AISB Journal

The Interdisciplinary Journal of
Artificial Intelligence and the Simulation of Behaviour

Volume 1 – Number 5 – June 2004

The Journal of the
Society for the Study of Artificial Intelligence
and the Simulation of Behaviour
http://www.aisb.org.uk

Published by
The Society for the Study of Artificial Intelligence and Simulation of Behaviour

http://www.aisb.org.uk/

ISSN 1476-3036

Contents

AISB Journal 1(5), © SSAISB, 2004
http://www.aisb.org.uk

Biologically-inspired Machine Vision, Theory and Application

Horst Holstein and Fred Labrosse

Department of Computer Science, University of Wales, Aberystwyth
Penglais, Aberystwyth, Ceredigion SY23 3DB, Wales, UK
hoh@aber.ac.uk ; ffl@aber.ac.uk

Editors' Introduction

Computer vision has developed into a mature science over the last forty years, but current computer vision systems are vastly different from, and in most cases lack the efficiency of, biological vision systems. Biological vision therefore remains a strong metaphor for the design of machines that simulate intelligent behaviour in visually sensed environments. It hints at information interchange between many cooperating parallel subsystems. This theme was strongly present in the keynote talk, "Action Representations", by Aloimonos and Fermüller. Immensely rewarding applications in human-machine interaction await advances in the multi-disciplinary threads of machine vision, perception and cognition.

The aim of the symposium was to promote a forum for a strong multi-disciplinary interaction in the state-of-the-art and opening directions of research and technology related to biologically-inspired machine vision and visualisation. The papers included for publication in the AISB Journal conveyed this aim in the oral and written presentations. Thus, Kolesnik and Barlit simulate image contrast detection by simple cells in the primary visual cortex through iterative orientation tuning, and exhibit noise suppression through the introduction of cross-oriented inhibition. Krüger, Lappe and Wörgötter propose a new multi-modal image representation that is also motivated by image processing in the human visual system. Li and Holstein demonstrate recognition of periodic human motion captured in sparse feature point data, without recourse to structure from motion

It is hoped that the selected papers will lead interested readers to access the full range of contributions presented in the Symposium Proceedings.

AISB Journal **1**(5), © SSAISB, 2004
http://www.aisb.org.uk

Perception of Human Periodic Motion in Moving Light Displays: a Motion-based Frequency Domain Approach

Baihua Li and Horst Holstein

Department of Computer Science, University of Wales, Aberystwyth
Ceredigion, Aberystwyth, UK, SY23 3DB
bbl00@aber.ac.uk ; hoh@aber.ac.uk

Abstract

We present a *motion-based frequency-domain* technique for modelling and recognising human periodic movements in moving light displays (MLDs). Periodic motions are modelled by motion templates composed of a set of feature power vectors. Feature power vectors are extracted from unidentified trajectories of feature points using motion power spectral analysis. Motion recognition therefore is carried out in the frequency domain by finding a best match among pre-stored templates. Recognition is demonstrated through examples of human periodic motion with data obtained from a marker-based motion capture system. Experimental results prove that motion characteristics exist not only in the spatio-temporal domain, but also in frequency domain. This method contrasts with common spatio-temporal approaches and avoids a time consuming recovery of underlying kinematic structures in visual analysis.

1 Introduction

The ability of humans to perceive structure and motion from sparse point feature motion cues has been demonstrated by Johansson's Moving light displays (MLDs) (Johansson, 1975). In the MLDs as shown in Figure 1, an image sequence was reduced to a set of moving light dots. These light dots were attached at the joint sites of a human subject, contrasted to a dark background. The dots acted as discrete feature-points presenting motion characteristics in spatio-temporal domain. The MLDs carried only motion information but no structural information, since the displayed points were discrete and unconnected. One frame of static dots remained meaningless to observers, while human observers were able to recognise activities such as walking, running or stair climbing from a sequence of relative movements of a small number of lights. Barclay et al. (1978), Cutting and Kozlowski (1977) also showed human observers can identify the actor's gender and even their friends by their gaits in MLDs. These pioneering works in psychology relating to human motion perception suggest that feature-based motion presentation plays an important role in recognition tasks. In the case of machine vision, the biological metaphor suggests that it is possible to use the reduced spatio-temporal information, such as embedded in MLDs, for recognition.

MLDs have been widely used in studies of perception in psychology (Johansson, 1975; Barclay et al., 1978; Cutting and Kozlowski, 1977); motion tracking, activity classification and recognition of human motions in computer vision (Goddard, 1992; Cédras

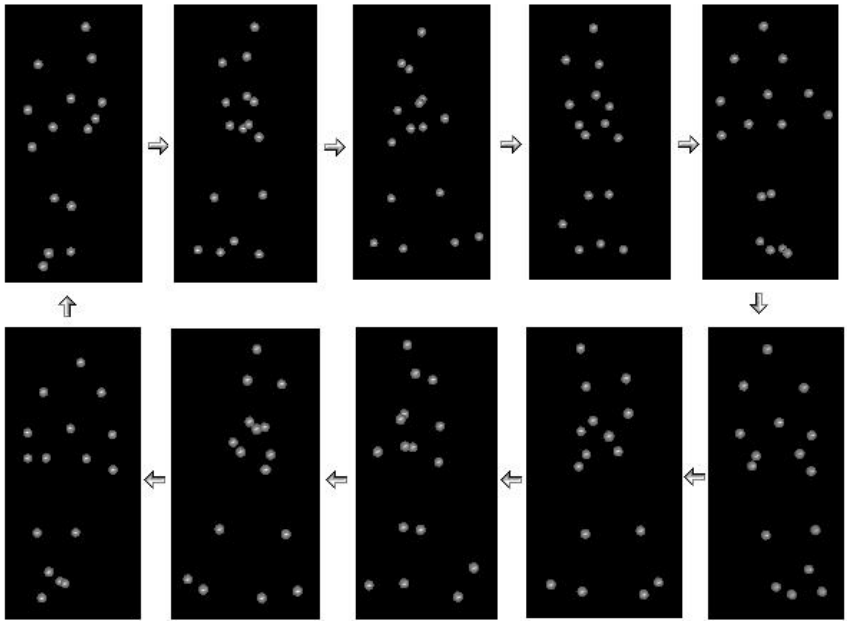


Figure 1: A clockwise circle-walking person in MLDs with 16 feature points.

and Shah, 1994; Campbell and Bobick, 1995; Cédras and Shah, 1995; Boyd and Little, 1997; Dorfmüller, 1999); clinical gait analysis and sports science (Ferrigno and Gussoni, 1988; Söderkvist and Wedin, 1993; Angeloni et al., 1994; Michael, 1996; Stoddart et al., 1999); computer game and animation industries (Gleicher, 1999; Richards, 1999); augmented reality and virtual reality (Dorfmüller, 1999); and more recently in humanoid robot design (Hill and Pollick, 2000). Motion analysis from MLDs uses concise and accurate data to investigate what are the essential recognition features for modelling motion, formulating kinematics and motion synthesis. The research has gained increasing attention in ever widening applications.

During the last two decades, MLD perception theory has emphasised *structure-based* recognition in computer vision (Cédras and Shah, 1994; Aggarwal and Cai, 1999; Gavrila, 1999; Moeslund and Granum, 2001). Researchers used various kinds of information from images to recover the time varying articulated structure of the human body. Subsequently, from the recovered underlying structure, parameters such as joint angles could be identified for motion interpretation and recognition. For example, Campbell and Bobick (1995) classified ballet dance steps using a phase space representation. The phase space was related to each degree of freedom (DOF) in the identified articulated human structure, using MLD data. Goddard (1992) proposed a computational model for visual motion recognition of gaits in walking, jogging, and running from MLDs. He used the joint angles and angular velocities as features for recognition. A difficulty in this work was the prior necessity to identify individual points in the images. He has argued the possibility for a perception directly from motion information.

Very few researchers have attempted motion recognition directly from motion information provided by MLDs (Cédras and Shah, 1995). However, recent work by Boyd and Little (1997), using global shape-of-motion features derived from MLD images, has

shown that it is possible to recognise individual people by their gait using non-structural means. This approach avoids the complex vision problem of kinematic structure recovery.

In this paper we propose a frequency domain approach for recognition of human periodic movements in MLDs. We believe that common periodic movements, such as walking, running and skipping, present human body segment movement in relative harmony. These motion characteristics exist not only in the spatio-temporal domain, but also in frequency domain. They are preserved even in the reduced information of MLDs, allowing motion recognition without knowledge of underlying structure.

The rest of the paper is organized as follows: Section 2 reviews related work on cyclic motion recognition. Section 3 states our method of data collection. Section 4 describes the frequency domain approach for recognition. Section 5 provides experimental results on recognition of human cyclic motion. We discuss and conclude our work in Section 6 and 7.

2 Related work

Human motion, specifically walking, has been studied extensively using various spatio-temporal cues from images or MLDs. Approaches directly using motion for periodic motion recognition are described in e.g. (Polana and Nelson, 1993; Tsai et al., 1994; Fujiyoshi and Lipton, 1998; Boyd and Little, 1997). Polana and Nelson (1993) propose a method for detecting periodic motion using Fourier transforms on several point trajectories. They indicated that in principle, the period of the movement could be inferred from averaging the fundamental frequencies of the point trajectories. Tsai et al. (1994) use the trajectory of one point on an object that performed some cyclic motion to compute curvature. An autocorrelation is performed to enhance self-similarity within the curvature function. The Fourier transform is finally used to detect the presence of a cycle and its period from the spatio-temporal curvature. Fujiyoshi and Lipton (1998) generate a “star” skeleton from the object boundary. They apply Fourier analysis to the skeleton for detecting periodic motion. Then they utilise both posture and motion cycle of the “star” skeleton to recognise activities such as walking and running.

In these approaches, Fourier transforms are used to detect or recognise periodicity. The detected periodicity is used to assist motion recognition. For spatio-temporal domain approaches, in order to deal with the problems of e.g. human motion irregularities or change in speed, techniques such as scale space or Dynamic Time Warping (DTW), considered computationally expensive, are usually used to match portions of scale space to find repeated patterns from identified curves.

The direct application of frequency domain analysis for motion recognition has received much less attention. Works emphasising frequency domain analysis are, for example, Angeloni et al. (1994), Köhle and Merkl (1996). Angeloni et al. use gait kinematic data from MLDs to analyse the frequency content of whole body movement. Their work presents the characteristic spectral distribution among articulated body parts. Köhle and Merkl demonstrate that the kinetic data from ground reaction force platforms can also be used to classify gait patterns in clinical gait analysis, through Fourier transforms of vertical force components and classification by self-organising maps. The works of both Angeloni and Köhle show that motion frequency spectra may include cues suitable for motion recognition.

We are pursuing the development of interpretation and recognition of human periodic motion. We investigate a pure frequency domain approach to model human periodic motion using kinematic data from MLDs.

