior in the reasoning process. For example, an alternative calculus may be the qualitative trajectory calculus [9].

[4] $\mathcal{OPRA}_4$ is actually a particular instance of the granulated $\mathcal{OPRA}_m$ calculus in which the granularity parameter m determines the number of base relations. An oriented point $\vec{O}$ can be described by its Cartesian coordinates $x_O, y_O \in \mathbb{R}$ and a direction $\phi_{\vec{O}} \in [0, 2\pi)$ with respect to an absolute frame of reference.

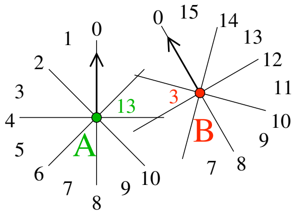


Figure 2. The $\mathcal{OPRA}_4$ relation $\vec{A}\,{}_4\angle^3_{13}\,\vec{B}$.

2.3 Modeling Spatio-Temporal Transitions by Conceptual Neighborhoods

Navigation rules restrict the possibilities of agents to act in space. For representing actions in a formal model, we must combine spatial and temporal information. An elegant way for accomplishing this spatio-temporal linkage is provided by so-called *conceptual neighborhoods* [3]. The idea of conceptual neighborhoods is to specify the discrete relation transitions that are possible due to continuous transformation [5]. Two base relations are considered *conceptually neighbored* if there can be a change-over due to an arbitrary small transformation of the objects. We denote the set of relations conceptually neighbored to a relation r by $\mathtt{cn}(r)$. In this context, such transformations are movements or changes of orientation of one or more of the involved objects. Depending on the transformations considered, different conceptual neighborhood structures can be induced [2].

A conceptual neighborhood graph of all base relations can be constructed interpreting the binary relation of "conceptually neighbored" as adjacency in the graph [3]. The neighborhood graph represents continuity aspects on the geometric or physical level of description in a discrete manner: Continuous processes map onto identical or neighboring classes of descriptions. A movement of an agent with respect to another agent can then be traced on the qualitative level as a sequence of neighboring spatial relations which hold for adjacent time intervals. Put differently, actions can be represented on a qualitative level as trajectories in the neighborhood graph. This provides us with an elegant approach to represent actions.

The basic idea underlying our approach is to consider rule-specific *transition systems* that differ from conventional neighborhood graphs in two aspects: First, we label edges in the graph by actions of the involved agents that cause the transition (thus, we obtain a directed graph), and second, we consider only edges that represent rule-compliant (or nearly rule-compliant) behavior of the agents. For example, a neighborhood transition $r_i \xrightarrow{(a_1,a_2)} r_j$ takes place when r_i represents "head-on", a_1 "turn to portside", a_2 "keep course", and r_j "on starboard side".

The starting point for defining these transiton systems is to identify an idealized transition sequence (the *idealized thread*), which may be considered a prototypical rule-compliant plan of maneuvers from a start to an end configuration if we observed the vessels in each point in time.

The idealized thread is not yet a suitable formalization of rule-compliant actions, as it abstracts from alternative action effects that need to be considered: depending on the precise position of the vessels, the same action may lead to different change-overs with respect to the qualitative relations as defined by the neighborhood graph. Therefore, the idealized thread is extended to a transition system that also includes neighbored configurations if they are still within the scope of the traffic rule at hand. For each of these added configurations, we derive actions that lead the vessels closer to the idealized thread. Analogously, we apply this method of neighborhood-based relaxation to start and end configurations.

3 Collision Regulations in Sea Navigation

In our investigations we focus on the domain of sea navigation. Traffic regulations for sea navigation have been defined in the *International Regulations for Preventing Collisions at Sea* (ColRegs) of the International Maritime Organization (IMO). For each pair of vessels, the rules define which one has to give way (burdened vessel) and which is the privileged one (it is possible that both vessels are burdened). Reasonable avoidance behavior of burdened vessels is described by specific patterns in supplemental textbooks.

In the following, we will focus on vessels in sight of one another. For each pair of boat types, the conditions "from port", "from starboard", "head-on", and "from rear" must be considered such that, for n different boat types, $4n^2$ cases can be distinguished. However, it is sufficient to first derive transition systems for the general avoidance patterns and then refine these for the concrete boat types and velocities.

In our scenario, every vessel has a goal point where it is directly heading to. If vessels are in danger of collision, they are able to perform one of the three actions: turning starboard (S), turning port (P), or keeping the course (midships, M). These steering actions have a temporary effect: The helm is put for a short period of time and afterwards the helm is put back to midships. In general, the motion of standard vessels can be compared to Ackermann kinematics, i.e., in general turning is not possible without any translational velocity, and sidewards motion is not possible at all. We assume all vessels moving with a constant translational velocity $v_t > 0$. Furthermore, we assume a prototypical velocity for each vessel type. Currently, speed changes are not considered.

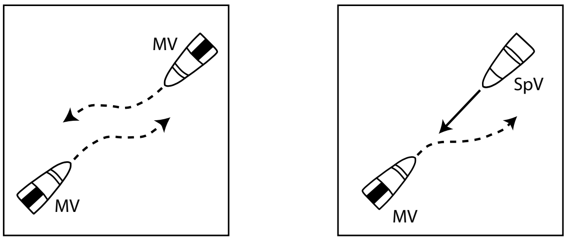
3.1 Kinematic Neighborhood Structure

The kinematics of the vessels induce a neighborhood structure in the underlying qualitative spatial representation (i.e., $\mathcal{OPRA}_4$) that is exploited for constructing the transition system (cf. Section 2.3). Since neighborhood transitions must correspond to physically possible behavior, the general neighborhood structure of $\mathcal{OPRA}_4$ takes three different aspects into account: superposition, simultaneous motion, and agent kinematics [2]. The general neighborhood structure for solid objects with unconstrained motion (where objects cannot superpose) $\mathtt{cn}_s$ (subscript s stands for "solid") is defined by

$$\begin{aligned}\mathtt{cn}_s({}_4\angle^j_i) = \{ & {}_4\angle^{j-1}_{i-1}, {}_4\angle^{j}_{i-1}, {}_4\angle^{j+1}_{i-1}, {}_4\angle^{j-1}_{i}, \\ & {}_4\angle^{j+1}_{i}, {}_4\angle^{j-1}_{i+1}, {}_4\angle^{j}_{i+1}, {}_4\angle^{j+1}_{i+1} \}.\end{aligned}$$

But since we assume different prototypical velocities for each vessel type, the neighbohood structure needs to be refined in order to match the kinematics of vessels. This means that we have to define a restricted neighborhood function $\mathtt{cn_s}'(r) \subseteq \mathtt{cn}_s(r)$ for each pair of vessel types. Put in other words, we need to capture the relation transitions corresponding to possible actions, i.e., if vessels S_1 and S_2 are in relation r then for every relation $r' \in \mathtt{cn_s}'(r)$ there exists at least one action pair that causes a transition into relation r'.

Due to lack of space we can just outline the general idea how this refined neighborhood structure can be determined. Consider a configuration with two vessels S_1 and S_2 of the same type with $S_1\,{}_4\angle^j_i\,S_2$.



(a) Two motor vessels (MVs): both have to alter their course starboard to pass each other on port side

(b) Motor vessel and Sport vessel (SpV): MV has to turn starboard, SpV holds course

Figure 3. Exemplary rule for two different kinds of vessels.

Because of type equality we presume equal translational velocities. According to [2], a turn to port by S_1 results in ${}_4\angle_{i-1}^{j}$, and a turn to starboard in ${}_4\angle_{i+1}^{j}$. A turn by S_2 results in the according changes of j. If both vessels perform a turn, e.g., S_1 to port and S_2 to starboard, the resulting relation can also be ${}_4\angle_{i-1}^{j+1}$.

We now need to determine neighboring relations for any of the $3^2 = 9$ potential action pairs. For the turning actions we assume an arbitrary rotation velocity $v_r > 0$. We denote the actions by (a_1, a_2) where a_1 is the action performed by S_1 and a_2 by S_2. If, for example, in a situation in which $S_1 \, {}_4\angle_2^2 \, S_2$ holds, one of the action pairs (S, S), (S, M), (M, S), or (M, M) is performed, the resulting configuration is unique, namely $S_1 \, {}_4\angle_3^3 \, S_2$. The actions (S, P) and (M, P) result in one of the three configurations: $S_1 \, {}_4\angle_3^1 \, S_2$, $S_1 \, {}_4\angle_3^2 \, S_2$, or $S_1 \, {}_4\angle_3^3 \, S_2$, depending on the relative differences in translational and rotational velocity which we do not take into further account here. The actions (P, S) and (P, M) result in the converse: $S_1 \, {}_4\angle_1^3 \, S_2$, $S_1 \, {}_4\angle_2^3 \, S_2$, or $S_1 \, {}_4\angle_3^3 \, S_2$. Only for the action tuple (P, P) the neighboring relations cannot be restricted compared to $\mathtt{cn}_s({}_4\angle_2^2)$.

Another interesting case is $S_1 \, {}_4\angle_0^0 \, S_2$. Due to $v_t > 0$ it is not possible to end up in $S_1 \, {}_4\angle_0^1 \, S_2$ if (M, S) is performed, or $S_1 \, {}_4\angle_0^{15} \, S_2$ for (M, P). The resulting configuration is definitely $S_1 \, {}_4\angle_1^1 \, S_2$ or $S_1 \, {}_4\angle_{15}^{15} \, S_2$, respectively. The above results need to be determined manually, which is of course a laborious task.

3.2 An Exemplary Rule in Sea Navigation

As mentioned before, different types of vessels require to apply different rules. For example, rule 14(a) of the ColRegs says: "When two power-driven vessels are meeting head-on or nearly head-on courses so as to involve risk of collision each shall alter her course to starboard so that each shall pass on the port side of the other." However, if a motor vessel meets a sport vessel, only the motor vessel has to turn starboard, and the sport vessel is the privileged one (these two rules are illustrated in Fig. 3).

We will now give a formalization of the collision avoidance pattern depicted in Fig. 3(a), which is in compliance to the ColRegs and built on the (refined) neigborhood structure presented above.

In the first stage, we generate the idealized thread for the rule, that is, the idealized course describing the transitions from a dangerous into a safe configuration. For this, reconsider the example in Fig. 3(a): First the vessels are head-on, then both must turn starboard. When they are not head-on anymore, they can go midships, and when they are just about side by side, they can turn port, heading to their original course. This idealized thread is depicted in Fig. 4. A box denotes a start configuration and a double circle a safe configuration denoting that the rule is processed and the boats are in no danger of a collision anymore.

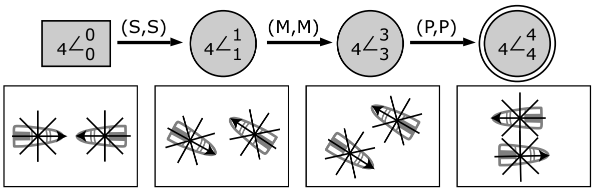


Figure 4. Idealized thread for the rule shown in Fig. 3 (a).

The idealized thread is not yet a suitable formalization of rule-compliant actions, as it abstracts from alternative action effects that need to be considered: Depending on the precise position of the vessels, the same action may lead to different change-overs with respect to the qualitative relations as defined by the neighborhood graph. That means, that observed real-world transitions are not necessarily neighborhood transitions (in particular, we can hardly observe transitions form region to line relations as considered in $\mathcal{OPRA}_4$). Therefore, relations that are neighbored to one in the idealized thread are included in the rule transition system as well. Fig. 5 shows the resulting rule transition system generated from the idealized thread. The idealized thread is highlighted by shaded boxes and circles. We note that relation ${}_4\angle_2^2$ has been included in this thread, because ${}_4\angle_1^1$ and ${}_4\angle_3^3$ are no direct neighbors and ${}_4\angle_2^2$ directly links these two relations. These transitions are derived under the premise of prototypical velocity. As we are considering same-type vessels in this rule model we presume the same velocity. But as soon as these actions are executed with different velocities the effects of executing them does not necessarily lead to perceiving the predicted relation. If, for example, ${}_4\angle_3^3$ holds, both vessels should turn port. Ideally, this results in ${}_4\angle_4^4$. But a slight difference in velocity may yield ${}_4\angle_4^3$, ${}_4\angle_3^4$, or one of the alternative end configuration. Assuming a velocity being *just about the same* for both vessels, we expect that the resulting relation r is at least a conceptual neighbor of the idealized relation. For models concerning different types of vessels with different prototypical assumptions on velocity we need to generalize: If the velocity proportion between two vessels is just about the same as assumed in the prototypical model, $r \in \mathtt{cn}_s(r_p)$ holds.

In this example, our start configuration ${}_4\angle_0^0$ marks linear regions. However, such situations are unlikely to occur and are "unstable", which means that any steering action or difference in velocity may lead to a neighboring relation (in this case, ${}_4\angle_{15}^{15}$ or ${}_4\angle_1^{15}$, e.g.). Therefore we have to consider those situations as start configurations as well (cf. Fig. 5).

Analogously, we add configurations neighbored to the idealized safe configuration. As the side-by-side relation ${}_4\angle_4^4$ is a linear relation (only linear regions occur), it is more likely that a neighboring relation is perceived. For being sure that we cannot go back into a collision situation we end a rule only if one vessel has already completely passed the other vessel, i.e., ${}_4\angle_5^3$, ${}_4\angle_3^5$, and ${}_4\angle_5^5$.

Incorporation of neighboring relations makes our formalization robust against noise in perception and action execution.

4 Reasoning for Agent Control

In the following, we briefly sketch a first concrete application of our formalization of the sea navigation rules. While acting according to the rules will avoid collisions in situations involving two vessels, this is generally not guaranteed when more than two vessels are involved. We therefore investigated how the formalization of the sea navigation

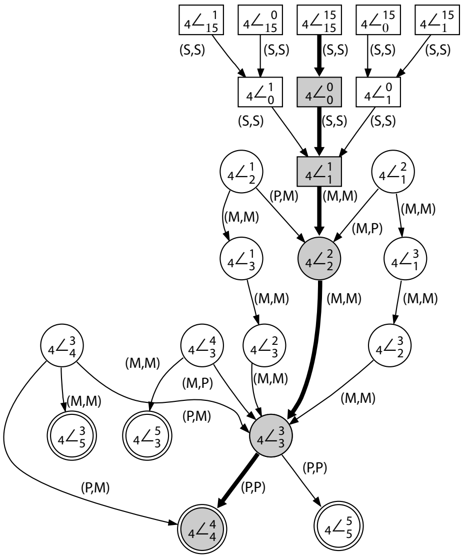


Figure 5. Complete transition system for the rule depicted in Fig. 4

rules can be employed to control and coordinate the vessels in order to avoid collisions in more complex situations.

In our approach, we combined the qualitative spatial relations describing the current situation and the relations describing possible future configurations between two boats as provided by the transition systems of the applicable rules to form a constraint network in which consistency corresponds to exemption from collisions. Constraint-based reasoning techniques [6] are then used to find a consistent and thus collision-free solution. The result is then repropagated to determine the suitable actions for the individual vessels that will lead to this particular constellation.

A simple example of the developed SailAway demonstrator depicted in Fig. 6 illustrates how the combination of the formalization with qualitative spatial reasoning techniques achieves collision-free navigation in a situation involving three boats.

5 Conclusion and Outlook

We investigated formalization of navigation rules, focusing on assessing the utility of qualitative representation and reasoning techniques in a real-world application scenario. Our investigation confirmed previous research in that qualitative representations enable mediation between real-world metric information and conceptual knowledge as used in communication or rule descriptions. It provides effective means to compile rules into a formal representation. Most notably, a qualitative representation is linked to a formal logic framework and realizes a tight integration of all components in a complex agent control application. Our approach is intimately linked to a high-level reasoning component and thereby differs from related approaches based on qualitative information, e.g. [10].

Currently, we only make use of comparatively simple reasoning techniques as we only aim at determining *some* action that is compliant with the rules. A more sophisticated approach would involve a planning component. So it appears promising to extend qualitative representation and reasoning to become an integral part of frameworks for reasoning about action and change and to be integrated into high-level agent control languages. We aim at advancing our approach in that direction.

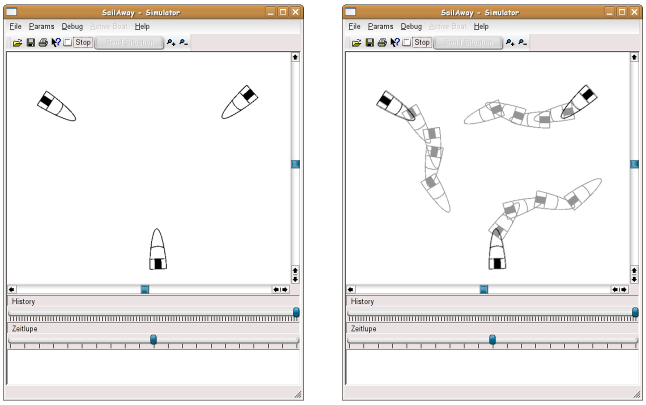


Figure 6. Configurations in the SailAway simulator window: left a situation with three motor vessels, right the resulting trajectories

6 Acknowledgments

The authors would like to thank Christian Freksa for fruitful discussions and impulses. Our work was carried out in the DFG Transregional Collaborative Research Center SFB/TR 8 Spatial Cognition, projects R3-[Q-Shape] and R4-[LogoSpace]. Financial support by the Deutsche Forschungsgemeinschaft is gratefully acknowledged.

REFERENCES

[1] Anthony G. Cohn and Shyamanta M. Hazarika, 'Qualitative spatial representation and reasoning: An overview', *Fundamenta Informaticae*, **46**(1-2), 1–29, (2001).

[2] Frank Dylla and Jan Oliver Wallgrün, 'Qualitative spatial reasoning with conceptual neighborhoods for agent control', *Journal of Intelligent and Robotic Systems*, (2007).

[3] Christian Freksa, 'Conceptual neighborhood and its role in temporal and spatial reasoning', in *Proceedings of the IMACS Workshop on Decision Support Systems and Qualitative Reasoning*, eds., Madan G. Singh and Luise Travé-Massuyès, pp. 181–187, North-Holland, Amsterdam, (1991). Elsevier.