3 Data collection from 3D-MLDs

All human kinematic data used in our investigation are acquired from a marker-based 3D optical motion capture (MoCap) system, the Vicon 512. The system provides 3D coordinates of unidentified trajectories of markers attached to a subject, in the manner of a 3D-MLD system. The data are not affected by the projective distortions of particular camera views. In this respect, we differ from other classical MLD investigations, which detect data from 2D projected image sequences. The high quality of our data allows us to apply a frequency domain approach for direct motion recognition from the captured data.

In our motion capture system, the world coordinate system has its origin on the ground. The xy-plane is parallel to the ground plane, and the Z-axis is vertical. Other conditions for data collection in our experiments are:

- Motions are captured in a control volume, about 4m (length)$\times$ 4m (width)$\times$2.5m (height). The measurement accuracy is to the level of a millimetre.
- Sixteen markers, regarded as feature points, are attached on human subjects at the following locations: head, back (anatomical T10), shoulders (L/R), elbows (L/R), wrists (L/R), AISs (L/R hips), knees (L/R), ankles (L/R), toes (L/R hallux big toes). They are effective in indicating motion cues in MLDs.
- The trajectories are nearly always uninterrupted. Some small trajectory gaps arising from occlusion are filled by interpolation during MoCap post-processing.
- The correspondence between a 3D trajectory and the marker identity is not assumed to be known in the motion to be identified. In Figure 4 and Figure 5, we have indicated feature point identity for display clarity, but identity information is not used in the recognition process.

4 A Frequency Domain Method

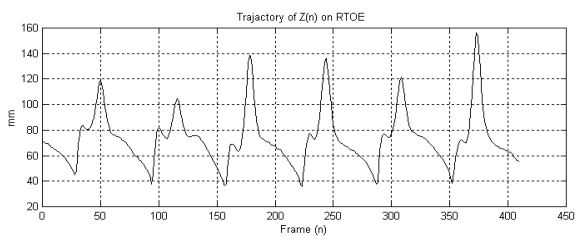


Figure 2: Right toe z-trajectory of a walking person.

The relative movements of feature points contain information both of motion and structure identity. For most common periodic activities carried out on a level (horizontal) floor, the vertical components, which are the z-coordinates of 3D-MLD trajectories, imply crucial cues relative to ground ($z = 0$) and provide a simple input for Fourier analysis. They can be used without transformation, because they contain no secular variation and are motion orientation invariant. In this study, we use the z-data as the only motion cue to be analysed. We find that cues from the unidentified z-trajectories suffice to discriminate between a number of simple periodic human activities. An example z-trajectory of right toe (hallux) of a walking person is shown in Figure 2.

4.1 Power spectral analysis for whole body movement

Our experiments assume availability only of unidentified trajectories of feature points, $i = 1 \ldots I$, obtained from 3D-MLDs. We apply spectral analysis to the vertical component $z_{i(n)}$ of the trajectory of each feature point i of frame samples $n = 0 \ldots N-1$, N being the trial length. The Fourier decomposition of the z-trajectory is expressed by

$$z_{i(n)} = \frac{1}{2} a_{i(0)} + \frac{1}{N} \sum_{k=1}^{N-1} a_{i(k)} \cos(2\pi nk/N) + b_{i(k)} \sin(2\pi nk/N), \tag{1}$$

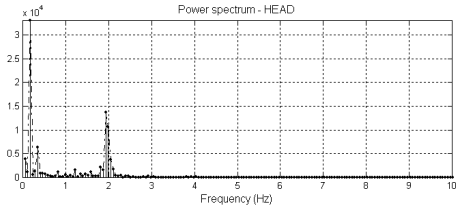
where $a_{i(k)}$ and $b_{i(k)}$ are the Fourier coefficients of feature point i. To achieve an adequate frequency resolution, the length of each trial is $N = 256$ to 1300 frames, ideally including at least 5 gait cycles (Gc) for a specific periodic movement. When the trial length N is an exact power of two, we adopt the Fast Fourier Transform (FFT) algorithm which is more efficient than the raw Discrete Fourier Transform (DFT).
The power spectrum for the kth frequency multiple of feature point i is defined by the Fourier coefficients $a_{i(k)}$ and $b_{i(k)}$ as

$$P_{i(k)} = a_{i(k)}^2 + b_{i(k)}^2, \;\; k = 1, 2 \ldots N/2\,. \tag{2}$$

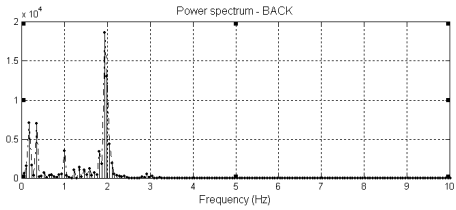
Examples of such power spectra for a clockwise circle-walking person are given in Figure 3.

From power spectral analysis of whole-body feature points for a number of common cyclic movements, such as walking, running, jumping, skipping, we find the dominant power of human movements occupies only a narrow bandwidth, with the upper limit about 10 Hz. The power spectral distribution shows clustering around some frequencies related to the fundamental activity frequency, such as the gait-cycle frequency in walking and running, and its multiples. The magnitude and envelope of a specific spectrum retain time-shift invariance regardless of where in time the Fourier transform is performed. We also observe that spectral patterns reflect the activeness of body parts, consistent with undergoing different intensity and character of motion. For example, as shown in Figure 3, each spectrum of head, shoulder, back and hip has a small total power and a very different distribution compared to each spectrum of elbow, wrist, knee, ankle and toe. Low frequency components well under 1Gc in spectra P_i reflect vertical motion noise associated with secular postural changes and human motion irregularity over the trial track. These low frequency noise components are relatively more evident for body parts undergoing small vertical movements, such as for head, back and shoulder during walking. The spectra of active body parts, such as elbow, wrist, knee, ankle and toe, show remarkable motion power clustering around the Gc and its multiples, and therefore show a relatively diminished motion noise at low frequency. We can also observe that the power components of the left toe (Figure 3(j)) are larger than that of the right toe (Figure 3(i)), though their spectral patterns are similar. The power discrepancy arises because clockwise circle-walking has a larger left foot movement than that of the right foot. For the same kind of motion in different subjects, spectral patterns of the same feature points are similar, hinting at the motion nature, differences being related to variation in individual speed and amplitude.

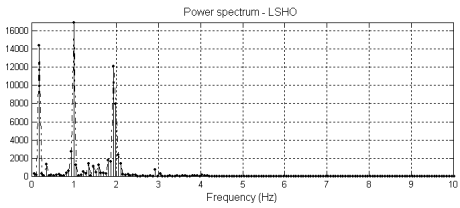
From the power spectral analysis of human periodic movements, we find human body parts present rhythmicity and harmony related to a fundamental activity cycle. To achieve speed-invariant representation for the same kind of movement, we normalise whole-body spectra to the fundamental activity cycle or generalised Gait-cycle (Gc). To obtain an accurate Gc, we sum the power spectra over all feature points i for frequency $k^* \Delta \mathrm{f}$ within



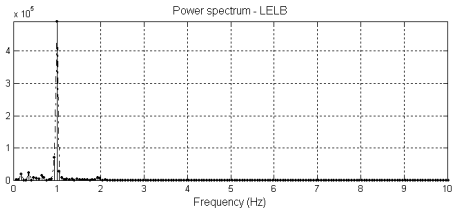
(a) head



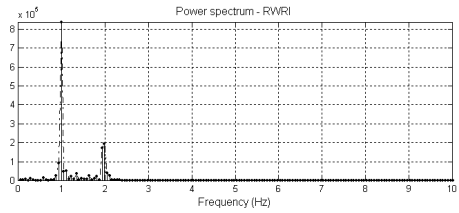
(b) back (anatomical T10)



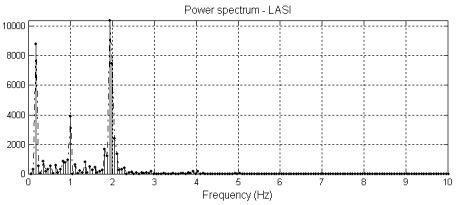
(c) shoulder



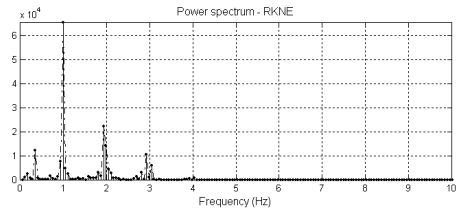
(d) elbow



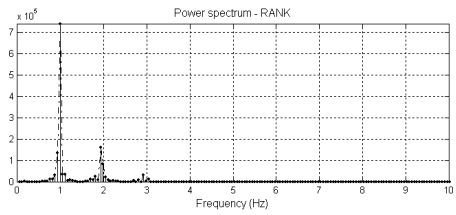
(e) wrist



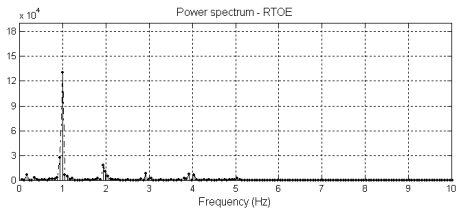
(f) hip



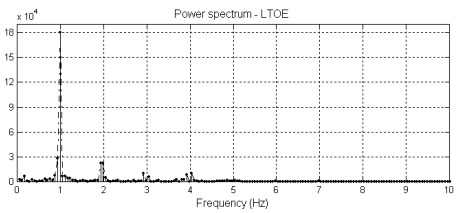
(g) knee



(h) ankle



(i) right hallux (big toe)



(j) left hallux (big toe)

Figure 3: Examples of vertical-component power spectra of a clockwise circle-walking person. $N = 1024$, $f_\delta = 60$ Hz, Gc ≈ 1 Hz.

a band-limited frequency $[0.4 \sim 5.0]$ Hz, $\Delta\mathrm{f} = \mathrm{f}_\delta / N$ denotes the frequency resolution, f_δ denotes the chosen sample rate, such as 60 Hz used in our experiments for human motion. The frequency corresponding to the maximum power magnitude in the first clustering of the resulting spectrum sum,

$$k^* := \max_k \left\{ \sum_i P_{i(k)} \right\} \tag{3}$$

is regarded as the activity cycle, or generalised gait-cycle (Gc= $k^* \Delta\mathrm{f}$).

The detected cycle frequency is subsequently used to normalise the frequency axis of power spectrum from Hz to generalised Gait-cycle (Gc). Figure 4 shows examples of Gc-scaled whole-body power spectra [1].

4.2 Feature power vector and motion template

Observably, frequency resolutions $\Delta\mathrm{f} = \mathrm{f}_\delta / N$ differ for the different trial lengths N. There is not a consistent number of spectral components among these spectra. It is therefore impossible to implement component-by-component comparison. A uniform motion template for trials is needed in order to make direct spectral comparison. Moreover in practice, a random trial length N may contain only a few motion cycles and result in an imprecise spectrum.

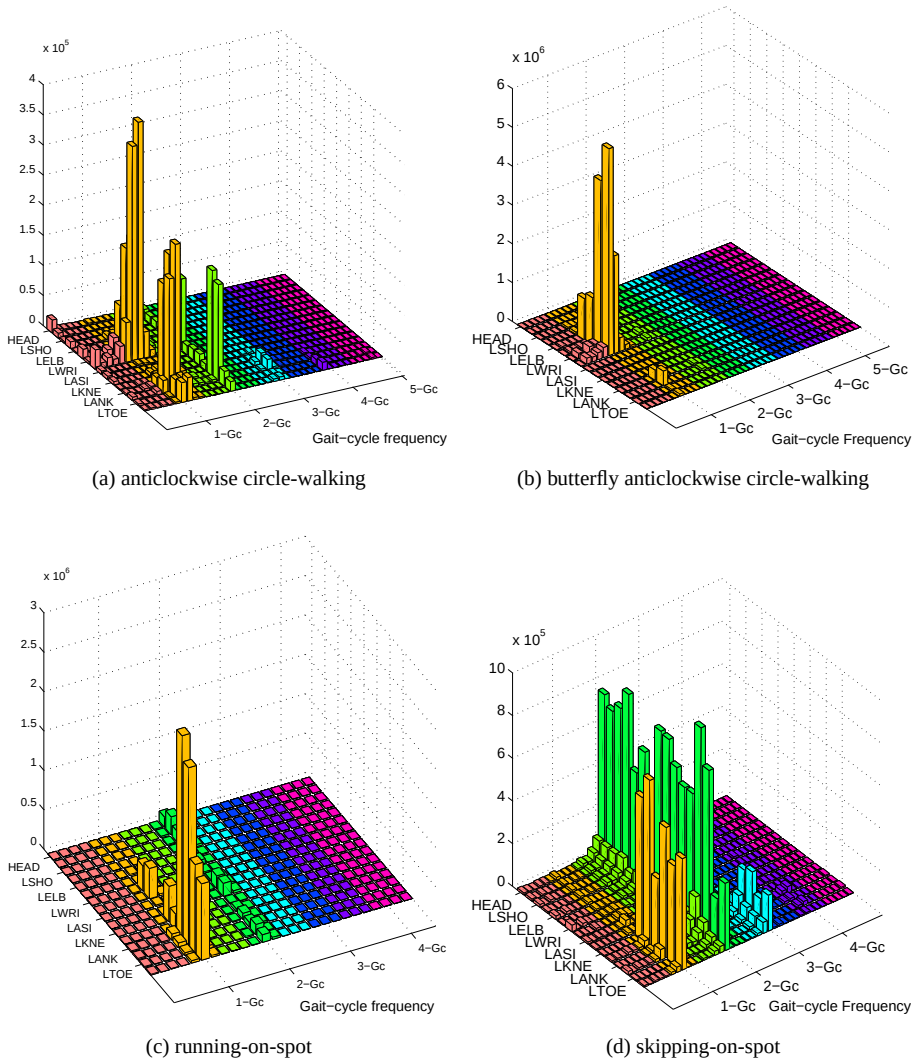
To obtain a uniform motion template, considering the nature of clustering distributions in power spectra, we extract a set of dominant power components around Gc and its multiples from each spectrum P_i, and regard the result as a *feature power vector* $\vec{\nu}_i$ of the feature point i:

$$\begin{aligned}
\nu_{i(0)} &= DC_i \ , \\
\nu_{i(n)}|_{n=1,\ldots 4} &= \sum_{k \in W_n} P_{i(k)} \ , \\
\nu_{i(5)} &= \sum_{k \notin W_{1,\ldots 4},\, k \neq 0} P_{i(k)} \ , \\
\nu_{i(6)} &= \sum_{k \neq 0} P_{i(k)} \ .
\end{aligned} \tag{4}$$

The first component $\nu_{i(0)}$ of the vector $\vec{\nu}_i$ is the DC component in the Fourier decomposition, equal to $\frac{1}{2} a_{i(0)}$, denoting the average vertical position of this point. To mitigate the frequency resolution problem, in $\nu_{i(n)}$ where $n = 1,\ldots 4$, we utilise *sum-windows* W_n to sum power components within the range of $\pm 20\%$Gc around the n^{th}Gc. The parameter $\nu_{i(5)}$ is used to represent the non-selected "small power" components that have not been included in the $\pm 20\%$Gc window powers $\nu_{i(1)}$ to $\nu_{i(4)}$. The last item $\nu_{i(6)}$ is used to represent the sum of motion powers over all frequencies of feature point i. *Feature power vectors* aggregate the frequency components into just 7 feature elements ($\nu_{i(0)}$ to $\nu_{i(6)}$). This makes comparisons with differently sampled trials possible, and condenses frequency power matching to efficient aggregate matching.

We stack I *feature power vectors* of feature points together as a *motion template*, $V = \{\vec{\nu}_i \mid i = 1 \ldots I\}$. The template can be viewed as an $I \times 7$ array. Each column in the

[1] In Figure 4 and Figure 5, 16 feature points are arranged in the order of HEAD, BACK, LSHO, RSHO, LELB, RELB, LWRI, RWRI, LASI, RASI, LKNE, RKNE, LANK, RANK, LTOE, RTOE.



(a) anticlockwise circle-walking

(b) butterfly anticlockwise circle-walking

(c) running-on-spot

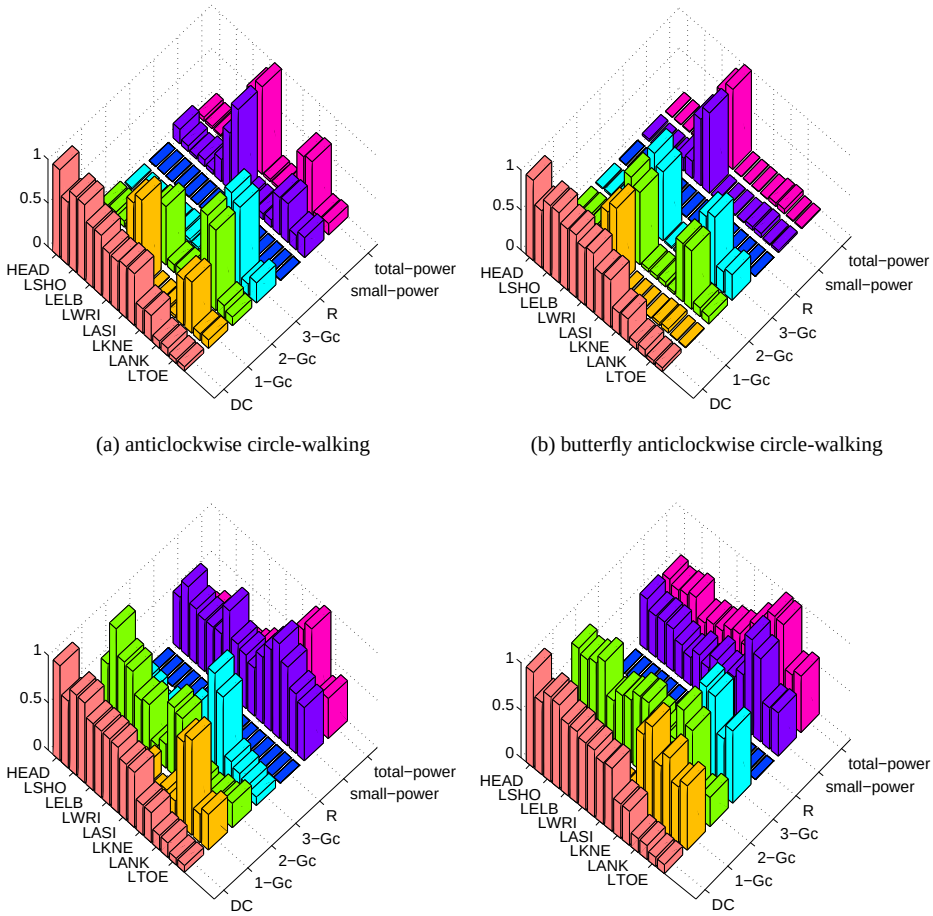
(d) skipping-on-spot

Figure 4: Gc-scaled whole-body power spectra.

motion template is scaled relative to the maximum value in this column. By this means, the power amplitudes are normalised for subjects. After normalisation, averaging is used to reduce small differences in motion templates among different subjects to generate a *standard motion template* for a specific motion. Some examples of motion templates are shown in Figure 5. In experiments, we found high-frequency feature power beyond 3^{rd}Gc are very small, so we set $\nu_{i(4)}$ to zero.

4.3 Motion recognition

Motion recognition is straightforward at this stage. It is achieved by finding the best match between an observed motion template and pre-stored standard motion templates.



(a) anticlockwise circle-walking

(b) butterfly anticlockwise circle-walking

(c) running-on-spot

(d) skipping-on-spot

Figure 5: Examples of motion templates.

We apply the algorithm stated above to an observed motion to generate its motion template $U = \{\vec{\mu}_j \mid j = 1 \ldots J\}$, with J unidentified *feature power vectors*. The feature points of the observed motion can be an adequate subset ($J \leq I$) of standard templates. We use an $J \times I$ *match matrix* $M = \{m_{(j,i)} \mid j = 1 \ldots J,\ i = 1 \ldots I\}$, see Equation 5, to store the weighted difference between each jth motion vector $\overrightarrow{\mu}_j$ in the observed template and each ith motion vector $\overrightarrow{\nu}_i$ in a model template,

$$m_{(j,i)} = \sum_{n=0}^{6} \left|\mu_{j(n)} - \nu_{i(n)}\right| \omega_n \ , \tag{5}$$

where $\vec{\omega} = [\omega_0,\ \omega_1,\ \ldots,\ \omega_6]$ is the weight vector. In our experiments, we used $|\vec{\omega}|_1 = 1$, $\vec{\omega} = [0.34,\ 0.2,\ 0.2,\ 0.03,\ 0,\ 0.03,\ 0.2]$.

In this motion template comparison, the first element related to DC-component is weighted by 0.34, as it significantly indicates the identity of a feature point from the

average normalised vertical position relative to the origin on the ground. We found from our experiments that principal spectral powers distributed around 1^{st}Gc and 2^{nd}Gc are nearly 10 times larger than the power around 3^{rd}Gc. We therefore set the comparison weights of 1^{st}Gc, 2^{nd}Gc and 3^{rd}Gc elements to be $0.2,\ 0.2$ and 0.03, respectively. The fifth weight in $\vec{\omega}$ with value 0.03 is used for the comparison of small power which may only occupy less than 10% of the total motion power. This element is very possibly interfered by motion irregularity and noise, and can not be highly weighted. The last element weighted by 0.2 implies motion intensity in terms of total power.

The best match of point j is taken to be the point corresponding to the minimum element $\min_i\{m_{(j,i)}\}$ in the jth row of *match matrix* M. This allows *motion power spectral similarity* $\mathcal{S}_{fft}$ to be defined from all best matches of the J feature points as

$$\mathcal{S}_{fft} = \left(1 - \frac{\sum_{j=1}^{J} \min_i\{m_{(j,i)}\}}{J} \right)^3 . \tag{6}$$

The motion with maximum similarity $\mathcal{S}_{fft}$ for all the searched templates is taken to indicate recognition.