[4] Christian Freksa, 'Using orientation information for qualitative spatial reasoning', in *Theories and methods of spatio-temporal reasoning in geographic space*, eds., A. U. Frank, I. Campari, and U. Formentini, 162–178, Springer, Berlin, (1992).

[5] Antony Galton, *Qualitative Spatial Change*, Oxford University Press, 2000.

[6] Peter Ladkin and Alexander Reinefeld, 'Effective solution of qualitative constraint problems', *Artificial Intelligence*, **57**, 105–124, (1992).

[7] Reinhard Moratz, 'Representing relative direction as binary relation of oriented points', in *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI 2006)*, Riva del Garda, Italy, (August 2006).

[8] Reinhard Moratz and Thora Tenbrink, 'Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations', *Spatial Cognition and Computation*, **6**(1), 63–106, (2006).

[9] Nico van de Weghe, Bart Kuijpers, Peter Bogaert, and Philippe Maeyer, 'A qualitative trajectory calculus and the composition of its relations', in *First International Conference on GeoSpatial Semantics (GeoS 2005)*, volume 3799, pp. 181–211. Springer, (2005).

[10] Michael F. Worboys, 'Event-oriented approaches to geographic phenomena', *International Journal of Geographical Information Science*, **19**, 1–28, (2005).

A Granular Point Position Calculus for solving ambiguous landmark problems in Cognitive Robotics

Reinhard Moratz

1 Introduction

A qualitative representation provides mechanisms which characterize central essential properties of objects or configurations. A quantitative representation establishes a measure in relation to a general standard of measure which is generally usable.

The constant general availability of common measures is now self evident. However, one needs only remember the example of the history of technologies of measurement of length to see that the more local relative measures, which are qualitatively represented, (for example, "one piece of material is longer than another" versus "this thing is two meters long") can be managed by biological/epigenetic cognitive systems much more easily as absolute quantitative representations.

The two main trends in Qualitative Spatial Reasoning are topological reasoning about regions [2, 10] and positional reasoning about point configurations [4, 11]. Especially positional reasoning is important for robot navigation [9].

Typically, in Qualitative Spatial Reasoning relatively coarse distinctions between configurations are made only. Applications exist in which finer qualitative acceptance areas are helpful. The possibility to use finer qualitative distinctions can be viewed as a stepwise transition to quantitative knowledge. The idea of using context dependant direction and distance intervals for the representation of spatial knowledge can be traced back to Clementini, di Felice, and Hernandez [1]. However, only special cases of reasoning were considered in their work. Here, we will propose a calculus that makes direct use of general purpose constraint propagation. Thereby robot applications including reasoning about ambigue perceptions will be made possible.

2 Generalizing ternary point configuration calculi

In 2-dimensional space, two points *A* and *B* can be used to “localise” a third point *C*; this is relative localisation, which means that no absolute reference system, such as in [3], is used: (1) *A* is the origin (which may be, for instance, the speaker’s location); (2) *B* is the relatum; and (3) *C* is the reference object. The localisation of *C* relative to *A* and *B* consists then of describing *C* relative to the reference system determined by *A* and *B*. We shall be considering two kinds of relative localisation:

1. Relative distance: how far is *C* from *A* compared to *B*? In other words, how does the distance from *A* to *C* compare with the distance from *A* to *B*?
2. Relative orientation: what is the angular distance of *C* from *B* for an observer placed at *A*? In other words, what is the angle determined by the directed straight lines (AB) and (AC)?

These two relative localisations will then be combined to lead to relative position.

The newly proposed calculus is called granular point configuration calculus GPCC. In this calculus two points are the basis for a reference system. The reference system can be interpreted as a partition of the plane into acceptance regions for a third point. All options for places of the third point which are in the same part of the partition are considered to be in an aquivalence class and are treated in the same way in categorization and reasoning tasks by subsequent modules. One variant of the GPCC calculus and its partition on the plane is shown in figure 1.

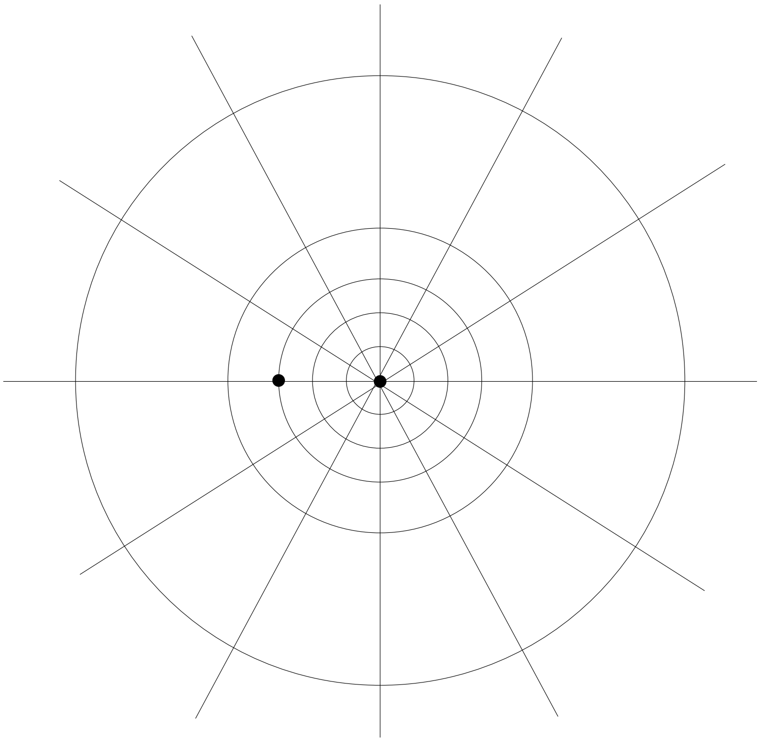


Figure 1. The partition of the GPCC$_3$-Calculus

To give a precise, geometric definition of the GPCC-relations we describe the corresponding geometric configurations in an analogue way to the TPCC calculus [7] on the basis of a Cartesian coordinate system represented by $\mathbb{R}^2$. First we define the special cases for $A = (x_A, y_A)$, $B = (x_B, y_B)$ and $C = (x_C, y_C)$.

$$\begin{array}{rcl} A, B \text{ dou } C & := & x_A = x_B \wedge y_A = y_B \wedge (x_C \neq x_A \vee y_C \neq y_A) \\ A, B \text{ tri } C & := & x_A = x_B = x_C \wedge y_A = y_B = y_C \end{array}$$

For the cases with $A \neq B$ we define a relative radius $r_{A,B,C}$

$$r_{A,B,C} \quad := \quad \frac{\sqrt{(x_C - x_B)^2 + (y_C - y_B)^2}}{\sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}}$$

$$A, B \text{ sam } C \quad := \quad r_{A,B,C} = 0$$

and for $A \neq B \neq C$ a relative angle $\phi_{A,B,C}$:

$$\phi_{A,B,C} \quad := \quad \tan^{-1} \frac{y_C - y_B}{x_C - x_B} - \tan^{-1} \frac{y_B - y_A}{x_B - x_A}$$

The further base relations have an acceptance area depending on the granularity of the calculus to be applied. The calculus shown in figure 1, GPCC$_3$, has a level of granularity of 3 and 267 relations. A calculus of the granularity level m, described below as GPCC$_m$, has $(4m - 1)(8m) + 3$ base relations. The base relations of GPCC$_3$ are thus defined:

$$\begin{array}{rcl} A, B \;{}_3\!\perp^1_0\; C & := & 0 < r_{A,B,C} \leq 1/3 \wedge \phi_{A,B,C} = 0 \\ A, B \;{}_3\!\perp^1_1\; C & := & 0 < r_{A,B,C} \leq 1/3 \wedge 0 \leq \phi_{A,B,C} \leq 1/6\pi \\ A, B \;{}_3\!\perp^1_2\; C & := & 0 < r_{A,B,C} \leq 1/3 \wedge \phi_{A,B,C} = 1/6\pi \\ A, B \;{}_3\!\perp^1_3\; C & := & 0 < r_{A,B,C} \leq 1/3 \wedge 1/6\pi \leq \phi_{A,B,C} \leq 2/6\pi \\ & \vdots & \\ A, B \;{}_3\!\perp^1_{23}\; C & := & 0 < r_{A,B,C} \leq 1/3 \wedge \\ & & 11/6\pi \leq \phi_{A,B,C} \leq 12/6\pi \\ A, B \;{}_3\!\perp^2_0\; C & := & r_{A,B,C} = 1/3 \wedge \phi_{A,B,C} = 0 \\ & \vdots & \\ A, B \;{}_3\!\perp^3_0\; C & := & 1/3 \leq r_{A,B,C} \leq 2/3 \wedge \phi_{A,B,C} = 0 \\ & \vdots & \\ A, B \;{}_3\!\perp^9_0\; C & := & 3/2 \leq r_{A,B,C} \leq 3/1 \wedge \phi_{A,B,C} = 0 \\ & \vdots & \\ A, B \;{}_3\!\perp^{11}_{23}\; C & := & 3/1 \leq r_{A,B,C} \wedge 11/6\pi \leq \phi_{A,B,C} \leq 12/6\pi \end{array}$$

This schema can be transferred and applied to arbitrary granularity m of a calculus GPCC$_m$. The general segments $A, B \;{}_m\!\perp^i_j\; C$ are then so defined:

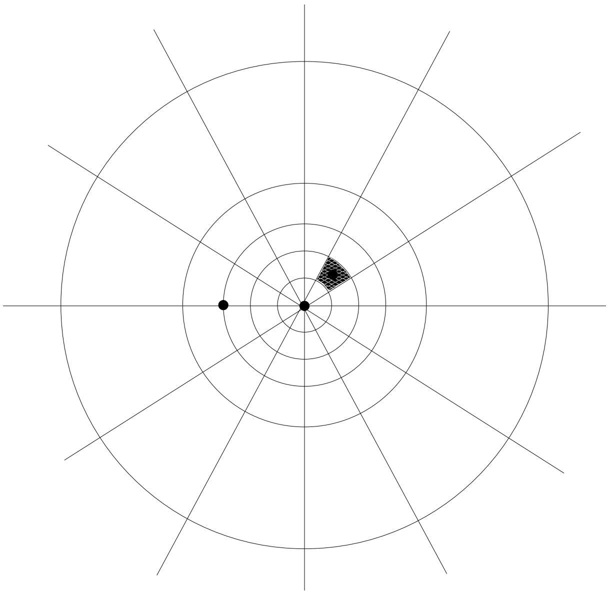


Figure 2. An example configuration of three points A, B, C. The depicted configuration corresponds to $A, B \;{}_3\!\perp^3_3$

$$\begin{array}{rcl} 0 \leq j \leq 8m - 2 \wedge j \bmod 2 = 0 & \rightarrow & \phi_{A,B,C} = \dfrac{j}{4m}\pi \\ 1 \leq j \leq 8m - 1 \wedge j \bmod 2 = 1 & \rightarrow & \dfrac{j-1}{4m}\pi < \phi_{A,B,C} < \\ & & \dfrac{j+1}{4m}\pi \\ 1 \leq i \leq 2m - 1 \wedge i \bmod 2 = 1 & \rightarrow & \dfrac{i-1}{2m} < r_{A,B,C} < \\ & & \dfrac{i+1}{2m} \\ 2 \leq i \leq 2m \wedge i \bmod 2 = 0 & \rightarrow & r_{A,B,C} = \dfrac{i}{2m} \\ 2m + 1 \leq i \leq 4m - 3 \wedge & & \\ i \bmod 2 = 1 & \rightarrow & \dfrac{m}{2m - \frac{i-1}{2}} < r_{A,B,C} < \\ & & \dfrac{m}{2m - \frac{i+1}{2}} \\ 2m + 2 \leq i \leq 4m - 2 \wedge & & \\ i \bmod 2 = 0 & \rightarrow & r_{A,B,C} = \dfrac{m}{2m - \frac{i}{2}} \\ i = 4m - 1 & \rightarrow & m < r_{A,B,C} \end{array}$$

Because we have three arguments, we have 3! = 6 possible ways of arranging the arguments for a transformation. Following Zimmermann and Freksa [13] we use the following terminology and symbols to refer to these permutations of the arguments (a,b : c):

term	symbol	arguments
identical	ID	a,b : c
inversion	INV	b,a : c
short cut	SC	a,c : b
inverse short cut	SCI	c,a : b
homing	HM	b,c : a
inverse homing	HMI	c,b : a

With ternary relations, one can think of different ways of composing them. However there are only a few ways to compose them in a way such that we can use it for enforcing local consistency [12]. In trying to generalize the path-consistency algorithm [6], we would like to enforce 4-consistency [5]. We then had to use the following (strong) composition operation:

$$\forall A, B, D:\ A, B\,(r_1 \diamond r_2)\,D \leftrightarrow \exists C:\ A, B\,(r_1)\,C \wedge B, C\,(r_2)\,D$$

Unfortunately, the $\textsc{Gpcc}_m$ calculi are not closed under strong composition. For that reason we can not directly enforce 4-consistency. But we can define a weak composition operation $r_1 \Diamond r_2$ of two relations r_1 and r_2. It is the most specific relation such that:

$$\forall A, B, D:\ A, B\,(r_1 \Diamond r_2)\,D \leftarrow \exists C:\ A, B\,(r_1)\,C \wedge B, C\,(r_2)\,D$$

While using the weak composition we can not enforce 4-consistency we still get usefull inferences.

The problem is calculating the permutation and composition results for such structures by machine. The operation tables can be approximated with the aid of a composition of distance orientation intervals (DOI) [8]. Thereby areal segments and their borders are summarized. Thus one obtains thereby a quasi-partition in which only linear overlappings occur.

The tables for approximate transformations and for the approximate compositions of the calculi $\textsc{Gpcc}_3$, $\textsc{Gpcc}_4$, and $\textsc{Gpcc}_5$ can be seen in the internet [**?**]. The calculi are, with respect to the transformation $\textsc{Hmi}$, closed:

$$\textsc{Hmi}\left({}_m\bot^{i}_{j}\right) \quad = {}_m\bot^{4m-i}_{8m-1-j}$$

3 Application in Robotics Contexts

3.1 Ambigue Landmark Problems

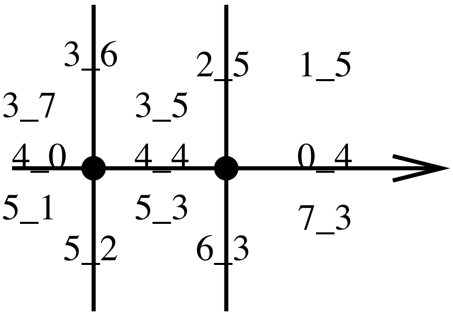


Figure 3. Double Cross reference system/partition

We can use the Double Cross calculus to represent our underdetermined spatial knowledge of the robotics example depicted in figure 4. The robot's observation at time point 1 (the red landmarks are close and can be distinguished, the green ones are to far away to be distinguished):

$$R1, R2 \quad (2\_5, 3\_6) \quad G1 \tag{1}$$

$$R1, R2 \quad (2\_5, 3\_6) \quad G2 \tag{2}$$

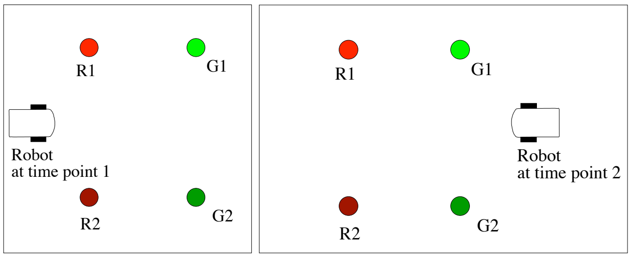


Figure 4. Two observation resulting in ambigue spatial knowlede

The robot's observation at time point 2 (the green landmarks are close and can be distinguished, the red ones are to far away to be distinguished):

$$G1, G2 \quad (5\_2, 6\_3) \quad R1 \tag{3}$$

$$G1, G2 \quad (5\_2, 6\_3) \quad R2 \tag{4}$$

The observation corresponding to equation (4) can be reformulated:

$$G1, R2 \quad (3\_5, 3\_6) \quad G2 \tag{5}$$

It follows:

$$R2, G1 \quad \textsc{Inv}\,(3\_5, 3\_6) \quad G2 \tag{6}$$

$$R1, R2 \quad (2\_5, 3\_6) \Diamond \textsc{Inv}\,(3\_5, 3\_6) \quad G2 \tag{7}$$

$$R1, R2 \quad (3\_5, 2\_5, 1\_5) \quad G2 \tag{8}$$

The conjunction (intersection) of equation (2) and equation (8) yields:

$$R1, R2 \quad 2\_5 \quad G2 \tag{9}$$

This manual deduction shows how the ambiguity is resolved in this landmark configuration. In general the observations can be represented in a constraint network and standard constraint propagation solves the ambiguity problem.

However, since the Double Cross calculus is coarse only special configurations of landmarks can be solved with this formalism. More fine grained calculi like the $\textsc{Gpcc}_m$ calculi are capable of solving much more general problems. This approach is ongoing work, first results are promising.

3.2 Representing perceived spatial knowledge

In robotic applications the relevant areal base relations with their borders are summarized into general relations. Out of this, one obtains a closed region in a plane (with the exception of its exterior segments which continue infinitely) as acceptance area for the third point of a ternary relational proposition. The bounded line segment acceptance areas belong to both neighboring segments and border points typically belong to four segments. All inner segments contain the point which corresponds to the relation sam.

The areal measure of these ambiguous acceptance areas is however 0. In the event that a corresponding border point triple is to be represented qualitatively, a disjunction of all bordering base relations must be used. As a result one obtains then a fine grained quasi-partition for the representation of the relative position of a point with respect to a reference system build by two points.

4 Conclusion

We presented a calculus for representing and reasoning about qualitative relative orientation information. We identified systems of atomic relations on different granularity levels.

Potential applications of the calculus were motivated by a robotics scenario. In the scenario, the disambiguation of landmarks is achived by constraint-propagation only, since the underdetermined spatial knowledge about the landmark position can be expressed as constraint networks.

Acknowledgment

The author would like to thank Frank Dylla, Marco Ragni, Diedrich Wolter, and Christian Freksa for interesting and helpful discussions related to the topic of the paper. The work was supported by the DFG Transregional Collaborative Research Center SFB/TR 8 “Spatial Cognition”.

REFERENCES

[1] E. Clementini, P. Di Felice, and D. Hernandez, ‘Qualitative represenation of positional information’, *Artificial Intelligence*, **95**, 317–356, (1997).

[2] A.G. Cohn, ‘Qualitative spatial representation and reasoning techniques’, in *KI-97: Advances in Artificial Intelligence*, eds., G. Brewka, C. Habel, and B. Nebel, Lecture Notes in Artificial Intelligence, 1–30, Springer-Verlag, Berlin, (1997).

[3] A. Frank, ‘Qualitative spatial reasoning with cardinal directions’, in *Proceedings of 7th Österreichische Artificial-Intelligence-Tagung*, pp. 157–167, Berlin, (1991). Springer.

[4] C Freksa, ‘Using orientation information for qualitative spatial reasoning’, in *Proceedings of International Conference on Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, eds., A U Frank, I Campari, and U Formentini, Berlin, (1992). Springer-verlag.

[5] A. Isli and A. Cohn, ‘Qualitative spatial reasoning: A new approach to cyclic ordering of 2d orientation’, *Artificial Intelligence*, **122**, 137–187, (2000).

[6] U Montanari, ‘Networks of constraints: Fundamental properties and applications to picture processing’, *Information Sciences*, **7**, 95–132, (1974).

[7] Reinhard Moratz, Bernhard Nebel, and Christian Freksa, ‘Qualitative spatial reasoning about relative position: The tradeoff between strong formal properties and successful reasoning about route graphs’, in *Lecture Notes in Artificial Intelligence 2685: Spatial Cognition III*, eds., C. Freksa, W. Brauer, C. Habel, and K.F. Wender, 385–400, Springer Verlag, Berlin/Heidelberg, (2003).

[8] Reinhard Moratz and Jan Oliver Wallgrün, ‘Spatial reasoning about relative orientation and distance for robot exploration’, in *Spatial Information Theory: Foundations of Geographic Information Science. Conference on Spatial Information Theory (COSIT)*, eds., W. Kuhn, M.F. Worboys, and S. Timpf, Lecture Notes in Computer Science, pp. 61–74. Springer-Verlag; D-69121 Heidelberg, Germany; http://www.springer.de, (2003).

[9] A. Musto, K. Stein, A. Eisenkolb, and T. Röfer, ‘Qualitative and quantitative representations of locomotion and their application in robot navigation’, in *Proceedings IJCAI-99*, pp. 1067 – 1072, (1999).

[10] J. Renz and B. Nebel, ‘On the complexity of qualitative spatial reasoning: A maximal tractable fragment of the region connection calculus’, *Artificial Intelligence*, **108(1-2)**, 69–123, (1999).

[11] C Schlieder, ‘Reasoning about ordering’, in *Spatial Information Theory: a theoretical basis for GIS*, ed., W Kuhn A Frank, number 988 in Lecture Notes in Computer Science, pp. 341–349, Berlin, (1995). Springer Verlag.

[12] A. Scivos and B. Nebel, ‘Double-crossing: Decidability and computational complexity of a qualitative calculus for navigation’, in *COSIT 2001*, Berlin, (2001). Springer.

[13] K. Zimmermann and C. Freksa, ‘Qualitative spatial reasoning using orientation, distance, path knowledge’, *Applied Intelligence*, **6**, 49–58, (1996).

Localization, Exploration, and Navigation Based on Qualitative Angle Information

Frieder Stolzenburg[1]

Abstract. Three of the major problems in qualitative spatial reasoning and robotics are localization, exploration, and navigation in known or unknown environments. This paper investigates how far different qualitative methods based on angle information, most of them originally invented for the representation of spatial knowledge only, are well-suited for these tasks. It turns out that with panoramas, which are special roundviews, the qualitative localization problem can be solved in a satisfiable manner. It is stated that the exploration problem, i.e. qualitative map building, remains difficult for all approaches. In addition, qualitative navigation operators for robot control are discussed.

1 Introduction

The main goal of qualitative spatial reasoning is to represent not only the everyday commonsense knowledge about the physical world, but also the underlying abstractions used by engineers in quantitative models (Cohn and Hazarika, 2001), in order to obtain a general model of a specific domain, as it is e.g. exemplified for robotic soccer by Dylla et al. (2005). A qualitative world model abstracts from the physical reality in order to obtain a robust and easily maintainable model. Furthermore, a qualitative model may still work, even if the exact (physical) laws are not known. Another motivation for a qualitative approach is that it is likely to be cognitively more adequate, e.g., human agents probably do not solve differential equations while interacting with their environment.

What does *qualitative* actually mean in this context? First, a qualitative representation normally is symbolic, without any continuous numerical values. The physical reality is approximated by a bounded number of states, i.e., the level of precision is decreased. Second, usually a qualitative world model yields only local information relative to the observer, i.e., the frame of reference corresponds to an egocentric point of view. Clementini et al. (1997) exemplify this e.g. for positional information. Both aspects are important for (qualitative) navigation in spatial environments: Since human and robot agents only have restricted computing resources for acting autonomously, an abstraction of the real world might be helpful. In addition, an agent has only access to local information, e.g. its internal state and a excerpt of the external world, hence it has essentially an egocentric view of the world. Following the lines of Levitt and Lawton (1990), the main questions that an agent in a spatial environment has to address are:

1. (self-)localization: where am I?
2. exploration: where are other places relative to me?
3. navigation: how do I get to other places?

In order to solve these tasks, *robot* agents may use sensors like e.g. compasses, digital video cameras, infrared sensors, laser range finders, or sonars. The corresponding numerical sensor data can be exploited by a robot. Together with probabilistic methods, imprecise or even inconsistent data can be processed (Thrun et al., 2005). However, *human* agents usually do not have access to quantitative metric information. Exact measuring of angles or distances is difficult for them. Therefore, it seems to be worthwhile to investigate qualitative approaches for localization, exploration, and navigation — not only for human but also for robot agents, whenever we are confronted with sensors, yielding very imprecise quantitative information.

In this paper, we will consider the scenario of an agent situated in a spatial environment. In order to keep things simple, we will restrict our attention mainly to two-dimensional environments or to the projection into two dimensions in this context, although most of the definitions stated here can easily be adapted to more (or less) dimensions. We assume that there is a certain number of natural or artificial landmarks in the environment, that are points without dimension, e.g. a mountain top or a beacon. Landmarks help agents to orient in known or unknown environments. A *landmark* must be a distinctive visual event, i.e., it defines a single direction, and it must be visually re-acquirable (Levitt and Lawton, 1990). Recognition of landmarks is a major problem in computer vision. We will not deal with this topic here, but suppose that, at least on a qualitative level, landmarks can be identified by the agents.

But what kind of qualitative, i.e. completely non-numerical information does an agent have available for orienting in a spatial environment? First, agents may estimate *distances* in categories like close and far. However, for humans distances are hard to measure without further remedies, and reliable distance sensors like laser range finders are nowadays still expensive. Using only odometry for (robot) navigation, i.e. tracking the distances the agent has moved among others, in order to determine the actual agent position, is often not successful. Second, agents may exploit *angle information*. Although it may be difficult to estimate angles exactly, too, qualitative angle information often is available: From a roundview, an agent may obtain (a) the (cyclic) ordering of visible landmarks. In addition, (b) left-right or other spatial relations among the landmarks (relative to the agent position) may be known. Therefore we will focus in the following on qualitative angle information (but see Sect. 7), treating both just mentioned aspects of qualitative angle information.

Let us now consider an example configuration with four landmarks, which are numbered from 1 to 4, as shown in Fig. 1. There, each of the six possible pairs of landmarks are connected by straight lines, leading to the shown tessellation, that (in this case) partitions the plane into 18 regions. By construction, all such regions are convex two-dimensional polytopes, i.e. polygons (e.g. triangles) or their unbounded counterparts. If e.g. an agent is somewhere located in region R in Fig. 1, then it sees landmark 2 left of landmark 3. We will discuss left-right relations further in Sect. 4. Furthermore, during a roundview from region R, the agent sees the landmarks in the order

[1] Hochschule Harz, Automation and Computer Sciences Department, Friedrichstr. 57-59, D-38855 Wernigerode, fstolzenburg@hs-harz.de

1234 (clockwise). Since the starting point of the roundview is not fixed, this order is cyclic and is equivalent to e.g. the order 2341. This cyclicity is problematic, as we will see in Sect. 3.

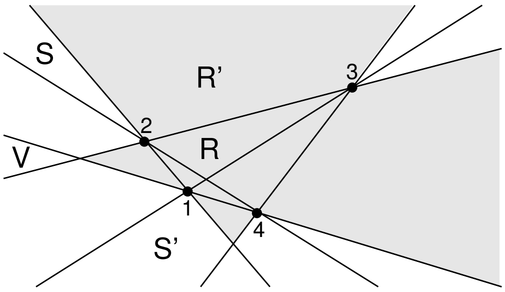


Figure 1. Example with four landmarks.

How can an agent now exploit such kind of information in qualitative localization, exploration, and navigation? This is the subject of the rest of the paper. In the next section (Sect. 2), we state more formally the basic definitions of point and line configurations based on landmarks. We discuss different means for qualitative spatial reasoning: cyclic orderings (Levitt and Lawton, 1990) and so-called aspects (Goodman and Pollack, 1993; Hübner and Wagner, 2005; Miene and Wagner, 2006) in Sect. 3, left-right relations (order types) and permutation sequences (Goodman, 1997; Goodman and Pollack, 1993) in Sect. 4, and panoramas (Schlieder, 1993, 1996) in Sect. 5. The complete set of panoramas of a configuration allows us a qualitatively exact mapping of the environment and thus addresses the exploration problem (also stated in Sect. 5). After that, qualitative operators for navigation such as crossing between two landmarks or heading toward a certain landmark are investigated (Sect. 6). Finally, we briefly compare the results stated here with related work on qualitative and also quantitative (robot) navigation (Sect. 7), including a brief excursus on distance-based configurations and position estimation. We end up with concluding remarks about qualitative spatial reasoning methods (Sect. 8).

2 Point Configurations and Line Arrangements

A *point configuration* is defined by a finite set of points P. A point configuration implicitly defines a *line arrangement*, that is the set L of all straight lines passing through pairs of points in P. Let P^* be the set of all intersection points of the straight lines in L. Clearly, it holds $P \subseteq P^*$. Note that the notion line arrangement can be defined without recursion to the notion point configuration (Goodman, 1997; Goodman and Pollack, 1993). The natural setting for line arrangements is the real projective plane. Nevertheless, we here think of them as lying in the Euclidean plane $\mathbb{R}^2$, to simplify matters. Let $n = |P|$, i.e. the number of points in P. Then, a point configuration is called *simple* iff each pair of straight lines intersects exactly once except for the points in P itself, which have multiplicity $n-1$, i.e., exactly $n-1$ lines from L are passing through them.

Each straight line of a point configuration passing through two points A and B in this direction (written as $\overline{AB}$) defines two half-spaces, namely the one to the left and the one to the right of the oriented straight line. A *region* of a configuration is a maximally connected component of the complement of the straight lines of the configuration. It can be defined as intersection of open half-spaces, i.e., a region can be identified by the set of straight lines in L, which are left or, alternatively, right of it. Tab. 1 shows the respective characterizations of the region R and R' from Fig. 1. Since both regions are immediate neighbors, their characterizations differ in only one relation: R is right of $\overline{23}$, whereas R' is left of it. Two configurations induced by the point sets P_1 and P_2, respectively, are called (qualitatively or combinatorially) *equivalent* iff they contain the same regions, where P_1 and P_2 must contain the same point identifiers.

region	left of	right of
R	$\overline{13}, \overline{14}, \overline{21}, \overline{24}, \overline{32}, \overline{43}$	$\overline{31}, \overline{41}, \overline{12}, \overline{42}, \overline{23}, \overline{34}$
R'	$\overline{13}, \overline{14}, \overline{21}, \overline{23}, \overline{24}, \overline{43}$	$\overline{31}, \overline{41}, \overline{12}, \overline{32}, \overline{42}, \overline{34}$

Table 1. Some left-right relations from Fig. 1.

Since there are $\binom{n}{2}$ pairs of landmarks and thus as many straight lines in L, in principle, $2^{\binom{n}{2}}$ different regions are possible. Obviously, not all of them occur in one configuration. In Fig. 1, there is e.g. no region that is left of $\overline{12}$, $\overline{23}$, and $\overline{31}$. But how many regions are there in a simple point configuration in general? Evidently, for $k = 2$ landmarks, we have 2 regions. For $k > 2$, already $\binom{k-1}{2}$ straight lines are there. A new point p_k introduces $k-1$ new straight lines. Each of them has $\binom{k-1}{2} - (k-2)$ intersection points for all but the first new straight line, which has one intersection point fewer. Each straight line leads to one plus the number of intersection points new regions. The total number of regions $\rho(n)$ can be computed as shown below.[2] Some values of $\rho(n)$ are listed in Tab. 2 together with other numbers that will be explained later on.

$$\begin{aligned}\rho(n) &= 2 + \sum_{k=3}^{n} \left((k-1) \left(\binom{k-1}{2} - (k-2) + 1 \right) - 1 \right) \\ &= \frac{1}{8}n^4 - \frac{3}{4}n^3 + \frac{23}{8}n^2 - \frac{13}{4}n + 1 \;=\; O(n^4)\end{aligned}$$

n	2	3	4	5	6
$\rho(n)$	2	7	18	41	85
$(n-1)!$	1	2	6	24	120
$n!$	2	6	24	120	720
$2^{n-1}(n-1)!$	2	8	48	384	3840
$\rho^*(n)$	2	6	18	46	101

Table 2. Numbers of regions for simple configurations.

3 Localization with Cyclic Orderings and Aspects

If quantitative data is not available, clearly the exact positions of landmarks cannot be computed. Only a qualitative map of the environment (i.e. the corresponding collection of regions) can be constructed by the agent. In particular, exact self-localization is not possible, if only qualitative angle information is available. Nevertheless, it should be possible to determine the region in which the agent is, where different regions should be represented by distinct identifiers. For this purpose, we will investigate roundviews at first, as done by Levitt and Lawton (1990) among others, that can be obtained by so-called omnivision cameras, allowing a view angle of 360° in total. Such cameras are used e.g. in RoboCup robotic soccer competitions. If we assume that the agent is never located exactly on one of the straight lines in L, then a *qualitative roundview* of the agent in clockwise order gives us the cyclic ordering of the n given, visible landmarks. It is characterized by a sequence of the n landmark identifiers. Because of the cyclicity, the first landmark in this sequence can be chosen arbitrarily. Therefore with the roundview approach, only $(n-1)!$ regions can be distinguished.

However, Tab. 2 reveals that this number is smaller than the number of regions $\rho(n)$ for $n \leq 5$. Hence, there must be regions in Fig. 1, which have identical cyclic orderings. In fact, this is the case e.g. for

[2] The number of (orientation) regions is given without proof in Levitt and Lawton (1990).

the regions R and R'. More generally, when crossing a connecting line between any two landmarks, then the cyclic ordering does not change. Hence, all shaded regions in Fig. 1 are characterized by one and the same cyclic ordering 1234. Even worse, for $n \geq 4$, it might be the case that regions, which are not even immediate neighbors, are associated with the same cyclic ordering. For the example in Fig. 1, this holds e.g. for the regions S and S' with associated cyclic orderings 3421 and 2134, respectively, which are identical.

It turns out, that the problem with cyclic orderings is their cyclicity. Alternatively, regions can be identified uniquely by means of what we call aspects (Goodman and Pollack, 1993), if we restrict ourselves to points in regions outside the convex hull of all landmarks. In this case, an *aspect* is defined by the permutation of the n landmarks in the order from left to right. For this, we first must define the *left-right relation* (called orientation of landmark pair boundaries by Levitt and Lawton, 1990) more precisely: landmark A is left of landmark B iff the azimuth (clockwise oriented) angle $\measuredangle(A,B)$ seen from the current point of view (agent position or region) is between 0° and 180° (exclusively). Hence, in order to measure a left-right relation, we must be able to detect 180° angles by means of sensors. By using a camera with view angle equal or smaller than 180°, this is certainly possible.

Since for all points in regions outside the convex hull, there always is a leftmost and rightmost visible landmark, thus aspects are in contrast to simple roundviews non-cyclic. Because of this, up to $n!$ aspects can be distinguished, which is always greater than $\rho(n)$ (see Tab. 2). Fig. 2 shows for the running example, that all regions outside the convex hull (shaded) have different aspects. This holds in general for any configuration, because, for regions outside the convex hull, the left-right relation is a linear order and can easily be read off the aspects: landmark A is left of landmark B or, stated differently, the point of view is right of $\overline{AB}$ iff A occurs before B in the respective permutation. Hence, each aspect uniquely determines a certain intersection of open half-spaces, i.e. a region.

In the context of the RoboCup robotic soccer scenario, aspects have been applied successfully in a case study with four-legged robots (Hübner and Wagner, 2005; Miene and Wagner, 2006). There, the robots move around the goal area and try to localize themselves.

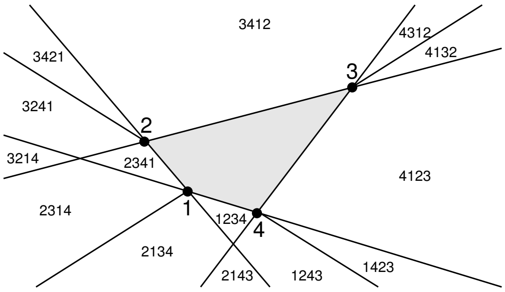


Figure 2. Aspects of regions outside the convex hull.

4 Order Types and Permutation Sequences

Cyclic orderings and aspects exploit the left-right relations between landmarks seen from a point of view from within a region. If we collect the order information of all triples of landmarks (i.e. allowing landmarks as points of view so to speak), we get the so-called *order type* of a configuration. We say, the triangle of landmarks ABC is ordered positively iff C is left of $\overline{AB}$, i.e., A, B, and C are oriented counter-clockwise. This can easily be generalized to any number of spatial dimensions $d \geq 1$ by means of determinants (see e.g. Goodman and Pollack, 1983): a sequence $p_0 \cdots p_d$ of $d+1$ points in $\mathbb{R}^d$ with $p_i = (x_{i1}, \cdots, x_{id}, 1)$ for $0 \leq i \leq d$ has *positive orientation* iff $\det(x_{ij}) > 0$.