5 Experimental Results

All experiments were conducted using real motion capture data from a marker-based optical motion capture system, the Vicon 512, as described in section 3. Recognition was tested on simple periodic activities, namely walking-on-spot (Walk-S), circle-walking (Walk-C, Walk-A for clockwise/anticlockwise directions), clockwise butterfly-walking (waving hands up and down, B-Walk-C), running-on-spot (Run-S), clockwise circle-running (Run-C), skipping type 1 (feet stepping alternately, Skip1), skipping type 2 (feet stepping together, Skip2), jumping type 1 (arms raised to horizontal level, Jump1), jumping type 2 (arms raised over head, Jump2).

Recognition is indicated by the *motion power spectral similarity* $\mathcal{S}_{fft}$ in Table 1. The results are averaged for a group of subjects that consisted of males and females, with ages from 5 to 60 years. The highest column entries, highlighted, occur when the observed activity matches the motion template activity. Because the algorithm utilises a whole-body power spectral analysis approach, recognition could be achieved by comparison of the movement on each feature point. For example, similarity parameters $\mathcal{S}_{fft}$ of an observed circle-walking with each model activity decrease in the order: circle-walking, walking-on-spot, clockwise butterfly-walking in a circle, circle-running, running-on-spot, skipping and jumping.

From the results above, we find that the proposed power spectral analysis approach is able to classify periodic motions solely on unidentified vertical-component trajectories. The different classes of movements, such as walking, running, jumping and skipping, are discriminated by the similarity parameter $\mathcal{S}_{fft}$. Even with similar movements, such as running-on-spot and circle-running, there is discrimination because the magnitudes of power spectra for left and right limbs have a bias in circular activities.

Table 1: Recognition of human periodic movements by $\mathcal{S}_{fft}$ and $\mathcal{S}_{fft\text{-}Gc}$.

	observed activity $\mathcal{S}_{fft}$ ($\mathcal{S}_{fft\text{-}Gc}$)								
activity motion template	walk-C .75∼1.1 Hz	walk-S .8∼1.0 Hz	B-walk-C .8∼1.1 Hz	run-C 1.3∼1.5 Hz	run-S 1.3∼1.5 Hz	jump1 .9∼1.2 Hz	jump2 .9∼1.1 Hz	skip1 .8∼1.1 Hz	skip2 1.4∼1.9 Hz
Walk-C (Gc=0.90 Hz)	**.88 (.87)**	.78 (.82)	.83 (.85)	.68 (.63)	.69 (.63)	.68 (.69)	.70 (.75)	.60 (.65)	.59 (.51)
Walk-A (Gc=0.90 Hz)	.87 (.86)	.78 (.82)	.83 (.85)	.68 (.63)	.69 (.63)	.67 (.68)	.70 (.75)	.59 (.64)	.59 (.50)
Walk-S (Gc=0.93 Hz)	.81 (.83)	**.79 (.83)**	.80 (.81)	.67 (.64)	.68 (.64)	.66 (.68)	.69 (.75)	.59 (.65)	.58 (.51)
B-Walk-C (Gc=0.92 Hz)	.75 (.78)	.72 (.76)	**.90 (.90)**	.60 (.58)	.61 (.57)	.71 (.73)	.72 (.77)	.49 (.56)	.57 (.50)
Run-C (Gc=1.39 Hz)	.73 (.71)	.70 (.69)	.74 (.72)	**.84 (.86)**	.77 (.79)	.70 (.72)	.72 (.72)	.69 (.68)	.58 (.61)
Run-S (Gc=1.42 Hz)	.71 (.69)	.71 (.70)	.63 (.63)	.75 (.78)	**.85 (.88)**	.65 (.67)	.63 (.64)	.70 (.69)	.50 (.53)
Jump1 (Gc=1.1 Hz)	.58 (.62)	.56 (.62)	.74 (.77)	.63 (.61)	.64 (.63)	**.90 (.91)**	.86 (.86)	.55 (.63)	.53 (.46)
Jump2 (Gc=0.92 Hz)	.60 (.66)	.62 (.69)	.73 (.76)	.63 (.60)	.63 (.58)	.83 (.82)	**.91 (.92)**	.54 (.62)	.45 (.37)
Skip1 (Gc=0.97 Hz)	.62 (.64)	.60 (.66)	.48 (.56)	.65 (.63)	.67 (.64)	.52 (.59)	.52 (.60)	**.77 (.81)**	.54 (.46)
Skip2 (Gc=1.70 Hz)	.63 (.60)	.61 (.60)	.60 (.59)	.61 (.65)	.62 (.67)	.56 (.58)	.54 (.54)	.57 (.56)	**.78 (.80)**

6 Discussion

We have found that the parameter $\mathcal{S}_{fft}$ reflects motion power characteristics of the whole-body, giving rise to recognition possibility. The parameter $\mathcal{S}_{fft}$ has been made insensitive to inter-subject variability for the same activity, by scaling with respect to the Gc. The same scaling, however, has also lost the important discriminating factor of speed among different activities, represented by the value of Gc itself. We therefore considered activity periodicity assisted recognition, defined as $\mathcal{S}_{fft\text{-}Gc} = \{\mathcal{S}_{fft},\ \mathcal{S}_{Gc}\}$, formally

$$\mathcal{S}_{Gc} = 1 - \frac{|Gc_{\text{observed}} - Gc_{\text{model}}|}{Gc_{\text{model}}}\ , \tag{7}$$

$$\mathcal{S}_{fft\text{-}Gc} = 0.8\mathcal{S}_{fft} + 0.2\mathcal{S}_{Gc}\ . \tag{8}$$

As shown in Table 1, the combined similarity parameter $\mathcal{S}_{fft\text{-}Gc}$ increases the ability to distinguish motions with largely different activity periodicity, such as running and walking. It will not contribute to distinguishing motions with similar activity periodicity, such as walking and Jump2, with activity periodicities Gc around 0.92 Hz[2].

For best frequency resolution, we use the Fourier transform for the whole sequence of length N when $N < 7 \times \mathrm{f}_\delta / G_c$, rather than the truncated sequence of length $2^{\lfloor \log_2 N \rfloor}$. In this case, the FFT is reduced to the raw Discrete Fourier Transform (DFT) algorithm. This is a tradeoff between frequency resolution and computational cost. Nevertheless, because each motion template is efficiently coded by a small number of parameters using *feature power vectors*, and because we also only use one standard template for indexing each kind of motion, a short matching time and small database are required. The method proved to be computationally efficient.

7 Conclusions

A frequency domain approach based on whole-body spectral analysis is proposed for efficient modelling and recognition of articulated periodic activities. The approach is motion-based using unidentified trajectories of feature points from MLDs. We apply *power spectral analysis* on each vertical-component trajectory from 3D-MLDs and detect a set of *feature power vectors*. These feature power vectors are used to generate a *motion template* for given periodic activity, averaged for a number of subjects, after normalisation both on frequency and power magnitude. This allows recognition to be carried out in frequency domain for a wide range of subjects. Recognition involves comparison of a template of the observed motion with standard motion templates to find a best match. The choice of feature points is not prescribed, the only requirements being that chosen feature points effectively reflect motion cues and are common to all templates.

The experimental results indicate that frequency domain analysis allows classification of periodic motions without identification of underlying kinematic structure. It demonstrates that inherent characteristics of human periodic movements exist not only in the spatio-temporal domain of MLDs, but also in the frequency domain. Frequency domain features hint motion nature which can be used to classify different periodic activities, and even discriminate similar movements within the same class.

[2] Averaged activity periodicity Gc of each periodic movement used the standard motion template and the Gc range of a group of observed motion are indicated in Table 1.

Whether frequency domain components also include information for individual subject recognition, and what kinds of features would be necessary for a more precise, reliable and unique interpretation to distinguish individuals, are still open questions.

Acknowledgements

All 3D-MLDs data of human periodic motion used in this paper were obtained with a 7-camera Vicon marker-based optical motion capture system, installed at the Department of Computer Science, UWA.

References

Aggarwal, J. K. and Cai, Q. (1999). Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440.

Angeloni, C., Riley, P. O., and Krebs, D. E. (1994). Frequency content of whole body gait kinematic data. *IEEE Trans. Rehabilitation Engineering*, 2(1):40–46.

Barclay, C. D., Cutting, J. E., and Kozlowski, L. T. (1978). Temporal and spatial factors in gait perception that influence gender recognition. *Perception and Psychophysics*, 23(2):145–152.

Boyd, J. and Little, J. (1997). Global versus structured interpretation of motion: moving light displays. In *Proc. IEEE Computer Vision and Pattern Recognition*.

Campbell, L. and Bobick, A. (1995). Recognition of human body motion using phase space constraints. In *Proc. IEEE Int. Conf. Computer Vision*, pages 624–630, Cambridge.

Cédras, C. and Shah, M. (1994). A survey of motion analysis from moving light displays. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 214–221, Washington.

Cédras, C. and Shah, M. (1995). Motion-based recognition: A survey. *Image and Vision Computing*, 13(2):129–155.

Cutting, J. E. and Kozlowski, L. T. (1977). Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, 9(5):353–356.

Dorfmüller, K. (1999). Robust tracking for augmented reality using retroreflective markers. *Computers and Graphics*, 23:795–800.

Ferrigno, G. and Gussoni, M. (1988). A procedure to automatically classify markers in biomechanical analysis of whole body movement in different sport activities. *Medical & Biological Engineering & Computing*, 26:321–324.

Fujiyoshi, H. and Lipton, A. (1998). Real-time human motion analysis by image skeletonization. In *Proc. IEEE Workshop on Applications of Computer Vison*.

Gavrila, D. M. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98.

Gleicher, M. (1999). Animation from observation: Motion capture and modtion editing. *Computer Graphics*, 33(4):51–55.

Goddard, N. H. (1992). *The Perception of Articulated Motion: Recognizing Moving Light Displays*. PhD thesis, University of Rochester.

Hill, H. H. and Pollick, F. E. (2000). Exaggerating temporal differences enhances recognition of individual from point light displays. *Psychological Science*, 11:223–228.

Johansson, G. (1975). Visual motion perception. *Scientific American*, 232(6):75–80, 85–88.

Köhle, M. and Merkl, D. (1996). Things we observed when watching people walk: Classification of gait patterns with self-organizing maps. In *Proc. 7th Australian Conf. Neural Networks ACNN'96*, Canberra.

Michael, M. W. (1996). *Gait analysis: An introduction*. Butterworth-Heinemann Ltd, Oxford.

Moeslund, T. B. and Granum, E. (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268.

Polana, R. and Nelson, R. (1993). Detecting activities. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 2–7.

Richards, J. (1999). The measurement of human motion: A comparison of commercially available systems. *Human Movement Science*, 18(5):589–602.

Söderkvist, I. and Wedin, P. A. (1993). Determining the movements of the skeleton using well-configured markers. *Journal of Biomechanics*, 26(12):1473–1477.

Stoddart, A. J., Mrázek, P., Ewins, D., and Hynd, D. (1999). Marker based motion capture in biomedical application. *IEE Electronics & Communications*, 103.

Tsai, P. S., Shah, M., Keiter, K., and Kasparis, T. (1994). Cyclic motion detection for motion based recognition. *Pattern Recognition*, 27(12):1591–1603.