Unfortunately, there are non-equivalent configurations which have the same order type. Thus in general, order types do not uniquely determine a configuration qualitatively. In order to see this, look at the configuration in Fig. 3, that shows the reflection of the one in Fig. 1, where the landmarks are numbered similarly in both cases. Both configurations have the same order type, and they consist of the same regions except for the regions V and V', respectively. Therefore in summary, it is not sufficient to get the left-right relations for the n kernel points (the landmarks), in order to identify a configuration qualitatively.

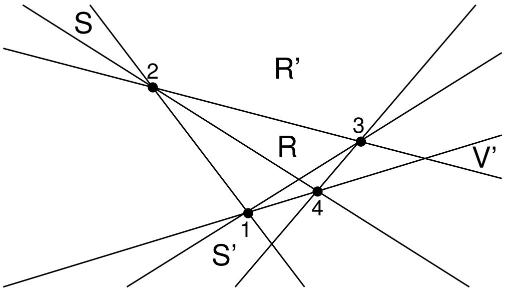


Figure 3. Reflection of the example in Fig. 1.

In order to overcome this problem, the notion permutation sequence of a configuration has been introduced in the literature (Goodman, 1997; Goodman and Pollack, 1993). For this, we project all landmark points orthogonally onto a straight reference line r, e.g. one of the two coordinate axes, thus again obtaining a permutation of the landmarks determined by the order in which the points fall onto r. Let now r rotate counter-clockwise. Then a new permutation arises whenever r passes through a direction orthogonal to one of the connecting lines $l \in L$, say $\overline{AB}$. If the landmarks A and B appear in the induced permutation in that order before r passed orthogonally to l, both landmarks will appear in reverse order in the permutation induced on r just after. The reversal of the landmarks in the *move* from one permutation to the next is called *switch*.[3] Allowing r to continue rotation, we obtain the circular, doubly infinite *permutation sequence* associated with the given configuration. This sequence is clearly periodic —the period corresponds to a full rotation of r—, and is in fact determined by a half period (Goodman and Pollack, 1993). Fig. 4 illustrates these definitions for the configuration of Fig. 1. Starting with the x-axis as rotating line r, we get the permutation 2143. After three moves with the switched lines $\overline{12}$, $\overline{42}$, and then $\overline{41}$, rotating r counter-clockwise by 90°, we arrive at the permutation 4123, which is the projection onto the y-axis.

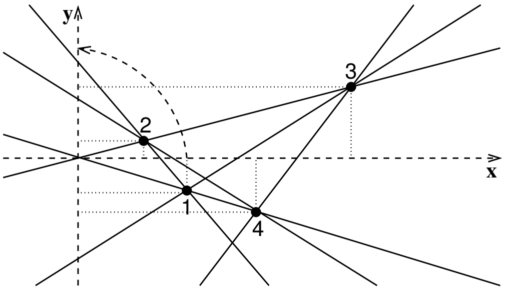


Figure 4. Projections on a straight line, rotating counter-clockwise.

Incidentally, the permutation sequences for the configurations in Fig. 4 and its reflection in Fig. 3 are different: in the original con-

[3] Note that we restrict attention to simple point configurations in this context.

figuration (Fig. 1 or 4) the switch wrt. $\overline{41}$ occurs just before the one wrt. $\overline{32}$, whereas this is the other way round in its reflection (Fig. 3). However, if we only consider the permutation sequences wrt. the given n landmarks, not taking into account the additional intersection points (as we have done so far), then the configuration is not always completely determined by its permutation sequence. To see this, look at the example with $n = 6$ landmarks, sketched in Fig. 5. There, the lines $\overline{12}$, $\overline{34}$, and $\overline{56}$ intersect in one point (namely where the shaded regions T and T' touch each other). However, slightly moving landmark 3 up or down, respectively, leads to configurations which have identical permutation sequences, although they are not equivalent, because one contains region T and the other region T' which are different. Therefore, the application of permutation sequences is of limited use, too. In addition, it remains an open practical question, how intersection points, orthogonal projections, let alone complete permutation sequences can be recorded by any concrete sensors. We have the same problems concerning symbolic projection (Schlieder, 1996) where only two qualitative projections on orthogonal axes as in architecture are considered.

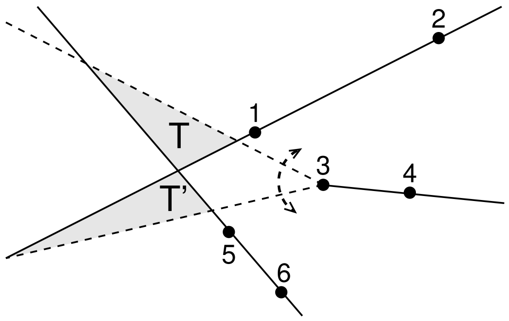


Figure 5. Problematic example for permutation sequences.

5 Panoramas and the Exploration Problem

From Sect. 4, we may conclude that a configuration can only be identified wrt. qualitative equivalence, if all regions contained in it are known. However, neither cyclic orderings nor aspects (see Sect. 3) allow us to name regions in a unique manner. To solve this problem, panoramas are introduced by Schlieder (1993, 1996): for the landmark points $p \in P$ and the point of view s, the clockwise oriented cyclic ordering of the $2n$ straight lines of the form $\overline{ps}$ and $\overline{sp}$ is called *panorama* of s. It corresponds to the order in which an agent sees the landmarks and their backs (although it is not easy to imagine how backs can be detected by concrete sensors, unless 180° angles can be measured). Therefore, a straight line of the form $\overline{ps}$ is abbreviated by $\overline{p}$, while $\overline{sp}$ is simply written as p.

Since panoramas have a period of length $2n$ where the second half is just the reverse of the first half, a convenient notation of a panorama is given by stating only the first half of a period starting with an arbitrary landmark or its back. Hence, up to $2^{n-1}(n-1)!$ panoramas can be distinguished, which is always greater than $\rho(n)$ (see Tab. 2). A panorama always defines a unique region in a configuration, because it determines all left-right relations as follows: a point s is left of the line $\overline{pq}$ iff $pq\overline{pq}$ occurs in the panorama of s. For the example configuration in Fig. 1, the panorama of region R is $1\overline{3}2\overline{41}3\overline{2}4$, that of R' is $12\overline{3412}34$. In both cases, stating only the first half would suffice. Note that deleting the backs $\overline{p}$ of all landmarks in the panoramas yields us the ordinary cyclic ordering of the respective region (here: 1234 in both cases).

In summary, panoramas finally solve the localization problem (see Sect. 1), because a unique identification of regions is possible with them. Can they also solve the exploration problem? The answer is yes, but in order to build a map of an unknown environment, a robot agent has to explore, i.e. to visit many regions. We have seen in Sect. 4, that the order type does not completely specify the configuration. In addition, the example in Fig. 2 teaches us that it is not sufficient to determine the aspects of some regions around the convex hull. For this example (see also Fig. 1) it is necessary to find out whether region V is contained in it or not, in order to distinguish this configuration from the one in Fig. 3 with the region V'. The example can easily be generalized by repeatedly putting one of both configurations (Fig. 1 or Fig. 3) on top of the other. Then at least all of the copies of the regions V or V', respectively, have to be visited, i.e. at least $\frac{1}{18}n$. Hence, the exploration problem, i.e. qualitative map building, turns out to be very expensive, since the number of regions to be visited cannot be bounded by a constant or a logarithmic function, as the example shows.

This negative result for qualitative exploration seems to be related to the stretchability problem in discrete geometry (Goodman, 1997; Goodman and Pollack, 1993). There, so-called pseudoline arrangements are considered. A *pseudoline* is a simple curve in the plane going to infinity in two directions, where any two members intersect each other at most once, and cross if they intersect. Like point configurations, a pseudoline arrangement induces a set of regions in the plane. It is called *stretchable* or realizable iff there is a straight line arrangement with the same combinatorial structure. This problems turns out to be NP-hard (Shor, 1991). If a configuration could be determined by knowing only a few regions of the configuration, the realizability problem could be answered efficiently, too.

As we have just seen, the exploration problem is difficult, if only qualitative information is available. With *quantitative* information, however, this appears to be different: in most cases, localization and map building is possible by making only a few snapshots of the environment. First, if e.g. the agent is able to measure distances and absolute orientations, i.e. wrt. to a global reference coordinate system, clearly the snapshot from only one point position is sufficient to localize the agent or one of the landmarks. Second, if the agent can measure absolute orientations only, e.g. by a compass, then two snapshots from known positions are enough for computing an unknown landmark position. Third, if only distance information is available, three or sometimes even two snapshots suffice (Levitt and Lawton, 1990). Last but not least, Betke and Gurvits (1997) describe a method for localizing a mobile robot in an environment with landmarks, where only their bearings relative to each other are given (i.e. relative angle information). Given such possibly noisy input, the algorithm estimates the robot position and orientation with respect to the map of the environment. The algorithm makes use of complex numbers and runs in time linear in the number of landmarks, employing a least squares approach.

6 Navigation

After having discussed the qualitative localization and exploration problem, let us now come to the navigation problem: how do I get to other places? If quantitative information is available and there are no obstacles around, the shortest way from one place to another obviously is simply traveling along the straight line segment connecting both places. Otherwise, exact quantitative distances cannot be measured. Then, a qualitative notion of distance is needed. Since qualitative localization cannot be more precise than determining the region the agent is in, a natural definition of qualitative distance is the number of regions that have to be passed (minimally) while going from one place to the other.

Formally, the *qualitative distance* between two regions can be defined as the number of point pairs inverted in their respective panoramas. Consider the regions X and X' in Fig. 6, whose respective panoramas are $2134\overline{2134}$ and $3241\overline{3241}$, respectively. The first panorama can be transferred into the second one by three inversions, namely 13, 23, and 14, since the corresponding lines $\overline{13}$, $\overline{23}$, and $\overline{14}$ must be crossed. Thus, the qualitative distance between the regions X and X' is 3. From this observation, Schlieder (1993) proposes a greedy algorithm for finding shortest qualitative paths. For this, the environment, i.e. the configuration has to be known in advance, however. In each step, a straight line of L is crossed, which is (still) inverted in the current panorama compared with the goal panorama. This simple greedy procedure solves the qualitative navigation problem efficiently. It is complete and optimal, i.e. it always finds a shortest path wrt. qualitative distance, provided that there are no obstacles in the way. In the latter case, a more general search procedure like A* (see e.g. Russell and Norvig, 1995), that is still optimal and complete, can be applied. In Fig. 6 one can see, that there are two possible shortest path from X to X' in the example: XUVX' and XUWX'. Note that the order in which the lines have to be crossed is not completely free.

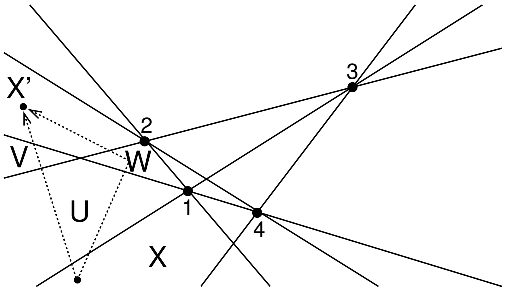


Figure 6. Finding the qualitatively shortest path.

Now the question remains, how can we come from one region to a neighbored one by means of qualitative operators. Schlieder (1993) does not answer this question, nevertheless he provides a clear representation of qualitative navigation. Levitt and Lawton (1990) propose several qualitative operators that could be used not only for cognitively adequate description (localization) but for controlling (robot) agent behavior, e.g. $at(p)$ (head toward landmark p), $btw(p_1,p_2)$ (crossing between landmarks p_1 and p_2 along the angular bisector, which leads to a hyperbolic trajectory), or $left(p_1,p_2)$ (crossing $\overline{p_1p_2}$ left of p_1). All operators are executed, until a straight line of L is reached. In the example, the operator sequence $at(2,1)$ $btw(2,1)$ $left(2,1)$ would do the job. Unfortunately, it is not possible to determine the appropriate operator from the actual panorama alone, unless the complete configuration is known. Therefore, we will not go into more details here.

7 Related Works

There are numerous related works in the field of qualitative spatial reasoning (Cohn and Hazarika, 2001), addressing the representation problem of configurations in the plane, exploiting only completely non-numerical information. Latecki and Röhrig (1993) e.g. start with the observation that humans are able to recognize the right angle, and so are able to distinguish an acute from an obtuse angle (i.e. less or more than 90°, respectively). They associate a triangle ABC with an ordered pair consisting of its triangle orientation (clockwise or counter-clockwise) and the qualitative angle at B (acute or obtuse). Although augmented configuration knowledge with this kind of qualitative angles allows a finer subdivision of the plane, it still does not enable us to distinguish the configuration in Fig. 1 from its reflection in Fig. 3, that are not qualitatively equivalent.

The CYCORD approach, based on the clockwise order of directions of straight lines (closely related to permutation sequences, see Sect. 4), allows the unified treatment of several qualitative spatial calculi (Röhrig, 1997). Its reasoning system finds out whether a set of orientations in the plane (called *constraints* in this context) is consistent (i.e. realizable in the plane), which however is NP-complete in general. A refinement of the theory (Isli and Cohn, 1998) makes not only use of the relations *left* and *right* between two directions, but also of *equal* and *opposite*. These four binary atomic relations lead to 24 ternary relations for cyclic ordering of directions in the plane, hence 2^{24} relations in total, namely including all possible unions of ternary relations. Isli and Cohn (1998) identify a subclass of the theory, whose constraint problem is tractable. Nonetheless, with this approach, a distinction of the two configurations sketched in Fig. 5 is not possible.

So far, we concentrated on exploiting only completely qualitative information, like left-right relations or ordering information. If more, yet noisy quantitative data is available, then more sophisticated procedures are possible. In Busquets et al. (2003) e.g., a multiagent approach to qualitative landmark-based navigation with triangulation is presented. Frommberger (2006) considers a simple goal-directed navigation task, where a robot agent must find a specific landmark in an unknown environment, avoiding collisions with obstacles, and proposes a solution that employs reinforcement learning. Yairi and Hori (2003) introduce a map learning method for mobile robots, employing probabilities for the co-visibility of objects. This approach makes use of the assumption that a pair of objects observed at the same time is likely to be located more closely together than others for which this is not the case. Furthermore, there are several, very successful probabilistic methods applied to what is called the simultaneous localization and map building (SLAM) problem in robotics (Thrun et al., 2005). Other methods solve the calibration problem for video cameras and allow to detect objects from photographs (photogrammetry) (Tsai, 1986).

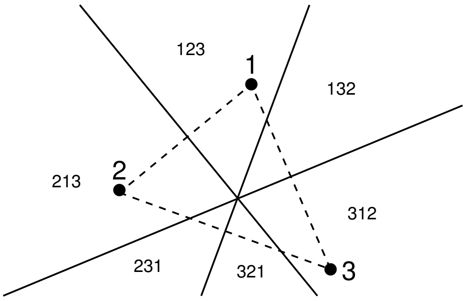


Figure 7. Regions induced by the perpendicular bisectors of all lines.

Dual to the approach followed in this paper, we may take qualitative distances into closer consideration instead of angle information. Then, from a set of n landmark points, again a configuration of regions is induced, this time by the perpendicular bisectors of the sides connecting all pairs of points. In this context, each region can uniquely be identified by one of the $n!$ orders of landmarks in increasing distance. This concept can easily be generalized to higher dimensions. Fig. 7 shows an example in the plane for $n = 3$ with six regions. Although distance-based configurations cannot be distinguished from its reflections, they may well be applied in practice, namely in robot localization based on the signal strength in wireless computer networks (WLAN) (Grabe, 2007; Ibach et al., 2004). Since the signal strength is very noisy, a qualitative approach may be a good idea there.

As in the angle-based setting, few landmarks already yield a fine tessellation of the plane. For each of the $\binom{n}{3}$ possible triangles of the given points, the three corresponding perpendicular bisectors of the side intersect in one point, namely in the center of the respective circumscribed circle. Because of this, we have $\binom{n}{3}$ regions less than in simple line arrangements with N lines in the plane. According to Graham et al. (1994), N lines lead to $\frac{N(N+1)}{2}+1$ regions in a simple line arrangement, where always exactly two lines intersect in one point. Since we have $N=\binom{n}{2}$ perpendicular bisectors of the sides, as the number of distance-based regions $\rho^*(n)$ we get the following formula (see also Tab. 2):

$$\begin{aligned}\rho^*(n) &= \left(\frac{N(N+1)}{2}+1\right)-\binom{n}{3}\\ &= \frac{1}{8}n^4-\frac{5}{12}n^3+\frac{7}{8}n^2-\frac{7}{12}n+1 \;=\; O(n^4)\end{aligned}$$

8 Conclusions and Future Work

In this paper, we discussed how far different qualitative spatial representation methods are suited to solve the localization, exploration, and navigation problem. It turns out that panoramas allow to solve the qualitative localization problem adequately. As shown in this paper, the qualitative and discrete exploration problem is difficult, compared with its quantitative counterpart, because always a certain fraction of regions has to be visited which cannot be bounded by a constant or a logarithmic function. Finally, qualitative navigation operators for robot control are discussed. In future work, it should be investigated how qualitative and quantitative spatial methods for localization, exploration, and navigation can benefit more from each other, e.g. in the context of photogrammetry, i.e. recognizing objects and positions from several snapshots in more than two dimensions.

REFERENCES

M. Betke and L. Gurvits. Mobile robot localization using landmarks. *IEEE Transactions on Robotics and Automation*, 13(2):251–263, 1997.

D. Busquets, C. Sierra, and R. L. de Màntaras. A multiagent approach to qualitative landmark-based navigation. *Autonomous Robots*, 15(2):129–154, 2003.

E. Clementini, P. Di Felice, and D. Hernández. Qualitative representation of positional information. *Artificial Intelligence*, 95(2): 317–356, 1997.