AISB Journal **1**(5), © SSAISB, 2004
http://www.aisb.org.uk

Biologically Motivated Multi-modal Processing of Visual Primitives

Norbert Krüger*, Markus Lappe† and Florentin Wörgötter‡

* Department of Computer Science and Engineering, Aalborg University Esbjerg, *nk@cs.aue.auc.dk*
† Psychologisches Institut, Universität Münster, Germany, *mlappe@uni-muenster.de*
‡ Department of Psychology, University of Stirling, Scotland, UK, *worgott@cn.stir.ac.uk*

Abstract

We describe a new kind of image representation in terms of local multi–modal Primitives. These Primitives are motivated by processing of the human visual system as well as by functional considerations. We discuss analogies of our representation to human vision and concentrate specifically on the implications of the necessity of communication of information in a complex multi-modal system.

1 Introduction

In this paper, we describe a new kind of image representation in terms of local multi–modal Primitives (see Figure 1). These Primitives are motivated by processing in the human visual system as well as by functional considerations. The work described here has been evolved from a project started in 1998 which has been focused on the integration of visual information (ModIP, 2003). The image representation described here is now a central pillar of the ongoing European project (ECOVISION, 2003) that focuses on the functional modelling of early visual processes.

In the human visual system beside local orientation also other modalities such as colour and optic flow (that are also part of our multi–modal Primitives) are computed in the hyper-columns of V1 (Hubel and Wiesel, 1969; Gazzaniga, 1995). *All these low level processes face the problem of an extremely high degree of vagueness and uncertainty (Aloimonos and Shulman, 1989).* This arises from a couple of factors. Some of them are associated with image acquisition and interpretation: owing to noise in the acquisition process along with the limited resolution of cameras, only erroneous estimates of semantic information (e.g., orientation) are possible. Furthermore, illumination variation heavily influences the measured grey level values and is hard to be modelled analytically (Ikeuchi and Horn, 1981). Information extracted across image frames, e.g., in stereo and optic flow estimation, faces (in addition to the above mentioned problems) the correspondence and aperture problem which interfere in a fundamental and especially difficult way (Ayache, 1990; Klette et al., 1998).

However, the human visual system acquires visual representations which allow for actions with high precision and certainty within the 3D world under rather uncontrolled conditions. *The human visual system can achieve the needed certainty and completeness*

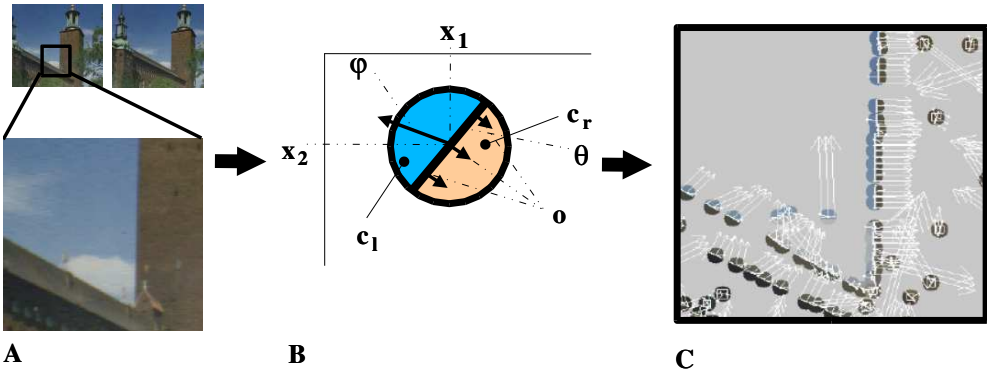


Figure 1: **A:** Image sequence and frame. **B:** Schematic representation of the multi–modal Primitives. **C:** Extracted Primitives at position with high amplitude.

by integrating visual information across modalities (Hibbard et al., 2000) and by utilising spatial and temporal interdependencies (Phillips and Singer, 1997; Hoffman, 1980). This integration is manifested in the huge connectivity between brain areas in which the different visual modalities are processed as well as in the large number of feedback connections from higher to lower cortical areas (Gazzaniga, 1995). The essential need for integrating visual information in addition to optimising single modalities to design efficient artificial visual systems has also been recognised in the computer vision community after a long period of work on improving single modalities (Aloimonos and Shulman, 1989).

However, integration of information makes it necessary that local feature extraction is subject to modification by contextual influences. As a consequence *adaptability* must be an essential property of the visual representation. Moreover, the exchange of information between visual events has necessarily to be paid for with a certain cost. This cost can be reduced by limiting the amount of information transferred from one place to the other, i.e. by reducing the bandwidth. This is the reason why we are after a *condensed* description of a local image patch, which however *preserves the relevant information*. Here relevance has to be understood not only in an information theoretical sense, but in a global sense (the system has to be subject to modifications by global interdependencies, in particular local entities have to be connectable to more complex entities) and action oriented sense (the transfered information has to be relevant for the actions the individual has to perform).

Taking the above mentioned considerations into account, the Primitives, which are the basic entities of our image representation, can be characterised by four properties:

> **Multi-modality:** Different domains that describe different kinds of structures in visual data are well established in human vision and computer vision. For example, a local edge can be analysed by local feature attributes such as orientation or energy in certain frequency bands (Krüger and Sommer, 2002). In addition, we can distinguish between line and step–edge like structures (contrast transition). Furthermore, colour can be associated to the edge. This image patch also changes in time due to ego-motion or object motion. Therefore time specific features such as a 2D velocity vector (optic flow) are associated to our Primitives (see Figure 1).

Adaptability: Since the interpretation of local image patches in terms of the above mentioned attributes as well as classifications such as 'edge–ness' or 'junction–ness' are necessarily ambiguous when based on local processing (Krüger and Felsberg, 2003), stable interpretations can only be achieved *through integration* by making use of contextual information (Aloimonos and Shulman, 1989). Therefore, all attributes of our Primitives are equipped with a confidence that is essentially *adaptable according to contextual information* expressing the reliability of the attribute. Furthermore, feature attributes themselves are subject to correction mechanisms that use contextual information.

Condensation: Integration of information requires *communication between Primitives* expressing spatial (Krüger and Wörgötter, 2002; Krüger et al., 2002b) and temporal dependencies (Krüger et al., 2002a). This communication has necessarily to be paid for with a certain cost (as will be made explicit in section 3). This cost can be reduced by limiting the amount of information transferred from one place to the other, i.e., by reducing the bandwidth. Therefore we are after a *condensed* representation. Also for other tasks it is essential to store information in a *condensed way*, e.g., for the learning of objects to reduce memory requirements.

Meaningfulness: Communication and memorisation not only require a reduction of information. We want to reduce the amount of information within an image patch *while preserving perceptually relevant information*. This leads to *meaningful* descriptors such as our attributes position, orientation, contrast transition, colour and optic flow.

We will describe our feature processing in section 2 and will compare it to early human visual processing in Section 3.

2 Feature Processing and Application

In this section we describe the coding of modalities associated to our Primitives. In addition to the position $\mathbf{x}$, we compute the following semantic attributes and associate them to our Primitives (see also Figure 1).

Frequency: We describe the signal on different frequency levels f independently. Often the decision in which frequency band the relevant information does occur is difficult, therefore we leave this decision open to be decided at later stages of processing. It may be even that for the same position on different frequency levels there occur different kinds of semantic information (for example, the top of the toy in Figure 2A on a high frequency level can be described as texture–like while on a lower frequency level it resembles an edge).

Orientation: The local orientation associated to the image patch is described by θ. The orientation θ is computed by interpolating across the orientation information of the whole image patch to achieve a more reliable estimate. This holds also true for the following feature attributes contrast transition, colour and optic flow.

Contrast transition: The contrast transition is coded in the phase φ of the applied filter (Felsberg and Sommer, 2001). The phase codes the local symmetry, for example a bright line on a dark background has phase 0 while a bright/dark edge has phase $-\pi/2$ (in Figure 3 the line that marks the border of the street is represented as a line or two edges depending

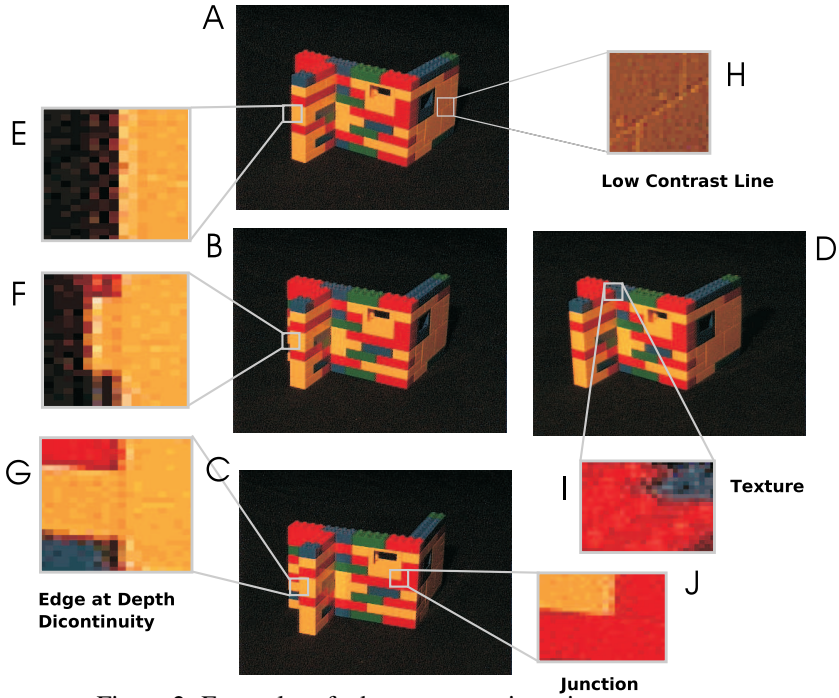


Figure 2: Examples of edge structures in an image sequence.

on the distance from the camera). In case of boundaries of objects, the phase represents a description of the transition between object and background (Kovesi, 1999; Krüger and Wörgötter, 2002).

Colour: Colour $(\mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r)$ is processed by integrating over image patches in coincidence with their edge structure (i.e., integrating separately over the left $(\mathbf{c}^l)$ and right $(\mathbf{c}^r)$ side of the edge as well as a middle strip $(\mathbf{c}^m)$ in case of a line structure). In case of a boundary edge of a moving object at least the colour at one side of the edge is expected to be stable (see Figure 2E–G) since it represents a description of the object.

Optic Flow: Local displacements $\mathbf{o}$ is computed by the well known optic flow technique (Nagel, 1987).

Furthermore, we represent the system's confidence c that the entity e does exist. We end up with a parametric description of a Primitive as

$$E = (\mathbf{x}, f, \theta, \varphi, (\mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r), \mathbf{o}; c).$$

In addition, to each of the parameters $\varphi, (\mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r), \mathbf{o}$ there exist confidences $c_i, i \in \{\varphi, \mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r, \mathbf{o}\}$ that code the reliability of the specific sub–aspects that is also subject to contextual adaptation.

We have applied our image representation to different contexts. First, an image patch also describes a certain region of the 3D space and therefore 3D attributes can be associated such as a 3D-position and a 3D-direction. In (Krüger et al., 2002b; Pugeault and Krüger, 2003), we have defined a stereo similarity function that makes use of multiple-modalities to enhance matching performance. Second, the Primitives can be subject to spatial contextual modification. We define groups of Primitives based on a purely statistical criterion in (Krüger and Wörgötter, 2002). Once these groups are defined, we

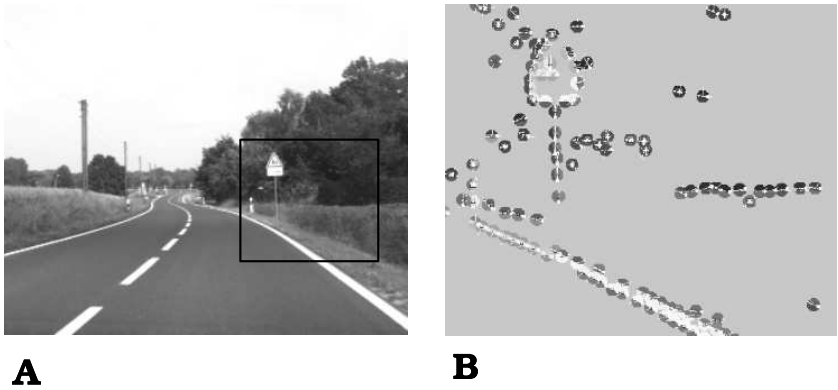


Figure 3: A: Original Image. B: Extracted Primitives with high amplitude.

modulate the confidences of our Primitives: confidences are increased if the Primitives are part of a bigger group, otherwise the confidences are decreased. Thirdly, we have stabilised features according to the temporal context. In (Krüger et al., 2002a; Krüger et al., 2002c), we make use of the motion of an object to predict feature occurrences and showed that we can stabilise stereo processing by modifying the confidences according to the temporal context.

3 Hyper-columns of Basic Processing Units in early Vision

In this section, we discuss aspects of the processing of visual information in the human visual system and draw analogies to our image representation.

The main stream of visual information in the human visual system goes from the two eyes to the LGN (Lateral Geniculate Nucleus) and then to area V1 in the cortex (see Figure 4 and (Wurtz and Kandel, 2000a)). There are two kinds of cell types involved (M (magnocellular) and P (parvocellular) cells) that have different response characteristics: M cells have a low spatial but high temporal resolution and are not colour sensitive. In contrast to M cells, P cells have a low temporal and high spatial resolution and are colour sensitive. Both kinds of cells project into two cortical pathways, the dorsal and ventral pathway (see Figure 4). The ventral pathway goes from the cortical area V1 to V2 to the Inferior Temporal Area (IT) and is believed to be mainly responsible for object recognition (Tanaka, 1993). In the dorsal stream information is transferred from V1 to MT (Middle Temporal Area) to MST (Medial Superior Temporal Area) and is believed to be involved in the analysis of motion and spatial information.

V1 (or Visual Area 1) is the main input of both pathways. The structure of V1 has been investigated by Hubel and Wiesel in their ground-breaking work (Hubel and Wiesel, 1962; Hubel and Wiesel, 1969). V1 is organised in a retinotopic map that has a specific repetitively occurring pattern of substructures called hyper-columns. Hyper-columns themselves contain so called orientation columns and blobs (see Figure 5). The main input of V1 comes from the LGN and targets to layer 4 to which information of both eyes projects (see Figure 5Aiii).

The orientation columns are organised in an ordered way such that columns representing similar orientations tend to be adjacent to each other (see Figure 5Ai). However, it is not only orientation that is processed in an orientation column but the cells are sensitive

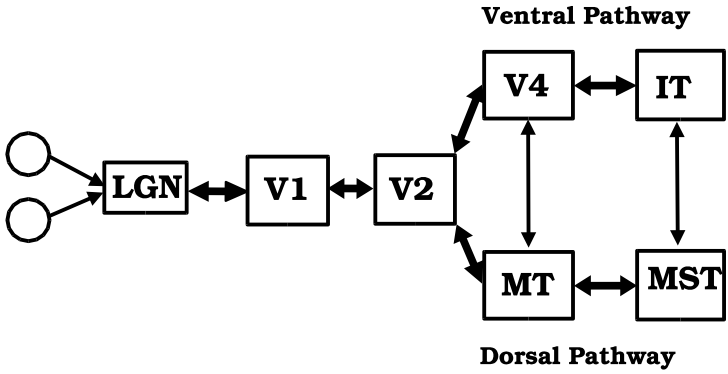


Figure 4: Flow of visual information in the human visual system (schematic).

to additional attributes (see Figure 5D) such as disparity (Barlow et al., 1967; Parker and Cumming, 2001), local motion (Wurtz and Kandel, 2000b), colour (Hubel and Wiesel, 1969) and phase (Jones and Palmer, 1987). Also specific responses to junction–like structures could be measured (Shevelev et al., 1995). Therefore, it is believed that in V1 basic local feature descriptions are processed similar to the feature attributes coded in our Primitives. However, since the processing is local,[1] the ambiguities of visual information is not resolved at this level. For example, response properties of neurons in V1 reflect the aperture problem (Stumpf, 1911). This holds also for our Primitives since the flow is also computed by a local operation.

It is believed that mainly form is processed in the ventral pathway. Neurophysiological equivalents of illusionary contours can be detected in V2 but not in V1 (von der Heydt et al., 1984). This is not surprising since illusionary contours like in the Kanizsa triangle (Kanizsa, 1976) presuppose an integration of information across a large spatial domain as well as across different feature types (e.g., edges and junctions) and can therefore only be processed at a later stage.

The different visual modalities are not computed independently but are combined. For example in V1 the processing of motion is necessarily intertwined with the processing of orientation because of the aperture problem. In V4, colour and orientation is combined (Wurtz and Kandel, 2000b). Accordingly, in our image representation the coding of colour is deeply intertwined with the coding of orientation. Colour is a feature that describes homogeneous surfaces. However, orientation describes discontinuities and can be used to separate the surfaces. In our image representation we therefore first compute orientation and then compute a left and a right colour according to this orientation.

In the dorsal pathway mainly motion is analysed. Like the occurrence of illusionary contours presuppose global interactions, the aperture problem can only be solved by taking the global context into account. This does not happen (and can not happen because of the local processing) in V1. However, in MT and MST many cell responses indicate a solution to aperture problem (Pack and Born, 2001; Wurtz and Kandel, 2000b). Similar to the cells in V1, our Primitives also reflect the aperture problem. However, we can use the output of our Primitives to apply global mechanisms that disambiguate the local flow.

As in the ventral pathway, cells in the dorsal pathway show multi-modal response

[1]There is a high connectivity within a hyper-column. There exist also connections across hyper-columns. However their distribution falls sharply with distance.

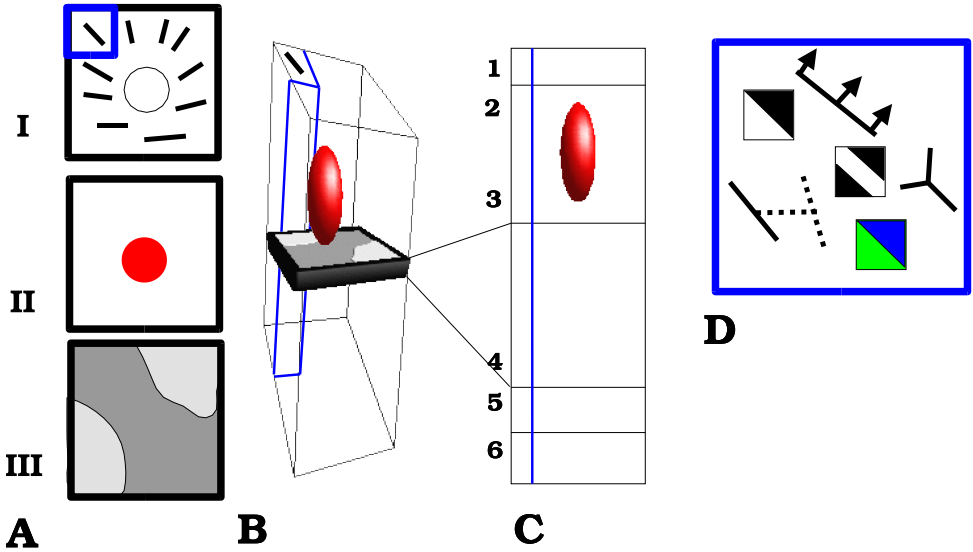


Figure 5: Hyper-columns in V1. A: There exist three physiological distinguishable sub-structure in a hyper-column: (i) in orientation columns information about oriented edge structure is represented in a topological way. (ii) Colour information is coded in so called 'blobs'. (iii) Information of both eyes are input to the fourth cortical layer (see also B). B: three–dimensional structure of a hyper-column. C: organisation in cortical layers. D: feature attributes that are coded in a hyper-column.

patterns. For example, a moving edge may not be visible as a luminance edge but can be constituted by colour or texture. MT cells respond to these kinds of structures although they are not sensitive to colour alone (Thiele et al., 1999; Wurtz and Kandel, 2000b).

Let us summarise. In V1 visual information is mainly locally processed. However, some semi–local interactions exist. The ambiguities of visual information can not be resolved at this stage of processing. A specialisation to form processing (along the ventral pathway V1–V2–V4–IT) and motion processing (along the dorsal pathway V1–V2–MT–MST) does occur.

As mentioned above, stable and reliable information can only be achieved by disambiguation through integration. However, this integration process makes the exchange of information within and across visual areas mandatory. As discussed before, intra–areal connections are very limited. However, inter–areal connection project to a much wider field of the next layer.

Regarding communication between visual areas we have to address two issues:

1) What is the bandwidth of information we want to transfer ("quantity")?

2) What kind of information do we want to communicate ("quality")?

The first question leads to a reflection about costs of communication. In any communication system transfer of information is associated to a cost which normally increases with the amount of information to be transferred and with the distance to be covered. This could concern the costs of "cables" but also the cost of the energy used for the transfer (Attwell and Laughlin, 2001).

In the brain, the communication between two neurons is realized by an axon docking to the soma or the dendrites of other neurons. Accordingly, the complexity and, thus, the "cost" of communication increases with the number of connections. This holds in a very general sense and may have been one driving force for the bandwidth reduction that is actually observed in neuronal visual processing. This bandwidth reduction most clearly manifests itself in mechanisms of visual attention and visual awareness. Focused attention is often taken as one central mechanisms used to reduce the bandwidth of computation as well as of information transfer in the brain to a manageable degree. Anatomically the bandwidth limitation requirement may be reflected by the density of fibres which connect different areas which is smaller than that which connects cells within a hyper-column.

A similar mechanism is also used in our image representation were we arrive at a significant reduction of information following the first processing stages. Compared to an average sized image patch of 15×15 pixels represented by a Primitive the output of a Primitive has less than 20 values, i.e., we have a compression rate of more than 96%. This rate becomes even higher when we compare the output of a Primitive to intermediate local stages of processing where feature attributes for all modalities are derived for each pixel.