A. G. Cohn and S. M. Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 43(1): 2–32, 2001.

F. Dylla, A. Ferrein, G. Lakemeyer, J. Murray, O. Obst, T. Röfer, F. Stolzenburg, U. Visser, and T. Wagner. Towards a league-independent qualitative soccer theory for RoboCup. In D. Nardi, M. Riedmiller, C. Sammut, and J. Santos-Victor, editors, *RoboCup 2004: Proceedings of International RoboCup Symposium*, LNAI 3276, pages 611–618. Springer, Berlin, Heidelberg, New York, 2005.

L. Frommberger. A qualitative representation of structural spatial knowledge for robot navigation with reinforcement learning. In *Proceedings of Workshop on Structural Knowledge Transfer for Machine Learning*, Pittsburgh, PA, 2006.

J. E. Goodman. Pseudoline arrangements. In J. E. Goodman and J. O'Rourke, editors, *Handbook of Discrete and Computational Geometry*, chapter 5, pages 83–109. CRC Press, Boca Raton, New York, 1997.

J. E. Goodman and R. Pollack. Multidimensional sorting. *SIAM Journal on Computing*, 12(3):484–507, 1983.

J. E. Goodman and R. Pollack. Allowable sequences and order types in discrete and computational geometry. In J. Pach, editor, *New Trends in Discrete and Computational Geometry*, chapter 5, pages 103–134. Springer, Berlin, Heidelberg, New York, 1993.

M. Grabe. Qualitative Distanz-basierte Positionsbestimmung. Diplomarbeit, Fachbereich Automatisierung und Informatik, Hochschule Harz, 2007. To appear.

R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, Reading, MA, 2nd edition, 1994.

K. Hübner and T. Wagner. An egocentric qualitative spatial knowledge representation based on ordering information for physical robot navigation. In D. Nardi, M. Riedmiller, C. Sammut, and J. Santos-Victor, editors, *RoboCup 2004: Proceedings of International RoboCup Symposium*, LNAI 3276, pages 134–149. Springer, Berlin, Heidelberg, New York, 2005.

P. Ibach, T. Hübner, and M. Schweigert. MagicMap – kooperative Positionsbestimmung über WLAN. In *Chaos Communication Congress Proceedings*, Berlin, 2004.

A. Isli and A. G. Cohn. An algebra for cyclic ordering of 2D orientation. In *Proceedings of 15th American Conference on Artificial Intelligence*, pages 643–649, Madison, WI, 1998. AAAI/MIT Press.

L. J. Latecki and R. Röhrig. Orientation and qualitative angle for spatial reasoning. In *Proceedings of 13th International Joint Conference on Artificial Intelligence*, pages 1544–1549, Chambéry, France, 1993. IJCAI Inc., San Mateo, CA, Morgan Kaufmann, Los Altos, CA. Volume 1.

T. S. Levitt and D. T. Lawton. Qualitative navigation for mobile robots. *Artificial Intelligence*, 44(3):305–360, 1990.

A. Miene and T. Wagner. Static and dynamic qualitative spatial knowledge representation for physical domains. *KI*, 2/06:30–35, 2006.

R. Röhrig. Representation and processing of qualitative orientation knowledge. In G. Brewka, C. Habel, and B. Nebel, editors, *KI-97: Advances in Artificial Intelligence – Proceedings of the 21st Annual German Conference on Artificial Intelligence*, LNAI 1303, pages 219–230, Freiburg, 1997. Springer, Berlin, Heidelberg, New York.

S. Russell and P. Norvig. *Artificial Intelligence – A Modern Approach*. Prentice Hall, Englewood Cliffs, NJ, 1995.

C. Schlieder. Representing visible locations for qualitative navigation. In N. Piera Carrete and M. G. Singh, editors, *Qualitative Reasoning and Decision Technologies*, pages 523–532. CIMNE, Barcelona, 1993.

C. Schlieder. Ordering information and symbolic projection. In S. K. Chang, E. Jungert, and G. Tortora, editors, *Intelligent Image Database Systems*, pages 115–140. World Scientific, Singapore, 1996.

P. W. Shor. Stretchability of pseudolines is NP-hard. In P. Gritzman and B. Sturmfels, editors, *Applied Geometry and Discrete Mathematics – The Victor Klee Festschrift*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science 4, pages 531–554. American Mathematical Society, Providence, RI, 1991.

S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, Cambridge, MA, London, 2005.

R. Y. Tsai. A versatile camera calibration technique for high accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4):323–344, 1986.

T. Yairi and K. Hori. Qualitative map learning based on co-visibility of objects. In *Proceedings of 18th International Joint Conference on Artificial Intelligence*, pages 183–188, 2003.

An ontology of spatial relations using fuzzy concrete domains

Céline Hudelot[1] and **Jamal Atif**[2] and **Isabelle Bloch**[3]

Abstract. We propose an ontology of spatial relations, in order to guide image interpretation and the recognition of the structures it contains using structural information on the spatial arrangement of these structures. This ontology is then enriched by fuzzy representations of concepts, using fuzzy concrete domains, which define their semantics, and allow establishing the link between these concepts (which are often expressed in linguistic terms) and the information that can be extracted from images. This contributes to reduce the semantic gap in image interpretation.

1 INTRODUCTION

The importance of semantics in images has been highlighted in different domains such as scene analysis, image interpretation, content based indexing of digital images, among others. The image semantics cannot be considered as being included explicitly in the image itself. It rather depends on prior knowledge on the domain and the context of the image. Introducing knowledge in the image interpretation process is not a new idea, as evidenced by the numerous work on knowledge based systems for computer vision (see for instance a review in [10]). However this type of approach suffers from several shortcomings, in particular because of the lack of genericity (many systems are rather ad hoc), and the difficulty of acquiring and representing prior knowledge. Recent developments in the field of knowledge engineering, including ontology engineering, allow answering some of these questions [19]. The use of ontologies is also widening in the domain of image indexation [29]. However the development of ontology based methods for image interpretation is still in its infancy.

As opposed to the domain of analysis and indexation of textual documents, in which ontologies are widely used and became an almost unavoidable support, the domain of image interpretation and semantic indexing has to face the difficult problem of matching the perceptual level and the conceptual level. The perceptual level consists of features, mainly pixels (in 2D), voxels (in 3D), or groups of pixels or voxels, while the concepts are usually expressed in plain text and have a linguistic nature (in our example, we deal with a controlled vocabulary). This problem is often referred to as the *semantic gap* [28]. It is close to the problem of symbol grounding or anchoring addressed in artificial intelligence [17] and in robotics [9].

An important type of knowledge that guides spatial reasoning (and therefore image interpretation) consists of spatial relations. Therefore our research focuses on image interpretation based on prior knowledge on the spatial organization of the observed structures.

In this paper, we propose to reduce the semantic gap between numerical information contained in the image and higher level concepts by enriching ontologies with a fuzzy formalism layer. Fuzzy representations have several advantages, including the representation of imprecision in the definition of concepts (in particular concerning spatial relations), in expert knowledge, and the reasoning under such imprecision. More specifically, we introduce an ontology of spatial relations (Section 2) and propose to enrich it by fuzzy representations of these relations in the spatial domain (Section 3). The integration between the ontology and the fuzzy models is performed through fuzzy concrete domains. As another contribution of this paper, we show how this enriched ontology can support the reasoning process in order to recognize structures in images. In Section 4, an example of brain structure recognition is presented. It illustrates the potential of the proposed fuzzy spatial relation ontology.

2 AN ONTOLOGY OF SPATIAL RELATIONS

As mentioned in [3], several ontological frameworks for describing space and spatial relations have been developed recently. In spatial cognition and linguistics, the project OntoSpace[4] aims at developing a cognitively-based commonsense ontology for space. Some interesting works on spatial ontologies can also be found in Geographic Information Systems (GIS) [7, 21], in object recognition in images or videos [12, 16], in robotics [13, 24], or in medicine concerning the formalization of anatomical knowledge [11, 14, 27]. All these ontologies concentrate on the representation of spatial concepts according to the application domains. They do not provide an explicit and operational mathematical formalism for all the types of spatial concepts and spatial relations. For instance, in medicine, these ontologies are often restricted to concepts from the mereology theory [14], and there is still a gap to fill before using them for image interpretation.

Moreover, to our knowledge, none of these ontologies takes into account the vagueness and the subjectivity of spatial information, even if a lot of frameworks for spatial knowledge representation and spatial reasoning under imprecision have been proposed. As mentioned in [18], a combination of both ontology of concepts and uncertainty management is necessary for real world applications. In this work, the authors propose to model uncertainty in the Dempster-Shafer framework, with applications to room concepts. The uncertainty is attached to each object of a scene, spatial relations are not directly involved. An interesting work dedicated to the representation of uncertain, subjective and vague temporal knowledge in ontologies has been proposed in [25]. A fuzzy temporal model is integrated into an ontology by using fuzzy concrete domains (fuzzy intervals). We propose to develop similar ideas for the representation

[1] Ecole Centrale Paris, France - celine.hudelot@ecp.fr

[2] Université des Antilles et de la Guyane, France - jamal.atif@guyane.univ-ag.fr

[3] Ecole Nationale Supérieure des Télécommunications (GET - Télécom Paris) - CNRS UMR 5141 LTCI - Paris, France - Isabelle.Bloch@enst.fr

[4] http://www.ontospace.uni-bremen.de/twiki/bin/view/Main/WebHome

of spatial knowledge. In particular, we propose a generic spatial ontology enriched with fuzzy representations of spatial concepts in the image concrete domain. This modular representation enables to keep abstract generic spatial concepts separated from their application dependent representation. Moreover, it provides a unified framework for the representation of spatial information in images and it makes image processing and interpretation easier.

An excerpt of the hierarchical organization of spatial relations in our ontology is displayed in Figure 1. It follows the distinction between topological and metric relations proposed in [22]. It should be noted that in both cases, relations can be given in a qualitative or quantitative form. In the following, we summarize the main concepts which have been highlighted in the literature for their importance for spatial reasoning, and which are therefore integrated in our ontology. As typical examples of relations for which all mentioned questions are raised, we have chosen "**close to**" and "**to the right of**" in order to illustrate our purpose all through the paper.

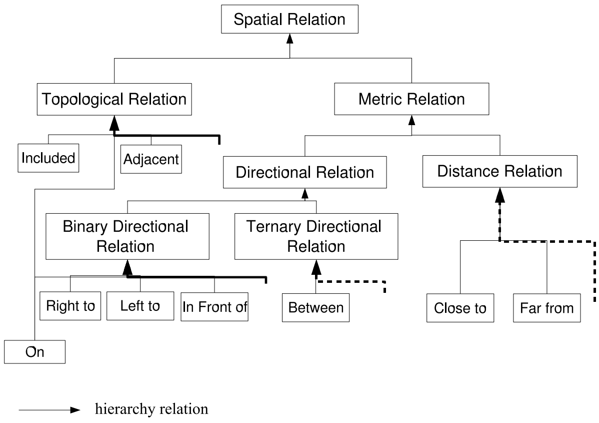


Figure 1. Excerpt of the hierarchical organization of spatial relations in our ontology.

2.1 Reference system

Expliciting spatial relations, in particular metric relations, requires a reference system. Let us consider the example of the directional relation "**x is to the right of y**" (see Figure 3). The semantics of the relation is not the same depending on whether the reference system is object **y** itself or an external observer. In order to define a binary relation between two objects, at least the three following concepts have to be specified: the target object, the reference object and the reference system. Works in spatial cognition have intensively addressed this question [20]. In general, a reference system is categorized either from the observer's point of view (which can be relative or absolute), or according to the way the relation is used (intrinsic, extrinsic or deictic use). It is therefore important to integrate the notion of reference system in the ontology. In the proposed spatial relation ontology, each metrical relation (directional relation or distance) is linked explicitly to a given reference system and the use of the relation requires defining the reference system associated to the relation. This view of reference system is very simple at this point, but sufficient for our actual framework. A deeper account of reference systems is left for future work.

2.2 Formal representation of spatial relations

We now describe the formalization of the different types of spatial relations which is necessary to clarify the user's diverse understanding of spatial relations and to automate spatial reasoning. The notations used in the following are the classical notations of description logics. One important entity of our ontology is the concept ***SpatialObject*** ($\mathsf{SpatialObject} \sqsubseteq \top$). Moreover, as mentioned in [23], the nature of spatial relations is twofold: they are concepts with their own properties but they are also links between concepts. For instance, the assertion "X **is to the right of** Y" can be interpreted and represented in two different ways:

1. as an "abstract" relation between X and Y that is either true or false;
2. as a physical spatial configuration between the two spatial objects X and Y.

As a consequence, we use a process of reification of spatial relations as in [23]. A spatial relation is not considered in our ontology as a role (property) between two spatial objects but as a concept on its own (***SpatialRelation***). Figure 2 represents the Venn diagram of the different concepts of the spatial relation ontology.

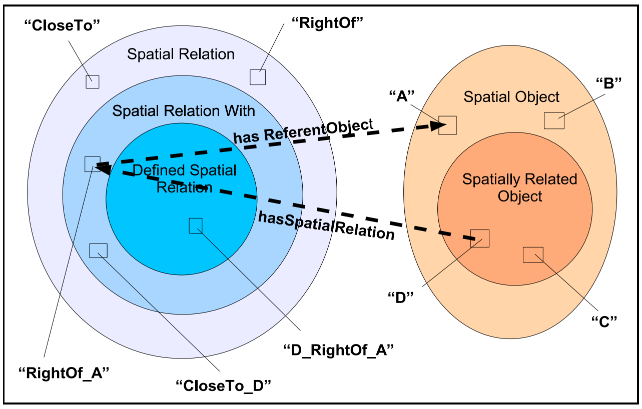


Figure 2. Representation of the main concepts of the spatial relation ontology.

- A ***SpatialRelation*** subsumes the general concept ***Relation***. It is defined according to a ***ReferenceSystem***.
$$\begin{aligned}\mathsf{SpatialRelation} \quad & \sqsubseteq \mathsf{Relation} \sqcap \\ & \ni \mathsf{type}.\{\mathsf{Spatial}\} \sqcap \\ & \exists\ \mathsf{hasReferenceSystem}.\mathsf{ReferenceSystem}\end{aligned}$$
SpatialRelation is subsumed into ***TopologicalRelation*** and ***MetricRelation*** which is itself subsumed into ***DirectionalRelation*** and ***DistanceRelation*** as shown in Figure 1. For ***BinarySpatialRelation***, we can also specify *inverse spatial relations* and properties such as *reflexivity, irreflexivity, symmetry, antisymmetry, asymmetry* useful for qualitative spatial reasoning as shown in [23].
- We define the concept ***SpatialRelationWith*** which refers to the set of spatial relations which are defined according to at least one or more reference spatial objects.
$$\begin{aligned}\mathsf{SpatialRelationWith} \quad & \equiv \mathsf{SpatialRelation} \sqcap \\ & \exists\ \mathsf{hasReferentObject}.\mathsf{SpatialObject} \sqcap \\ & \geq 1\ \mathsf{hasReferentObject}\end{aligned}$$
- We define the concept ***SpatiallyRelatedObject*** which refers to the set of spatial objects which have at least one spatial relation with

another spatial object. This concept is useful to describe spatial configurations.

$$\begin{aligned} \text{SpatiallyRelatedObject} \quad & \equiv \text{SpatialObject} \sqcap \\ & \exists\ \text{hasSpatialRelation.SpatialRelationWith} \sqcap \\ & \geq 1\ \text{hasSpatialRelation} \end{aligned}$$

It should be noted that this concept is generic. Depending on the context, we may have knowledge about the relations an object may have to another one or not. This concept allows representing this knowledge if it is available, but we may also have object concepts in the ontology for which no spatial relation to another object is defined.

- At last, the concept ***DefinedSpatialRelation*** represents the set of spatial relations for which target and reference objects are defined.

$$\begin{aligned} \text{DefinedSpatialRelation} \quad & \equiv \text{SpatialRelation} \sqcap \\ & \exists\ \text{hasReferentObject.SpatialObject} \sqcap \\ & \geq 1\ \text{hasReferentObject} \sqcap \\ & \exists\ \text{hasTargetObject.SpatialObject} \sqcap \\ & = 1\ \text{hasTargetObject} \end{aligned}$$

This distinction between ***SpatialRelation***, ***SpatialRelationWith***, ***SpatiallyRelatedObject*** and ***DefinedSpatialRelation*** is important. Indeed, the meaning of *Right_Of*, *Right_Of_Y* and *X is to the Right_Of_Y* is not the same as illustrated in Figure 3 where an absolute frame of reference is considered.

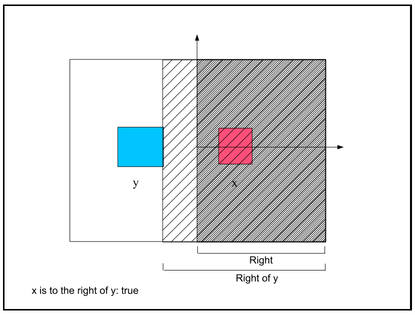
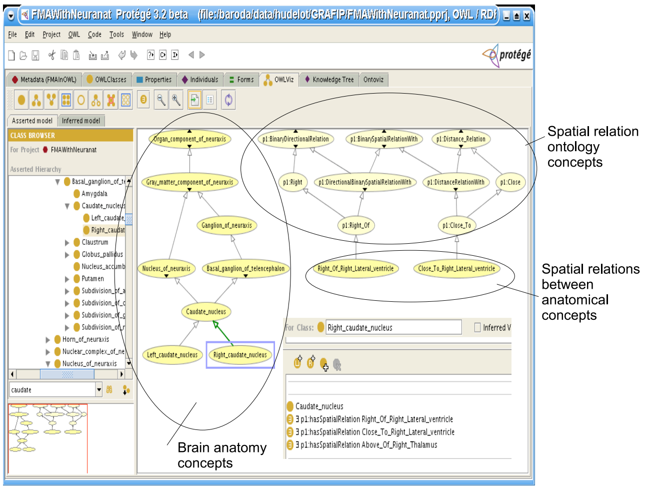


Figure 3. Three different concepts: **right** (according to a reference frame, **right of y** (with respect to a reference object), and **x is to the right of y**.