The second question above concerns the quality of information which needs to be transferred between the different stages of visual processing. Here we refer back to what we have said above noting that pre-processed visual information is exceedingly ambiguous as the consequence of fundamental problems in image data acquisition as well as resulting from the intrinsic structure of the detectors (receptive fields). This leads to the situation that redundant information must be transferred because only through redundancy it can be assured that erroneous information can be disambiguated. For this it is required that a visual event which is represented by the firing of neuron A has a relevance for the event represented by B. Since event A is supposed to be used to correct event B both events need to be highly correlated. This can be quantified by the following measure of statistical interdependencies:

$$\frac{P(B|A)}{P(B)}. \tag{1}$$

If this term takes a high value then there is a high likelihood of the occurrence of event B when we know event A has occurred compared to the likelihood of the occurrence of the event B without prior knowledge. In this case, events A and B can be used to mutually correct each other because they are carrying shared (i.e., redundant) information. The expression (1) has been called 'Gestalt coefficient' in (Krüger, 1998) where it was shown that applying binarised Gabor wavelets to natural images, a high Gestalt coefficient corresponds to the Gestalt laws Collinearity and Parallelism. As an extension of (Krüger, 1998), it has been shown in (Krüger and Wörgötter, 2002) that by using our multi–modal Primitives we can increase the statistical interdependencies measured by (1) significantly compared to using orientation only (Krüger, 1998). That means that by using our Primitives we can increase interdependencies of visual events. In this way in our Primitives not only information is condensed but transferred to *more meaningful descriptors*.

4 Conclusion

We have introduced a novel way to compute visual Primitives which are motivated by early processing in the human visual system in analogy to the output of V1 hyper-columns. These Primitives are multi-modal and give a dense and meaningful description of a scene. Our Primitives can adapt according to the spatial and temporal context that is realized in

the human visual system through a high synaptic connectivity. In this way the locally unreliable feature extraction can be disambiguated and stable feature representations can be achieved.

References

Aloimonos, Y. and Shulman, D. (1989). *Integration of Visual Modules — An extension of the Marr Paradigm*. Academic Press, London.

Attwell, D. and Laughlin, S. (2001). An energy budget for signalling in the grey matter of the brain. *Journal of Cerebral Bloodflow and Metabolism*, 21:1133–1145.

Ayache, N. (1990). *Stereovision and Sensor Fusion*. MIT Press.

Barlow, H., Blakemore, C., and Pettigrew, J. (1967). The neural mechanisms of binocular depth discrimination. *Journal of Physiology (London)*, 193:327–342.

ECOVISION (2003). Artificial visual systems based on early-cognitive cortical processing (EU–Project). *http://www.pspc.dibe.unige.it/ecovision/project.html*.

Felsberg, M. and Sommer, G. (2001). The monogenic signal. *IEEE Transactions on Signal Processing*, 49(12):3136–3144.

Gazzaniga, M. (1995). *The Cognitive Neuroscience*. MIT Press.

Hibbard, P., Bradshaw, M., and Eagle, R. (2000). Cue combination in the motion correspondence problem. *Proceedimgs of the Royal Society London B*, 267:1369–1374.

Hoffman, D., editor (1980). *Visual Intelligence: How we create what we see*. W.W. Norton and Company.

Hubel, D. and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Phyiology*, 160:106–154.

Hubel, D. and Wiesel, T. (1969). Anatomical demonstration of columns in the monkey striate cortex. *Nature*, 221:747–750.

Ikeuchi, K. and Horn, B. (1981). Numerical shape from shading and occluding boundaries. *Artificial Intelligence*, 17:141–184.

Jones, J. and Palmer, L. (1987). An evaluation of the two dimensional Gabor filter model of simple receptive fields in striate cortex. *Journal of Neurophysiology*, 58(6):1223–1258.

Kanizsa, G. (1976). Subjective contours. *Scientific American*.

Klette, R., Schlüns, K., and Koschan, A. (1998). *Computer Vision - Three-Dimensional Data from Images*. Springer.

Kovesi, P. (1999). Image features from phase congruency. *Videre: Journal of Computer Vision Research*, 1(3):1–26.

Krüger, N. (1998). Collinearity and parallelism are statistically significant second order relations of complex cell responses. *Neural Processing Letters*, 8(2):117–129.

Krüger, N., Ackermann, M., and Sommer, G. (2002a). Accumulation of object representations utilizing interaction of robot action and perception. *Knowledge Based Systems*, 15:111–118.

Krüger, N. and Felsberg, M. (2003). A continuous formulation of intrinsic dimension. *Proceedings of the British Machine Vision Conference*.

Krüger, N., Felsberg, M., Gebken, C., and Pörksen, M. (2002b). An explicit and compact coding of geometric and structural information applied to stereo processing. *Proceedings of the workshop 'Vision, Modeling and VISUALIZATION 2002'*.

Krüger, N., Jäger, T., and Perwass, C. (2002c). Extraction of object representations from stereo imagesequences utilizing statistical and deterministic regularities in visual data. *DAGM Workshop on Cognitive Vision*, pages 92–100.

Krüger, N. and Wörgötter, F. (2002). Multi modal estimation of collinearity and parallelism in natural image sequences. *Network: Computation in Neural Systems*, 13:553–576.

Krüger, V. and Sommer, G. (2002). Wavelet networks for face processing. *JOSA*, 19:1112–1119.

ModIP (2003). Modality Integration Project. *www.cn.stir.ac.uk/ComputerVision/Projects/ModIP/index.html*.

Nagel, H.-H. (1987). On the estimation of optic flow: Relations between different approaches and some new results. *Artificial Intelligence*, 33:299–324.

Pack, C. C. and Born, R. T. (2001). Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature*, 409:1040–1042.

Parker, A. and Cumming, B. (2001). Cortical mechanisms of binocular stereoscopic vision. *Prog Brain Res*, 134:205–16.

Phillips, W. and Singer, W. (1997). In search of common foundations for cortical processing. *Behavioral and Brain Sciences*, 20(4):657–682.

Pugeault, N. and Krüger, N. (2003). Multi–modal matching applied to stereo. *Proceedings of the BMVC 2003*.

Shevelev, I., Lazareva, N., Tikhomirov, A., and Sharev, G. (1995). Sensitivity to cross–like figures in the cat striate neurons. *Neuroscience*, 61:965–973.

Stumpf, P. (1911). über die Abhängigkeit der visuellen Bewegungsrichtung und negativen Nachbildes von den Reizvorgangen auf der Netzhaut. *Zeitschrift fur Psychologie*, 59:321–330.

Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science*, 262:685–688.

Thiele, A., Dobkins, K., and Albright, T. (1999). The contribution of color to motion processing in macaque area mt. *J. Neurosci.*, 19:6571–6587.

von der Heydt, R., Peterhans, E., and Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science*, 224:1260–62.

Wurtz, R. and Kandel, E. (2000a). Central visual pathways. In Kandell, E., Schwartz, J., and Messel, T., editors, *Principles of Neural Science (4th edition)*, pages 523–547.

Wurtz, R. and Kandel, E. (2000b). Perception of motion, depth and form. In Kandell, E., Schwartz, J., and Messel, T., editors, *Principles of Neural Science (4th edition)*, pages 548–571.

AISB Journal **1**(5), © SSAISB, 2004
http://www.aisb.org.uk

A Computational Model for Contrast Detection in Images: Iterative Tuning and Cross-Orientation Inhibition of Simple Cells in V1

Marina Kolesnik, Alexander Barlit

Fraunhofer Institute for Media Communication, Schloss Birlinghoven,
D-53754 Sankt-Augustin, Germany
marina.kolesnik@imk.fraunhofer.de

Abstract

The orientation selectivity of simple cells in visual cortex gives a striking example of the biological system perfectly adapted to the perception of oriented stimuli. Several models have employed major principles of orientation selectivity for the processing of contrast variations in images. We have recently suggested a model for iterative orientation tuning, in which the astonishingly regular layout of simple cells is explicitly involved in the processing of oriented stimuli. In this work we extended the iterative model by incorporation a mechanism of cross-oriented inhibition. We then investigated the two models using synthetic, noisy and natural images. We found that the two models account for a large fraction of the contrast invariance of orientation selectivity – another striking aspect of the behaviour of simple cells. Our results indicate that the iterative processing of visual stimuli combined with local amplification of proximate simple cells is responsible for ~75% of the contrast invariance. Contrary to some earlier studies, the cross-oriented inhibition did not have any significant contribution to the contrast invariance but accelerated the convergence of the iterative processing on a stable solution. When probed with different images, the new model with cross-oriented inhibition generated a clear pattern of object contours.

1 Introduction

Edge detection is a cornerstone processing stage in the analysis of visual information by humans. This has triggered the development of numerous algorithms for detection of local luminance changes in images. The most popular edge detectors include Marr-Hildreth zero-crossings [1], Canny [2], Haralick [3], Deriche [4] approaches, and a full list of suggested algorithms would go on. From the sheer number of suggested edge detectors, one may conclude that the detection of contours of objects in images is not an easy task for a computer. The fact that a person can effortlessly find the contours of objects has inspired the investigation, and modelling of a contrast detection circuitry in mammalian visual system. This work attempts at developing a processing algorithm for

contrast detection in images built upon principles of the physiological orientation selectivity in mammals.

Forty years ago, Hubel and Wiesel ([5], 1962) discovered that *simple cells* in cat primary visual cortex (V1) are tuned for the orientation of light/dark borders. The inputs to V1 come from the lateral geniculate nucleus (LGN), whose cells are not significantly orientation selective [6]. LGN cells themselves get their input from the retinal ON and OFF *ganglion cells* with centre-surround receptive fields (RFs), first discovered by Kuffler ([7], 1953). The orientation selectivity of simple cells in V1, as proposed by Hubel and Wiesel [5], derives from an oriented arrangement of input from the LGN: ON-centre LGN inputs have RFs centres aligned over simple cells ON subregions, and similarly for OFF-centre inputs. Because of this input arrangement, simple cells perform a linear spatial summation of light intensity in their fields and have an elongated shape of their RFs. The orientation preference of simple cells is quite narrow, and turning a bar-stimulus through more than about 20° from the preferred orientation, greatly reduces the cell's firing rate.

A traditional *feed-forward model* of the orientation selectivity performs linear spatial summation of input signals from the LGN followed by a non-linear rectification stage, in which a threshold filters out small inputs evoked by improperly oriented stimuli [8], [9]. Although many aspects of simple cell responses are consistent with this linear model, there also are important violations of linearity. For example, scaling the contrast of a stimulus would identically scale the responses of a linear cell. At high contrasts, however, the responses of simple cells show clear saturation. Such behaviour of the simple cells is referred to as *contrast invariance* of orientation selectivity [10].