The ontology of spatial relations has been developed with the software Protégé OWL[5] and can be obtained on demand. Figure 4 represents how this ontology can be used to describe structures in specific domains. In this figure, the ontology is imported in an ontology of the brain anatomy (excerpt of the Foundational Model of Anatomy (FMA) [26]) and is used to describe the spatial organization of brain anatomical components. We consider that each physical anatomical component is a spatial object. Then, spatial relations between these different spatial objects are described by using the spatial relation ontology. For instance, as illustrated in Figure 4, the right caudate nucleus is to the right and close to the right ventricle and above the right thalamus.

3 FUZZY REPRESENTATIONS OF SPATIAL RELATIONS IN THE ONTOLOGY

We propose to introduce the imprecision in the ontology of spatial relations through fuzzy concrete domains and by using fuzzy representations of spatial relations [5]. The integration of the fuzzy model in the ontology follows a similar approach as the one in [25]. The



Caudate_nucleus
∃ p1:hasSpatialRelation Right_Of_Right_Lateral_ventricle
∃ p1:hasSpatialRelation Close_To_Right_Lateral_ventricle
∃ p1:hasSpatialRelation Above_Of_Right_Thalamus

Figure 4. Part of an ontology of the brain anatomy (excerpt of the FMA [26]). The concepts of the spatial relation ontology are prefixed by **p1**.

syntactic integration relies on concrete domains: it defines how to physically link a concept of the ontology with the fuzzy model. As opposed to the approach of [25], we propose a multiple integration. We present in this section our fuzzy models of spatial relations and their integration in the ontology.

Each object is associated to its fuzzy concrete domain. In the proposed approach, it is a spatial fuzzy set and the association is performed through the attribute *has for fuzzy concrete domain*. A spatial fuzzy set is a fuzzy set defined on the image space, denoted by $\mathcal{S}$. Its membership function μ (defined from $\mathcal{S}$ into $[0, 1]$) represents the imprecision on the spatial definition of the object (its position, size, shape, boundaries, etc.). For each point x of $\mathcal{S}$ (pixel or voxel in digital 2D or 3D images), $\mu(x)$ represents the degree to which x belongs to the fuzzy object. Objects defined as classical crisp sets are but particular cases, for which μ takes only values 0 and 1. In the following, all definitions will include the crisp case as a particular case, so that the complete framework applies for both crisp and fuzzy objects and relations. The examples in Section 4 use crisp objects and fuzzy spatial relations.

The complement of an object defined by its membership function μ is classically defined by the membership function $c(\mu)$ where c is a fuzzy complementation (typically $c(a) = 1 - a$).

In a similar way, concepts representing spatial relations are associated to concrete domains which are fuzzy sets. They can be of various natures: fuzzy number, spatial fuzzy set, interval, angle histogram, etc. The choice of the representation depends on the relation but also on the type of question raised and the type of reasoning one wants to perform. Typically, in spatial reasoning, questions and reasoning may concern:

1. the relations that are satisfied or not between two given objects (or satisfied to some degree) (Figure 5 a);
2. the area of the space $\mathcal{S}$ where a relation to one reference object is satisfied (to some degree) (Figure 5 b, c).

It is out of the scope of this paper to detail the fuzzy definitions we rely on (see e.g. [5] for a synthesis of the existing fuzzy definitions of spatial relations). These definitions allow answering both types of questions.

[5] http://protege.stanford.edu/plugins/owl/

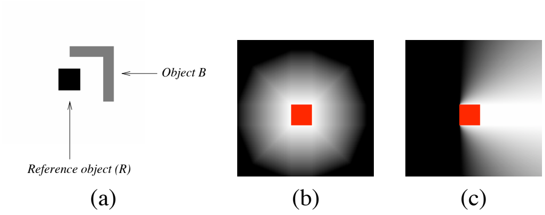


Figure 5. (a)Illustration of the first type of question: Given two objects R and B, what are the relations between them? For instance, is object B close to reference object R? Is it to the right of R? (b, c) Illustration of the second type of question: given a reference object R (the red square), what are the regions of space that satisfy a relation to it? (a) Area of space close to R, represented as a spatial fuzzy set where the membership degree at each point represents the degree to which the relation is satisfied at this point (white = 1, black = 0). (b) Area of space representing the region to the right of R.

As in [25], we follow an approach of modular semantics for integrating the fuzzy model of spatial relations with the spatial relation ontology. We combine the two various formalisms in a modular way, thus we can combine and use the best of each of them. Moreover, the separation of the abstract domain (the spatial relation ontology) from its fuzzy concrete image domain contributes to reduce the semantic gap. This integration consists in linking concepts of the spatial relation ontology to their corresponding physical fuzzy representation in the image domain. Of course, the fuzzy representation depends on the type of question. For instance, for the relation "**Right of** R", we are interested in the area of the image space where the relation right of R can be satisfied. Therefore this concept is linked to a fuzzy landscape representation, whereas the relation "**Right of**" is linked to a fuzzy subset of the set of angles representing the semantics of the relation.

Figure 6 represents the nature of integration links for directional relations. These links are implemented by the relation *has for fuzzy concrete domain*. In this figure, operators correspond to comparison operators.

As the introduction of concrete domains in OWL is based on XML Schema datatypes, we have defined a set of XML Schema datatypes in order to describe fuzzy sets, fuzzy numbers, fuzzy intervals and spatial fuzzy sets. The actual computation of the spatial relations is based on C/C++ programs. The semantics of the relations is captured by the association of the OWL representation in Protégé and the fuzzy representations coded as XML Schema datatypes.

Whereas the ontology of spatial relations is generic, the semantics of some relations can vary according to the field of application. For example a relation such as "**close to**" will not have the same meaning in a GIS context or in the context of interpretation of satellite images or of medical images. This difference is expressed in the fuzzy model, whereas the ontology of spatial relations remains a support for more general reasoning.

4 APPLICATION TO THE SEGMENTATION OF BRAIN STRUCTURES IN MRI

In this section, we show how the proposed approach can be exploited in the context of structural pattern recognition. We consider a real medical problem in brain imaging: internal brain structure recognition in magnetic resonance volumes, where the use of spatial relations is of prime importance. The elaboration of the domain ontology can benefit from the existing knowledge formalization models (such as the FMA [26]), that emerged from the medical informatics research field. While neuro-anatomy has not been much developed in these models, it is largely described in textbooks [30] and dedicated sites[6], in linguistic form. These models involve concepts that correspond to anatomical objects, their characteristics, or the spatial relations between them. This motivates the use of the ontology of spatial relations for enriching the cerebral anatomy ontology. Moreover, the semantic enrichment by the fuzzy representations of spatial relations makes it possible to formalize the ontology concepts in an operational way, that facilitates object recognition and image interpretation. In a given context, the parameters of the fuzzy representations defining the semantics of the spatial relations in this particular context can be learned, as proposed in [1] for instance.

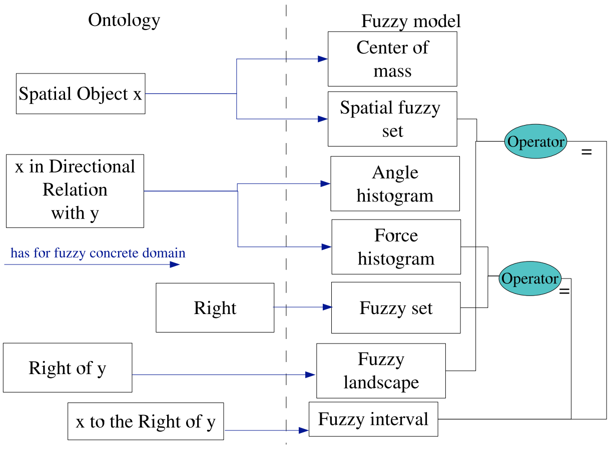


Figure 6. Syntactic integration between the spatial relation ontology and the fuzzy representation model for directional relations.

In previous work [4, 6, 8], two methods have been proposed for recognizing brain structures, a global one and a sequential one. The choice of the structures to recognize and the spatial relations that guide the recognition was entirely supervised. This constraint can now be relaxed by exploiting the features of the proposed ontology, and this constitutes an important and concrete outcome of this paper. In the following, we consider crisp spatial objects and fuzzy spatial relations.

In a **sequential approach** [6, 8], the structures are recognized successively. To detect a structure, its spatial relations with the previously recognized structures are used to reduce the search space to image areas that satisfy these relations. Let us detail the process in the case of the detection and recognition of the right caudate nucleus assuming that the right lateral ventricle has already been extracted. The situation is represented in Figure 7.

- A first step consists in extracting information from the domain ontology by querying it. The goal of the query is to find the spatial relations involving the right lateral ventricle and the right caudate nucleus. As the first one is already extracted and recognized, it is taken as a reference object. As a querying language, we use the nRQL language provided by RACER [15].
  An answer to a query concerning the caudate nucleus using our enriched domain ontology is: ***Right_Of_Right_Lateral_ventricle*** and ***Close_To_Right_Lateral_ventricle***. Indeed, according to the domain ontology "the right caudate nucleus is **to the right** and

[6] http://www.chups.jussieu.fr/ext/neuranat/index.html for instance

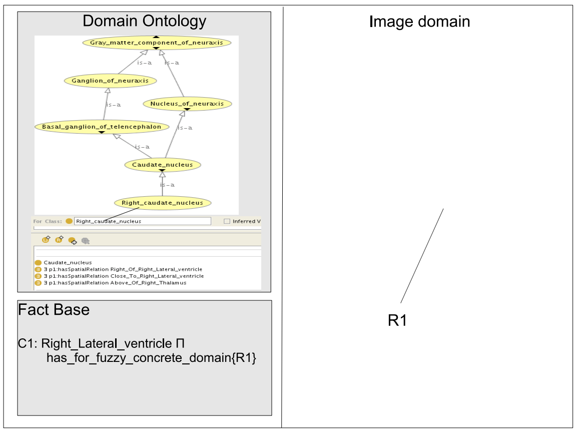

Domain Ontology
Image domain
Gray_matter_component_of_neuraxis
Ganglion_of_neuraxis
Nucleus_of_neuraxis
Basal_ganglion_of_telencephalon
Caudate_nucleus
Right_caudate_nucleus
Caudate_nucleus
∃ p1 hasSpatialRelation Right_Of_Right_Lateral_ventricle
∃ p1 hasSpatialRelation Close_To_Right_Lateral_ventricle
∃ p1 hasSpatialRelation Above_Of_Right_Thalamus
Fact Base
C1: Right_Lateral_ventricle ⊓
has_for_fuzzy_concrete_domain{R1}
R1

Figure 7. The right lateral ventricle corresponds to the spatial region R1 in the image. The domain ontology describes spatial relations between the right caudate nucleus and the right lateral ventricle. These relations will be exploited to segment the right caudate nucleus.

close to the right lateral ventricle and **above** the right thalamus" (see Figure 7). Note that the last part of this knowledge is not used here since the thalamus is not recognized yet.

- Then, according to the ontology of spatial relations, concepts such as ***Right_Of_Right_Lateral_ventricle*** or ***Close_To_Right_Lateral_ventricle*** are derived from the concept ***SpatialRelationWith*** and their syntactic integration (i.e fuzzy semantics in the image domain) corresponds here to a *fuzzy landscape* (see Figure 6). The fuzzy semantics is used to guide the operating mode (in this case, a fuzzy dilation with a structuring element defining the right direction). A similar reasoning is used for the relation **close to**, leading to another morphological operation.
- In the image domain, the search space of the "right caudate nucleus" corresponds to the area to the right and close to the right lateral ventricle, derived from the conjunctive fusion of the results of the two morphological operations, still performed in the spatial domain (Figure 8).

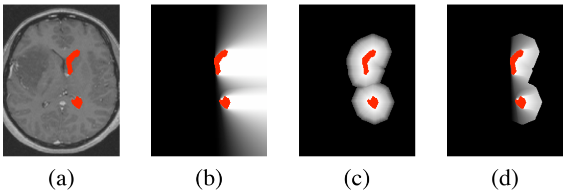


Figure 8. (a) The right ventricle is superimposed on one slice of the original image. The search space of the object "caudate nucleus" corresponds to the conjunctive fusion of the spatial relations "**to the right of the right ventricle**" (b) and "**close to the right ventricle**" (c). The fusion result is shown in (d).

The next step consists in segmenting the caudate nucleus. The fuzzy region of interest derived from the previous steps is used to constrain the search space and to drive the evolution of a deformable model [2, 8].

While in the sequential approach, segmentation and recognition are performed simultaneously, in a **global approach** [4], several objects are first extracted from the image using a segmentation method, and then recognized. The recognition can be achieved by assessing if the spatial relations between two objects x and y are those existing in the domain ontology.

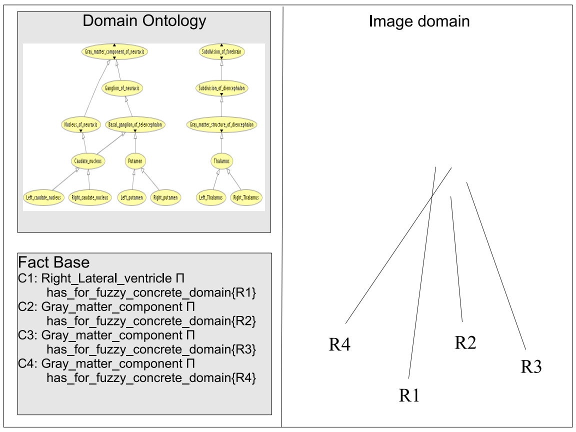

Domain Ontology
Image domain
Fact Base
C1: Right_Lateral_ventricle ⊓
has_for_fuzzy_concrete_domain{R1}
C2: Gray_matter_component ⊓
has_for_fuzzy_concrete_domain{R2}
C3: Gray_matter_component ⊓
has_for_fuzzy_concrete_domain{R3}
C4: Gray_matter_component ⊓
has_for_fuzzy_concrete_domain{R4}
R4
R2
R3
R1

Figure 9. The right lateral ventricle corresponds to the spatial region R1 on image. The domain ontology describes spatial relations between several grey nuclei and the lateral ventricles. These relations will be exploited to identify each individual structure.

- From the segmentation process (not described here), three structures that belong to the grey nuclei are extracted. The first step consists in assessing spatial relations between these structures. For the sake of simplicity we focus on relative directions. The situation is represented in Figure 9.
- We are interested in finding all the directional spatial relations between R1, R2, R3, R4, where R1 represents the lateral ventricles and R2–R4 the three regions to be labeled. The ontology of spatial relations is used to select an adequate representation for question 1, i.e the fuzzy representation of concepts "***X in directional relation with Y***" (see Section 3). The derived syntactic integration corresponds for instance here to a *histogram of angles* (see Figure 6). By using a *fuzzy interval* operating mode, the degrees of satisfaction of several directional relations between the segmented regions are computed. In this example, the following assertions yield high degrees of satisfaction: "***R2 is to the right of R1***", "***R2 is below R4***", "***R3 is to the right of R1***", "***R3 is to the right of R4***", "***R4 is to the right of R1***".
- The description of the concepts C1, C2, C3, C4 (Figure 9) is completed with the predominant directional relations between R1, R2, R3, R4 and then are classified in the hierarchy using reasoners. This allows us to label, i.e. to recognize each individual structure. In the example, structures R2, R3 and R4 are recognized as thalamus, putamen and caudate nucleus, respectively.

5 CONCLUSION

The contribution of this paper is twofold. First, an ontology of spatial relations is proposed, along with its integration with existing domain ontologies, such as the FMA for anatomical concepts. Second, this ontology is linked to fuzzy representations which define the semantics of the spatial concepts, in particular the spatial relations. This link is implemented via concrete domains. This allows adapting the semantics to a particular application, while the ontology remains general. Different types of reasoning become then possible: (i) a quite general reasoning may consist in classifying or filtering ontological concepts to answer some queries; (ii) at a more operational way, the ontology and the fuzzy representations can be used to deduce spatial reasoning operations in the images and to guide image interpretation tasks such as localization of objects, segmentation, recognition. The potential of these types of reasoning and of the proposed approach has been illustrated on a simple example in brain imaging. The enriched ontology contributes to reduce the semantic gap, which is a difficult and still open problem in image interpretation, and provides tools both for knowledge acquisition and representation and for its operational use. It has an important potential in model-based recognition that deserves to be further explored.

Acknowledgments: This work has been partly supported by grants from Région Ile-de-France, GET and ANR (post-doctoral stays of J. Atif and C. Hudelot at ENST).

REFERENCES

[1] J. Atif, C. Hudelot, G. Fouquier, I. Bloch, and E. Angelini, 'From Generic Knowledge to Specific Reasoning for Medical Image Interpretation using Graph-based Representations', in *International Joint Conference on Artificial Intelligence IJCAI'07*, pp. 224–229, Hyderabad, India, (jan 2007).

[2] J. Atif, O. Nempont, O. Colliot, E. Angelini, and I. Bloch, 'Level Set Deformable Models Constrained by Fuzzy Spatial Relation', in *Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU*, pp. 1534–1541, Paris, France, (2006).

[3] J. Bateman and S. Farrar, 'Towards a generic foundation for spatial ontology', in *International Conference on Formal Ontology in Information Systems (FOIS-2004)*, pp. 237–248, Trento, Italy, (2004).

[4] E. Bengoetxea, P. Larranaga, I. Bloch, A. Perchant, and C. Boeres, 'Inexact Graph Matching by Means of Estimation of Distribution Algorithms', *Pattern Recognition*, **35**, 2867–2880, (2002).

[5] I. Bloch, 'Fuzzy Spatial Relationships for Image Processing and Interpretation: A Review', *Image and Vision Computing*, **23**(2), 89–110, (2005).

[6] I. Bloch, T. Géraud, and H. Maître, 'Representation and Fusion of Heterogeneous Fuzzy Information in the 3D Space for Model-Based Structural Recognition - Application to 3D Brain Imaging', *Artificial Intelligence*, **148**, 141–175, (2003).

[7] R. Casati, B. Smith, and A.C. Varzi, 'Ontological Tools for Geographic Representation', in *Formal Ontology in Information Systems*, ed., N. Guarino, 77–85, IOS Press, Amsterdam, (1998).

[8] O. Colliot, O. Camara, and I. Bloch, 'Integration of Fuzzy Spatial Relations in Deformable Models - Application to Brain MRI Segmentation', *Pattern Recognition*, **39**, 1401–1414, (2006).

[9] S. Coradeschi and A. Saffiotti, 'Anchoring Symbols to Vision Data by Fuzzy Logic', in *ECSQARU'99*, eds., A. Hunter and S. Parsons, volume 1638 of *LNCS*, pp. 104–115, London, (July 1999). Springer.

[10] D. Crevier and R. Lepage, 'Knowledge-based image understanding systems: a survey', *Computer Vision and Image Understanding*, **67**(2), 160–185, (1997).

[11] O. Dameron, B. Gibaud, and X. Morandi, 'Numeric and symbolic knowledge representation of cerebral cortex anatomy: methods and preliminary results', *Surgical and Radiologic Anatomy*, **26**(3), 191–197, (2004).

[12] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V.K. Papastathis, and MG Strintzis, 'Knowledge-assisted semantic video object detection', *IEEE Transactions on Circuits and Systems for Video Technology*, **15**(10), 1210–1224, (2005).

[13] P.F. Dominey, J.D. Boucher, and T. Inui, 'Building an adaptive spoken language interface for perceptually grounded human-robot interaction', in *4th IEEE/RAS International Conference on Humanoid Robots*, volume 1, pp. 168–183, (2004).

[14] M. Donnelly, T. Bittner, and C. Rosse, 'A formal theory for spatial representation and reasoning in biomedical ontologies.', *Artificial Intelligence in Medicine*, **36**(1), 1–27, (January 2006).

[15] V. Haarslev and R. Moller, 'RACER system description', *Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2001)*, (2001).

[16] D. Han, B.J. You, Y.S.E. Kim, and I.L.H. Suh, 'A generic shape matching with anchoring of knowledge primitives of object ontology', in *ICIAR 2005*, eds., M. Kamel and A. Campilho, volume LNCS 3646, pp. 473–480, (2005).

[17] S. Harnad, 'The symbol grounding problem', *Physica*, **42**, 335–346, (1990).

[18] Joana Hois, Kerstin Schill, and John A. Bateman, 'Integrating uncertain knowledge in a domain ontology for room concept classifications', in *The Twenty-sixth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, Research and Development in Intelligent Systems, Cambridge, UK, (December 2006). Springer-Verlag.

[19] C. Hudelot, *Towards a Cognitive Vision Platform for Semantic Image Interpretation; Application to the Recognition of Biological Organisms.*, Phd in computer science (in english), Université de Nice Sophia Antipolis, April 2005.

[20] R. L. Klatzky, 'Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections', in *Spatial Cognition*, volume LNCS 1404, pp. 1–18, (1998).

[21] E. Klien and M. Lutz, 'The Role of Spatial Relations in Automating the Semantic Annotation of Geodata', in *Conference on Spatial Information Theory (COSIT 2005)*, eds., A. G. Cohn and D. M. Marks, volume LNCS 3693, pp. 133–148, (2005).

[22] B. J. Kuipers and T. S. Levitt, 'Navigation and Mapping in Large-Scale Space', *AI Magazine*, **9**(2), 25–43, (1988).

[23] F. Le Ber and A. Napoli, 'The design of an object-based system for representing and classifying spatial structures and relations', *Journal of Universal Computer Science*, **8**(8), 751–773, (2002).

[24] R. Moratz and T. Tenbrink, 'Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations', *Spatial Cognition and Computation*, **6**(1), 63–106, (2006).

[25] G. Nagypal and B. Motik, 'A fuzzy model for representing subjective and vague temporal knowledge ontologies', in *International Conference on Ontologies, Databases and Applications of Semantics*, Catania, Sicily, Italy, (2003).

[26] C. Rosse and J. L. V. Mejino, 'A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy', *Journal of Biomedical Informatics*, **36**, 478–500, (2003).

[27] S. Schulz, U. Hahn, and M. Romacker, 'Modeling anatomical spatial relations with description logics', in *Annual Symposium of the American Medical Informatics Association. Converging Information, Technology, and Health Care (AMIA 2000)*, pp. 779–783, Los Angeles, CA, (2000).

[28] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, 'Content-based image retrieval at the end of the early years', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(12), 1349–1380, (2000).

[29] C. Town, 'Ontological inference for image and video analysis', *Machine Vision and Applications*, **17**(2), 94–115, (2006).

[30] S. G. Waxman, *Correlative Neuroanatomy*, McGraw-Hill, New York, 24 edn., 2000.

Semantic Similarity of Natural Language Spatial Relations

Angela Schwering [1]

Abstract. Communication problems between humans and machines are often the reason for failures or wrong computations. While machines use well-defined languages and rules in formal models to compute information, humans prefer natural language expressions with only vaguely specified semantics. Similarity comparisons are a central construct of the human way of thinking. For instance, humans are able to act sensible in completely new situations by comparing them to similar experiences in the past. Similarity is used for reasoning on unknown information. It is necessary to overcome the differences in representing and processing information to avoid error-prone communication. A machine being able to understand natural language and detect the semantic similarity between expressions would be the key to eliminate human-machine communication problems.

This paper addresses human-machine communication about spatial configurations in natural language. We propose a computational model to capture the semantics of natural language spatial relations and to compare them with respect to their semantic similarity. The semantic description is based on an approach developed by Shariff, Egenhofer and Mark which describes natural language spatial relations via a combination of several formal spatial relations. The semantic similarity measure is inspired by Gärdenfors' conceptual spaces: we model the formal relations as quality dimensions of a geometric space, describe the natural language expressions as regions in the multidimensional space and determine their similarity via spatial distances.

1 Introduction

Any computer system working with spatial data or interacting with the environment - ambient intelligence, robots or traditional geographic information systems - must have the capability to communicate with humans about environments. Spatial human-machine communication via natural language poses great problems: natural language spatial expressions are highly ambiguous with only vaguely specified semantics and therefore cannot be easily formalized. However, a computer requires a formal model to understand and process the semantics of spatial relations. A computer being able to understand and express spatial configurations in natural language would simplify greatly the communication between humans and machines.

Shariff, Egenhofer and Mark [24, 8, 23] developed a formal model to describe natural language spatial relations: following the premise *topology matters, metric refines* they investigated several topologic and metric properties of natural language expressions in a human subject test. Based on their findings, we propose to specify the semantics of natural language spatial relations by describing all possible formal spatial relations that apply to the spatial configuration described by the natural language term. We use formal spatial relations as qualities to describe natural language spatial relations.

While equality and inequality is very easy to detect for computers, similarity is a vague and relatively undefined measure to compare entities. However, similarity plays an important part in the human cognition: the sense of similarity is the foundation for the human ability to classify similar entities, to reason on similar situations and for learning. The vagueness of similarity-based reasoning often leads to cognitively more plausible results than formal reasoning does.

Gärdenfors proposed conceptual spaces [11, 12, 13] as a cognitively plausible framework for representing information at a conceptual level. We apply the theory of conceptual spaces to a geometric similarity measure for natural language spatial relations: the conceptual space is formed by a set of quality dimensions which are grounded in well-defined formal spatial relations. Each natural language spatial relation is described by values on the quality dimensions and therefore occupies a region within the conceptual space. The geometric space is then used to determine the semantic distance between the natural language expressions.

The reminder of the paper is structured as follows: section 2 gives of an overview on the related work: we start with outlining different formal spatial relations and explain how the model by Shariff, Egenhofer and Mark use formal relations to describe natural language spatial relations. Then we shortly introduce Gärdenfors' conceptual spaces as framework for similarity measurement. In section 3 we describe our approach for determining the semantic similarity of natural language spatial relations. We explain how the quality dimensions of a conceptual space are constructed and afterwards outline the semantic distance measure for quality dimensions. Section 4 presents the results of our experiment: we compute the semantic similarity of spatial relations based on the findings in the human subject test by Mark, Egenhofer and Shariff and discuss our results. Section 5 summarizes the outcome of the paper and outlines directions for future work.

2 Related Work

To enable machines to capture semantics of natural language spatial relations, the relations must be described in a formal model. In this section we describe formal spatial relations which build the basis for the computational model by Shariff, Egenhofer and Mark [24, 8, 23] to describe natural language spatial relations. Afterward we describe Gärdenfors' theory of conceptual spaces which will be used as framework for similarity measurement.

[1] University of Osnabrueck, Germany, email: aschweri@uos.de

2.1 Formal Spatial Relations

There exist three different types of spatial relations: topologic, metric and direction relations. Topologic relations describe the position and arrangement of features in space, which are invariant under continuous transformations, such as rotation, translation and scaling. Containment, overlap and disjointness are examples for topologic relations. Egenhofer et al. developed an influential categorization of topologic relations: the 9-Intersection model describes properties of relations formally by a 3 × 3 matrix, which indicates whether the interior (A^0), exterior (A^-) and boundary (δA) of both objects intersect or not (figure 1). Eight relations were identified between two regions and nineteen between lines and regions [4, 5, 6] (figure 2).

$$I(A,B) = \begin{pmatrix} A^0 \cap B^0 & A^0 \cap \delta B & A^0 \cap B^- \\ \delta A \cap B^0 & \delta A \cap \delta B & \delta A \cap B^- \\ A^- \cap B^0 & A^- \cap \delta B & A^- \cap B^- \end{pmatrix}$$

Figure 1. Matrix characterizing the topologic relation [23, p. 257].

Metric relations focus on the distance of features. The absolute value of the quantitative distance of two features can be measured. Since they are based on the existence of a metric, they change under scaling but are preserved under translation and rotation.

Direction relations [10, 18] describe the orientation of spatial objects. Depending on the reference frame, the orientation is described by fixed directions such as cardinal directions [9] or by directions relative to some object or the intrinsic axes of an object. Direction relations are invariant under translation and scaling of the reference frame. Using relative orientation, direction relations are also invariant under rotation.

2.2 Semantics of Natural Language Spatial Relations

While formal spatial relations have well-defined semantics, natural language spatial relations have more complex semantics and often imply more than one type of formal spatial relation. People are more familiar with using spatial terms in their natural languages, but systems use definitions based on a computational model for spatial relations. To bridge this gap Shariff, Egenhofer and Mark developed a model defining the geometry of spatial natural language relations following the premise topology matters, metric refines [7]. The computational model for spatial relations [8, 24] consists of two layers: first it captures the topology between lines and regions based on the 9-Intersection model. The second layer analyzes the topologic configuration according to a set of metric properties: splitting, closeness and approximate alongness.

Topologic Properties. Shariff et al. use the 9-Intersection model to describe topologic properties of natural language spatial relations: there exist 19 different relations between lines and regions. The conceptual neighborhood graph illustrated in figure 2 arranges the different relations according to the topologic differences in the 3 × 3 matrix. The arrangement of relations also largely corresponds to the human similarity perception: the nearer two relations are within the network, the more similar they are (see [6] for further discussion of conceptual neighborhood graphs between lines and regions).

A natural language spatial relation can be described by one or several topologic relations: a human subject test showed that humans associate the expressions *crosses* and *intersects* both with the topologic relation A in figure 2 and the expression "along edge" with relation B and C.

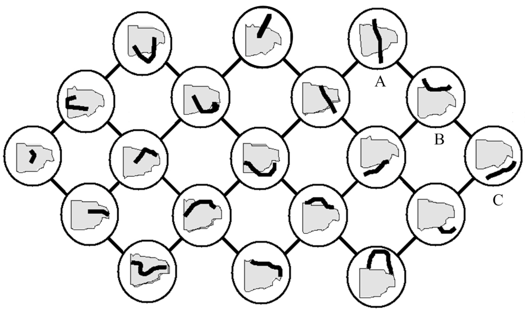


Figure 2. Conceptual neighborhood graph of topologic relations between line and region [8, p. 301].

Metric Properties. The metric properties of natural language spatial relations refer to the ratios of certain distances, length or size differences of the region and the line.

Splitting determines the way a region is divided by a line and vice versa. The intersection of the interior, exterior or boundary of a line and a region is either one- or two-dimensional. In the one-dimensional case, the length of the intersection is measured, in the two-dimensional case the size of the area. To normalize length and area, they are divided by either the region's area or the length of the line or the region's boundary.

- Interior/Exterior Area Splitting: describes how the line separates the interior / exterior of the region (figure 3 illustrates the Interior Area Splitting IAS).
- Interior/Exterior Traversal Splitting: describes the ratio of the line lying inside/outside the region to the length of the whole line (figure 3 illustrates the Exterior Traversal Splitting ETS).
- Perimeter/Line Alongness: describes the ratio of the line lying on the region's boundary to the length of the region's boundary or to the line's length (figure 3 illustrates the Line Alongness LA).
- Region Boundary Splitting: describes how the boundary of the line splits the boundary of the region.

Closeness describes the distance of a region's boundary to the disjoint parts of the line. This set of measures distinguishes between four different configurations:

- Inner/Outer Closeness: the distance between the line's and the region's boundary with the line's boundary being inside/outside the region.
- Inner/Outer Nearness: the distance between the line's interior and the region's boundary with the line being completely inside/outside the region (figure 3 illustrates the Outer Nearness ON).

Approximate alongness is a combination of the closeness measures and the splitting ratios: it assesses the length of the section where the line's interior runs parallel to the region's boundary. Four types of approximate alongness are of interest:

- Inner/Outer Approximate Perimeter Alongness assesses how the line's interior splits a buffer zone around the region's boundary with the buffer zone being inside/outside the region (figure 3 illustrates Outer Approximate Perimeter Alongness OPA).

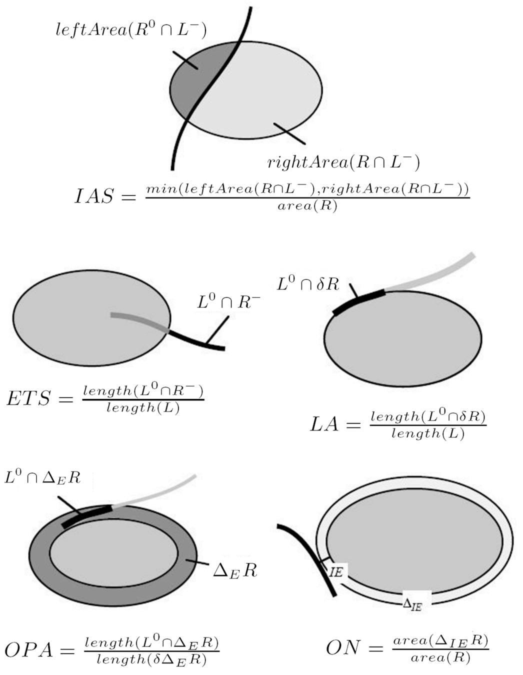

$leftArea(R^0 \cap L^-)$
$rightArea(R \cap L^-)$
$IAS = \frac{min(leftArea(R \cap L^-), rightArea(R \cap L^-))}{area(R)}$
$L^0 \cap R^-$
$L^0 \cap \delta R$
$ETS = \frac{length(L^0 \cap R^-)}{length(L)}$
$LA = \frac{length(L^0 \cap \delta R)}{length(L)}$
$L^0 \cap \Delta_E R$
$\Delta_E R$
IE
Δ_{IE}
$OPA = \frac{length(L^0 \cap \Delta_E R)}{length(\delta \Delta_E R)}$
$ON = \frac{area(\Delta_{IE} R)}{area(R)}$

Figure 3. Illustration of several metric measures: Interior Area Splitting (IAS), Exterior Traversal Splitting (ETS), Line Alongness (LA), Outer Approximate Perimeter Alongness (OPA) and Outer Nearness (ON) [24, pp. 212–214].

- Inner/Outer Approximate Line Alongness assesses the length of the line's interior falling within a buffer zone around the region's boundary with the buffer zone being inside/outside the region.

We are aware of the fact that this computational model does not include a "complete" semantic description of natural language spatial relations. It does not consider directional properties, neither other aspects such as functional properties of spatial relations [2]. However, the combination of topologic and metric properties already serves as a good approximation of the semantics inherent in natural language spatial relations and has proven very useful for the similarity comparison.

2.3 Gärdenfors' Conceptual Spaces

The notion of a conceptual space was introduced by Peter Gärdenfors as a framework for representing information at the conceptual level [11, 12, 13]. Conceptual spaces can be utilized for knowledge representation and support semantic similarity measurement. A conceptual space is formalized as a multidimensional space consisting of a set of quality dimensions [2]. The quality dimensions can have any geometric or topologic structure: in figure 4 are shown two linear dimensions. Each concept is described on the quality dimensions with either a single value or an interval. Concepts are therefore represented via n-dimensional convex regions within the conceptual space.

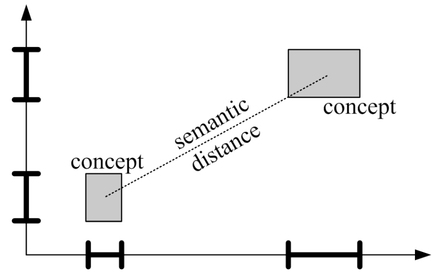


Figure 4. Two concepts and their semantic distance in a two-dimensional conceptual space.

The fact that conceptual spaces have a metric allows for the calculation of spatial distances between concepts in the space. The spatial distance is interpreted as the semantic distance: the nearer two concepts are in the conceptual space, the lower is their semantic distance and the higher is their similarity. The similarity is considered as a decaying function of the semantic distance [1, 11]. Gärdenfors proposes to use the Minkowski metrics to calculate the semantic distance.

$$distance_{ij} = [\sum_{k=1}^{n} |x_{ik} - x_{jk}|^r]^{\frac{1}{r}}$$

The Minkowski metrics are described via a generic formula: $r = 1$ results in the city-block distance and $r = 2$ in the Euclidean distance. According to the city-block metric, the distance equals the sum of the absolute distances of each dimension and the Euclidean distance is computed as the square root of the sum of the dimension-wise squared differences [25]. Johannesson and Gärdenfors demonstrated in experimental studies, that the Euclidian metric is more appropriate when stimuli are composed of integral, perceptually fused dimensions and the city-block metric is appropriate when the stimuli are composed of separable dimensions [14, 15].

3 A Semantic Similarity Measure for Natural Language Spatial Relations

Our semantic similarity measure for natural language spatial relations combines the above mentioned approaches: Shariff, Egenhofer and Mark showed how natural language spatial relations can be described by their topologic and metric properties. We use their model to determine relevant qualities for the semantic description of spatial relations. For the similarity comparison we apply Gärdenfors' theory of conceptual spaces as framework to represent the topologic and metric properties and determine the similarity.

Via this formalization of natural language spatial relations a machine cannot only "understand" the meaning of the spatial expressions, but also compare two natural language expressions with respect to their similarity. A semantic similarity measure therefore enables a machine to compare instructions from humans: when it receives a new instruction it can search for similar instructions in the past and reuse its "experiences". Similarity comparisons are important to simulate human reasoning processes and offer the vagueness and flexibility that formal-logic reasoning is often lacking. The ability of judging similarity is necessary to react adequately to new situations by comparing them to experiences learned in the past. We consider similarity comparisons as an important form of non-classical reasoning.

[2] Gärdenfors' conceptual spaces consist of a more complex structure based on domains represented through a set of integral dimensions, which are distinguishable from all other dimensions. Since this paper focuses on conceptual spaces as technique for knowledge representation and similarity measurement, we do not focus on the cognitive foundation of conceptual spaces, but concentrate only on the methodology to formalize conceptual spaces. We therefore describe a "simplified" version of conceptual spaces consisting of one-dimensional scientific dimensions. The complete cognitive foundation of conceptual spaces can be found in [11].

3.1 Topologic Properties of Natural Language Spatial Relations

Shariff et al. proposed 16 different properties to specify the semantics of natural language spatial relations. The first property describes the topologic relation between the region and the line.

Representing the semantics. We use the conceptual neighborhood graph between lines and regions as shown in figure 2 as one quality dimension of the conceptual space. The topologic properties of natural language spatial expressions are described by different values on this dimension. While most qualities are described on a linear dimension, this quality dimension has a network structure. This is necessary to reflect the complex similarity relations between the different topologic configurations. As mentioned above, there exist different ways to construct the conceptual neighborhood graph between lines and regions. In [6], Egenhofer and Mark propose the snapshot model, which focusses on the differences in the topologic relations, and the smooth-transition model, which explicitly represents the change process between the configurations. The selection of the conceptual neighborhood graph has an influence on the similarity value computed by our model, because the arrangement and therefore also the measured distances differ.

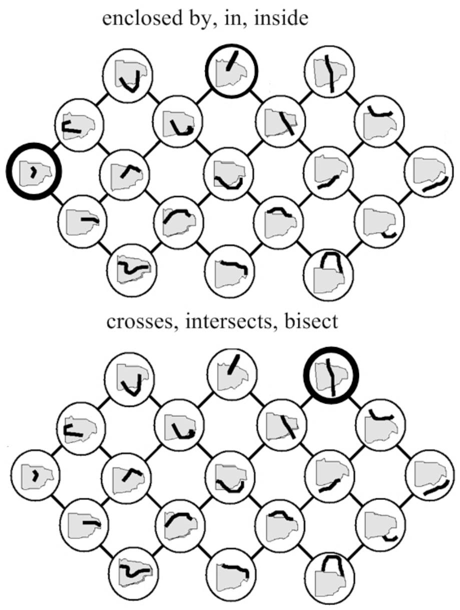


Figure 5. Examples for the topologic description of spatial relations.

Measuring the similarity. The semantic distance between two different properties on a network dimension are the shortest distance between the nodes in the network. Figure 5 shows the relations $enclosed\,by$, in and $inside$ described by a line lying completely inside the region and the relations $crosses$, $bisects$ and $intersects$ described by a line starting outside of the region, crossing the region and ending outside of the region again. The semantic distance between these relations equals six, because the nodes are six steps away in the network.

$$distance(x_1 \wedge ... \wedge x_k, y_1 \wedge ... \wedge y_m) = \frac{1}{km} \sum_{i=1}^{k} \sum_{j=1}^{m} distance(x_i, y_j)$$

If a concept is represented by a conjunction of several nodes (i.e. several topologic configurations are possible for one expression) we compute the average distance as shown in the above formula. This measure was proposed by Rada [19] as semantic distances in a semantic network.

3.2 Metric Properties of Natural Language Spatial Relations

Shariff et al. propose further 15 metric properties to describe natural language spatial relations.

Representing the semantics. The 15 different metric properties are represented on separate dimensions: seven for splitting, four for closeness and four for alongness. These dimensions are linear: each natural language spatial relation is described by an interval on those dimensions which are applicable. Table 1 shows the specification for the same six relations as above. For instance, the dimension Outer Closeness (OC) is not determinable for the relations $enclosed\,by$, in and $inside$, because the line is completely inside the region.

	IAS	IC	OC	IN
crosses	0.03- 0.50		0.29- 8.62	
enclosed by		0.19- 0.79		0.19- 0.79
in		0.03- 0.78		0.03- 0.69
inside		0.02- 0.87		0.02- 0.77
intersect	0,02- 0,50		0,25- 9,10	
bisect	0,05- 0,48	0,40- 8,78		

Table 1. Examples for the metric description of spatial relations.

Measuring the similarity. The semantic distance on linear dimensions is measured as the spatial distances. If the spatial relations are described via intervals, the semantic distance is computed between the mean values of the intervals. In the multidimensional space, this corresponds to the semantic distance between the center points of the regions describing the spatial relations. The center point of a concept is often interpreted as the prototypical instance of this concept.

3.3 Adaptation of the Theory of Conceptual Spaces

Gärdenfors describes conceptual spaces only at a theoretical level. To develop a computational model of conceptual spaces we need to extend and adapt slightly the theory to enable calculations with real world data.

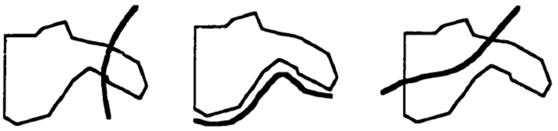
To use conceptual spaces for our similarity comparison of spatial relations, we have to consider the fact, that the topologic properties are decisive and the metric properties serve only as refinement. The calculation accounts for the importance of dimensions by assigning weighting factors to each quality dimension reflecting the importance of the dimension.

In order to calculate the semantic distances between concepts it is required that all quality dimensions of the space are represented in the same relative unit of measurement. This is ensured by calculating the z scores for these values, also called z-transformation [3]. We compute the spatial distance based on the standardized dimension values.

Moreover, we have to consider that not all spatial expressions are described by the same set of quality dimensions: some quality dimensions may not be applicable at all or their values do not matter for a specific configurations. Therefore the similarity comparison must be able to compare descriptions of spatial relations based on different conceptual spaces. In [22, 20, 21] we proposed a two-step process for computing semantic distances in conceptual spaces based on different dimensions: in the first step we check whether both concepts are described by the same dimension. For this purpose we introduce a boolean dimension with the value 'yes' for dimension existence and 'no' indicating that the dimension is not applicable. If the dimension is applicable, we compare the values for this quality. This way we can compare concepts described in non-identical conceptual spaces. The weighting factors are chosen in a way, that a non-applicable dimension increases the semantic distance more than different values on the same quality dimension.

4 Evaluation

To evaluate the quality of our similarity comparison we test our model with a set of spatial relations.



The road crosses the park.

The road runs along the edge of the park

The road splits the park.

Figure 6. Examples from the human subject test [17, 23].

Mark and Egenhofer [17] investigated the underlying topologic relations of different natural language spatial relation in a human subject test: the subjects were presented with a picture of a park and a sentence describing a particular spatial relation between a park and a road. The subjects had to draw a road in the picture such that the resulting drawing would fit the spatial relation. Figure 6 shows several examples of these drawings.

Shariff et al. analyzed the drawings from the experiment for the metric properties of the investigated spatial relations and constructed a complete topologic and metric description of 59 natural language spatial relations. Figure 5 and table 1 show only a small set of relations as examples; the complete data can be found in [24] and [23].

We used Shariff et al.'s specification of natural language spatial relations as data basis for our similarity comparison. The similarity comparison was done for all 59 spatial relations tested by Shariff and showed good results. Here however, we can present the results only for the small subset of 15 spatial relations [3]. Table 2 shows the resulting similarity values. To increase the readability we present the results in a condensed way: we assigned the spatial relations to three groups: *similar* relations have a semantic distance less than 1, *partially similar* relations have a semantic distance between 1-4.5 and *not similar* relations have a semantic distance greater 4.5.

	along edge	avoid	bisect	bypass	cross	enclosed by	enter	intersect	in	inside	near	within	splits	transsect	traverse
along edge	s	p	n	p	n	n	n	n	n	n	p	n	n	n	n
avoid	p	s	n	s	n	n	n	n	n	n	s	n	n	n	n
bisect	n	n	s	n	s	n	p	s	n	n	n	n	s	s	s
bypass	p	s	n	s	n	n	n	n	n	n	s	n	n	n	n
cross	n	n	s	n	s	n	p	s	n	n	n	n	s	s	s
enclosed by	n	n	n	n	n	s	n	n	s	s	n	s	n	n	n
enter	n	n	n	n	p	n	s	p	n	n	n	n	p	p	n
intersect	n	n	s	n	s	n	p	s	n	n	n	n	s	s	s
in	n	n	n	n	n	s	n	n	s	s	n	s	n	n	n
inside	n	n	n	n	n	s	n	n	s	s	n	s	n	n	n
near	p	s	n	s	n	n	n	n	n	n	s	n	n	n	n
within	n	n	n	n	n	s	n	n	s	s	n	s	n	n	n
split	n	n	s	n	s	n	p	s	n	n	n	n	s	s	s
transect	n	n	s	n	s	n	p	s	n	n	n	n	s	s	s
traverse	n	n	s	n	s	n	p	s	n	n	n	n	s	s	s

Table 2. Results of the similarity comparison (s means similar, p means partially similar and n means not similar.

Obviously every relation is the most similar to itself, because it has a semantic distance of 0 to itself. We will discuss the results now in detail for the relations *avoid*, *cross* and *in*. We chose three very different relations to show that our similarity measure can handle different spatial configurations. The reader may have a look at table 2 for the complete results.

The most similar relations to *avoid* are the relations *bypass* and *near*. Both relations describe a line (the road in the human subject test by Mark et al.) which is entirely outside of the region (the park respectively). The relation *along edge* is only classified as partially similar: in the human subject test the subjects described *along edge* as a either touching or non-touching line along the region. Therefore our computational model determines a higher semantic distance to *avoid* than to *bypass* and *near*. All other relations are classified as non-similar: they describe relations where the line is at least partly inside the region (the subjects described *enclosed by* topologically as a line being in the region).

The most similar relations to *cross* are the relations *traverse, intersect*, *split*, *transsect* and *bisect*. All these relations describe a line starting from outside of the region, going through the region and ending again outside of the region. The relation *enter* is classified as partially similar: it describes a line starting outside the region, going into it and ending inside of the region. All other relations indicate spatial configurations where the line is either entirely inside or outside the region. Our computational model classifies these relations as not similar.

The most similar relations to *in* are the relations *enclosed by*, *within* and *inside*. Although *enclosed by* and *in* mean something different, they both indicate almost the same spatial configuration. All other relations have not been classified as similar, because they all indicate a spatial configuration where at least some part of the line is outside of the region.

The test shows that our computational model for semantic similarity measurement between spatial relations comes up with sensible results. Of course, this approach is limited to the comparison of the spatial configuration only and does not account for other semantic properties of spatial relations.

[3] The complete data-set can be found online http://www.cogsci.uni-osnabrueck.de/ aschweri/SimilarityComparison.xls

5 Summary and Future Work

In this paper we developed a semantic similarity measure for natural language spatial relations and evaluated the measure based on a set of English spatial relations. The aim of this computational model is to easy human-computer communication about spatial configuration: using our computational model, a computer becomes able to analyze the semantics of natural language terms and compare them with respect to their semantic similarity. A computer can "understand" spatial expressions in natural language and reason on them.

We combined a framework for semantic similarity measurement, namely Gärdenfors conceptual spaces, with Shariff et al.'s model to describe natural language spatial relations. Shariff et al.'s description is based on topologic properties and various measures for metric properties. Each property is represented on a different quality dimension: the topologic properties are modeled on a network dimension while the metric properties are modeled on a linear dimension.

In our experiment we computed the similarity of various spatial relations for which have been produced formal descriptions before in a human subject test by Mark et al. Although the results are already very convincing, we will have to examine in the future, how well the computed similarity values match human similarity judgement.

Moreover, the formal model of Shariff, Egenhofer and Mark is limited to lines and regions. More work has to be done to test to what extend this approach can be applied to region-region and line-line spatial relations as well. Furthermore, the formal model by Shariff et al. is based only on a single park-road domain. It should be examined whether there are any scale or domain dependencies (see for example [16]).

REFERENCES

[1] F. Attneave, 'Dimensions of similarity', *American Journal of Psychology*, **63**, 516–556, (1950).

[2] K. R. Coventry and S. C. Garrod, *Saying, seeing, and acting: The psychological semantics of spatial prepositions*, Essays in Cognitive Psychology, Psychology Press, Hove, UK, 2004.

[3] J. Devore and R. Peck, *Statistics - The exploration and analysis of data*, Duxbury, Pacific Grove, CA, 4th edn., 2001.

[4] M. Egenhofer and R. Franzosa, 'Point-set topological spatial relations', *International Journal of Geographical Information Systems*, **5**(2), 161–174, (1991).

[5] M. J. Egenhofer and K. K. Al-Taha, 'Reasoning about gradual changes of topological relationships', in *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, eds., A. U. Frank, I. Campari, and U. Formentini, volume 1329 of *Lecture Notes in Computer Science*, 196–219, Springer, Berlin, Germany, (1992).

[6] M. J. Egenhofer and D. M. Mark, 'Modeling conceptual neighborhoods of topological line-region relations', *International Journal of Geographical Information Systems*, **9**(5), 555–565, (1995).

[7] M. J. Egenhofer and D. M. Mark, 'Naive geography', in *Proceedings of the International Conference on Spatial Information Theory (COSIT95)*, eds., A. U. Frank and W. Kuhn, volume 988 of *Lecture Notes in Computer Science*, pp. 1–15, Berlin, Germany, (1995). Springer.

[8] M. J. Egenhofer and A. R. Shariff, 'Metric details for natural-language spatial relations', *ACM Transactions on Information Systems*, **16**(4), 295–321, (1998).

[9] A. Frank, 'Qualitative spatial reasoning: cardinal directions as an example', *International Journal of Geographical Information Systems*, **10**(3), 269–290, (1996).

[10] C. Freksa, 'Using orientation information for qualitative spatial reasoning', in *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, eds., A. U. Frank, I. Campari, and U. Formentini, volume 639 of *Lecture Notes in Computer Science*, pp. 162–178, Pisa, Italy, (1992). Springer.

[11] P. Gärdenfors, *Conceptual spaces: The geometry of thought*, MIT Press, Cambridge, MA, 2000.

[12] P. Gärdenfors, 'Conceptual spaces as a framework for knowledge representation', *Mind and Matter*, **2**(2), 9–27, (2004).

[13] P. Gärdenfors, 'How to make the semantic web more semantic', in *Formal Ontology in Information Systems, Proceedings of the Third International Conference (FOIS 2004)*, eds., A. Varzi and L. Vieu, volume 114 of *Frontiers in Artificial Intelligence and Applications*, 153–164, IOS Press, Amsterdam, NL, (2004).

[14] M. Johannesson, 'Combining integral and separable subspaces', in *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, eds., J. D. Moore and K. Stenning, pp. 447–452. Lawrence Erlbaum, (2001).

[15] M. Johannesson, *Geometric models of similarity*, Ph.D. dissertation, PhD Thesis. Lund University, 2002.

[16] A.-K. Lautenschütz, C. Davies, M. Raubal, A. Schwering, and E. Pederson, 'The influence of scale, context and spatial preposition in linguistic topology', in *Proceedings of the International Conference on Spatial Cognition*, Lecture Notes in Artificial Intelligence, Bremen, (2006). Springer.

[17] D. Mark and M. J. Egenhofer, 'Topology of prototypical spatial relations between lines and regions in english and spanish', in *Proceedings of the Twelfth International Symposium on Computer- Assisted Cartography*, volume 4, pp. 245–254, Charlotte, North Carolina, (1995).

[18] R. Moratz and T. Tenbrink, 'Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations', *Spatial Cognition and Computation*, **6**(1), 63–106, (2006).

[19] R. Rada, H. Mili, E. Bicknell, and M. Blettner, 'Development and application of a metric on semantic nets', *IEEE Transactions on systems, man, and cybernetics*, **19**(1), 17–30, (1989).

[20] A. Schwering, 'Hybrid model for semantic similarity measurement', in *Proceedings of the 4th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE05)*, eds., R. Meersman and Z. Tari, volume 3761 of *Lecture Notes in Computer Science*, pp. 1449–1465, Agia Napa, Cyprus, (2005). Springer.

[21] A. Schwering, *Semantic Similarity Measurement including Spatial Relations for Semantic Information Retrieval of Geo-Spatial Data*, Phd thesis, University of Münster, 2006.

[22] A. Schwering and M. Raubal, 'Spatial relations for semantic similarity measurement', in *2nd International Workshop on Conceptual Modeling for Geographic Information Systems (CoMoGIS2005)*, eds., J. Akoka, S. W. Liddle, I.-Y. Song, M. Bertolotto, I. Comyn-Wattiau, W.-J. vanden Heuvel, M. Kolp, J. Trujillo, C. Kop, and H. C. Mayr, volume 3770 of *Lecture Notes of Computer Science*, pp. 259–269, Klagenfurt, Austria., (2005). Springer.

[23] A. R. Shariff, *Natural-language spatial relations: Metric refinements of topological properties*, Phd thesis, University of Maine, 1996.

[24] A. R. Shariff, M. J. Egenhofer, and D. M. Mark, 'Natural-language spatial relations between linear and areal objects: The topology and metric of english language terms', *International Journal of Geographical Information Science*, **12**(3), 215–246, (1998).

[25] P. Suppes, D. M. Krantz, R. D. Luce, and A. Tversky, *Foundations of measurement - geometrical, threshold, and probabilistic representations*, volume 2, Academic Press, Inc, San Diego, California, USA, 1989.