Several neuronal models have attempted to address the nonlinearities of simple cell responses by extending the linear model to include a gain control stage [11], [12], [13]. It is suggested the response of a simple cell is governed by *shunting inhibition* - the divisive normalization of the cell activity due interaction with other cells. The shunting inhibition controls the gain of the transformation of the cell's input current to output membrane potential [14]. A followed rectification stage converts the latter into a firing rate of the cell. The models with shunting inhibition predict response saturation because the divisive normalization increases with stimulus contrast. Another class of models, which also exhibit the contrast invariance rely on the amplification of LGN input by recurrent excitation occurring within the cortical column [29], [30], [31], [32]. This amplification is gated selectively by intracortical inhibition and thereby sharpens weak and poorly oriented LGN input. To arrive at these results the models make an assumption that LGN synapses comprise 5%-10% of the total excitatory synapses present in layer 4. However, recordings of visually evoked membrane potential changes in simple cells [33] indicate that the LGN input is responsible for generating approximately 35% to 46% responses of simple cells.

Despite remaining controversy over the details of synaptic mechanisms underlying orientation selectivity, advances in understanding of major principles of its functionality, laid the foundation for the development of computational models for contrast detection in images [15], [16], [17]. The typical architecture of such a model is built upon a simple cell circuit, which is composed of segregated ON- and OFF-data streams interacting via mechanism of opponent inhibition as suggested by Ferster [18].

None of these models, however, explicitly employed the impressive regularity in a spatial layout of simple cells called by Hubel and Wiesel [5], the *functional organisation* of visual cortex. Visual cortex has a distinctive striped appearance in cross-section, caused by the arrangement of cells in layers of different densities and for this reason it is also known as the *striate* cortex. Simple cells that respond more strongly to stimuli in one eye than in the other, and are said to show *ocular dominance*, are

aligned into *ocular dominance stripes*. Moreover, when the orientation preference of cells in the ocular dominance stripes was related to their position, an astonishingly systematic organisation has emerged: the orientation preference changed linearly with position across V1 [19], [20]. After some distance where cells had shown a systematic clockwise stepping of their orientation preferences, the sequence would reverse to anticlockwise. Hubel and Wiesel therefore suggested that orientation-selective cells are organised in columns or "slabs", in which all cells have the same preferred orientation, and that adjacent slabs represent adjacent orientation. Because, furthermore, the orientation slabs tended to be at right angles to ocular dominance stripes, their regular structure was nicknamed as *Icecube* layout.

Since axons and dendrites take up a significant fraction (about 60%) of the cortical volume [21], limitations on the brain size require keeping neuronal processes as short as possible. Evolution was likely to select in favour of developmental rules that produce orientation preference maps which are sufficiently optimised in terms of length of neuronal connections. Numerical simulations relating orientation preference maps to the length of intracortical wiring have shown that the optimised layout is the Icecube if the strength of local connections is Gaussian [22]. Therefore we assume that local interactions of spatially close simple cells within the Icecube ought to be important for their functionality. A first attempt at utilizing the regularity of orientation preference maps for contrast detection in images, has been made in [23]. It is suggested that the processing of visual input undergoes several iterative cycles. The responses of simple cells at different iterations change due to local interactions of proximate cells. The model takes advantage of the regular layout of orientation preference in a very explicit way: each simple cell is sending activation into a regular net of local connections and amplifies the activity of spatially close cells. The model achieves a significant level of contrast invariance of orientation selectivity due to the iterative amplification of cells of similar orientations at retinotopically close positions.

The work here discusses an extension of the iterative model [23] by incorporating a mechanism of orthogonal suppression of spatially close simple cells of V1. The new model accounts for another aspect of the behaviour of simple cells, namely that simple cells are subject to *cross-oriented inhibition*; the responses to an optimally oriented stimulus can be diminished by superimposing an orthogonal stimulus that is ineffective in driving the cell when presented alone [24], [25]. We suggest that a highly systematic wiring of simple cells in neighbouring ocular dominance bands is involved in the transition of inhibiting signals to nearby cells of orthogonal orientation.

We test the performance of both models using a selected set of two-dimensional stimuli as well as noisy and natural images. Comparison of processing results reveals a very similar behavioural pattern. Both models account for a large fraction of the contrast invariance of orientation selectivity. Our results indicate that incorporation of the cross-oriented inhibition does not significantly improve its performance in terms of contrast invariance, rather stabilises the model and accelerates its convergence on equilibrium. The new model is robust to noise and, when probed with natural images, it generates a clear pattern of contours.

2 Model 1: the iterative orientation tuning of simple cells

The first model 1 has been introduced in [23] and is built upon the idea that visual perception is a continuous process of interpretation of incoming visual data. We adopt the view that the brain has no internal representation of the outside world because it is continuously available "out there" for active perception. While the eye is fixating a particular object, the low-level processing of constant visual input may be undergoing several iterative cycles. Every moment when light hits retina it would cause a different neural activity in the underlying visual circuitry, the activity which depends on a level of current neuronal excitation. Consequently, neural responses to the same visual input at different processing cycles would vary.

This iterative approach to the low-level visual processing gains a further meaning when the role of feedback connections is considered. It is logical to suggest that feedback projections, activated at subsequent iterations, would alter cell responses to the visual input, which itself remains constant in time. It is even more likely that the regular layout of simple cells in V1 reinforces responses of simple cells at subsequent iterations. In the model, we assume that the activation of a simple cell amplifies the activity level of proximate cells, so as to tune these neighbouring cells to a local orientation pattern. After several cycles of iterative tuning, the whole system reaches equilibrium and responses of simple cells to visual input stabilise.

A neural circuit of the model for iterative orientation tuning (Fig. 1) consists of two ON- and OFF-pathways interacting via a mechanism of opponent inhibition. Visual input is processed sequentially, first by retina-LGN followed by a simple cell circuit in V1. A key feature of the model is the iterative processing of visual input, which imitates an instance when the eye is fixating a particular object and the processing of still visual input might undergo several iterative cycles. Local intracortical interaction of simple cells changes their responses to the visual input at subsequent iterations. The local interaction of simple cells is governed by their spatial layout. We adopt an *Icecube* model to describe the spatial layout of simple cells. In the process of local interaction the activation of each simple cell causes excitation of close cells in the *Icecube* layout.

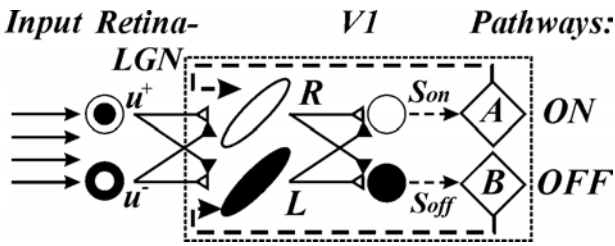


Figure 1. The architecture of the model neural network. The network consists of two major stages - the retina-LGN stage, followed by a simple cell circuit. Triangles at the end of lines denote the excitatory input; filled-in-black triangles denote the inhibitory input. Two dashed lines show the iterative interaction of spatially close simple cells.

The first processing stage deals with responses of retinal ganglion cells with centre-surround receptive fields (RFs) [7]. The retinal ganglion cells are modelled at each spatial position by the difference of input stimulus and its convolution with a 2-dimensional Gaussian kernel (A1), [27]. Retinal ON and OFF ganglion cells synapse mainly onto respective ON and OFF cells of the LGN. In the model, retinal inputs do

not change while passing through the LGN.Simple cells of V1 are driven by oriented input from the LGN. Physiological studies on simple cell responses recorded in cat striate cortex suggest that elongated sensitivity profile of a simple cell subfield is best modelled by a difference of two elongated Gaussians (A2). Each simple ON cell receives excitatory input from the LGN ON cells beneath it and is inhibited by LGN OFF cells at the same retinotopic position (A3).

In addition, simple cells undergo local interaction, which amplifies the activity of cells belonging to a same channel. All simple cells are considered to be stacked into a 3-D array (*Icecube* layout, Fig. 2), in which two coordinates define the spatial (retinotopic) position of the cell and the remaining third coordinate is related to the cell's preferred orientation. Each simple ON cell undergoes additive amplification received from those simple ON cells that are spatially close in the 3-D array; the same ON cell is inhibited by proximate simple OFF cells (A3). The reverse arrangement holds true for simple OFF cells.

Final activation of a simple cell results from the cross-channel inhibition obtained as the steady-state solution of inhibitory shunting interaction (A4), [28].

Because the strength of amplification is an inverse function of squared distance (A6), the activation of proximate cells effectively decays within the distance of 3 units. Due to the weighting factor $\omega = 16$ in (A6), the effect of amplification affects only one neighbouring cell in all 8 directions within the spatial layer, and about 6 neighbouring cells in the orientation column (3 orientations both up and down the column). This local amplification enhances responses of both retinotopically proximate cells and cells tuned to similar orientations.

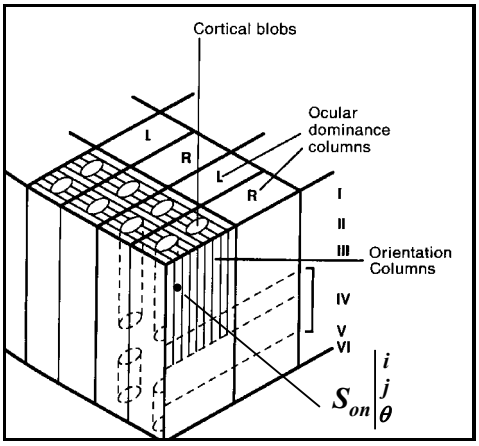


Figure 2. A schematic diagram of the spatial layout of V1 - *Icecube* layout. The primary visual cortex is divided into ocular dominance columns; running perpendicular to these are orientation columns. Orientation preference within columns changes systematically so that each column represents directions from 0° to 180°. The Icecube layout of simple cells is modelled by the 3-D array consisting of spatial layers and orientation columns. Each element of the array, (i,j,θ) has two spatial coordinates, i and j, for position within the layer and one orientation coordinate, θ, for position in the orientation column. It is assumed that each 3-D position contains a pair of ON and OFF cells, S_{on} and S_{off}. This array layout is repeated twice for both contrast polarities $p = 1,2$.

The local amplification changes responses of simple cells to the same visual input over time. This is mediated by the iterative processing of visual input: amplification functions (A6) for the ON and OFF cells are fed into (A3) and the processing cycle is repeated. As the model proceeds through iterations, responses of simple cells would increase. It is however important that the model reaches equilibrium and the amplification of proximate cells stabilises. The corresponding balancing mechanism is provided by the cross-channel interaction (A4), which does not let responses of simple cells to grow indefinitely. However, the cross-channel interaction alone cannot fully prevent a small growth of cell responses cells in the vicinity of sharp luminance changes.

3 Model 2: the iterative tuning with cross-orientation inhibition

Incorporation of the mechanism of cross-oriented inhibition into the model 2 is based on considerable experimental evidence suggesting that stimuli at non-optimal orientations suppress the background activity of simple cells [13], [24], [25]. Clearly, the cross-oriented inhibition has the potential to suppress weak responses in the vicinity of contrastive edges, thus increasing model's robustness to noise. Similar to the local amplification, the cross-oriented inhibition is governed by the spatial layout of simple cells in V1.

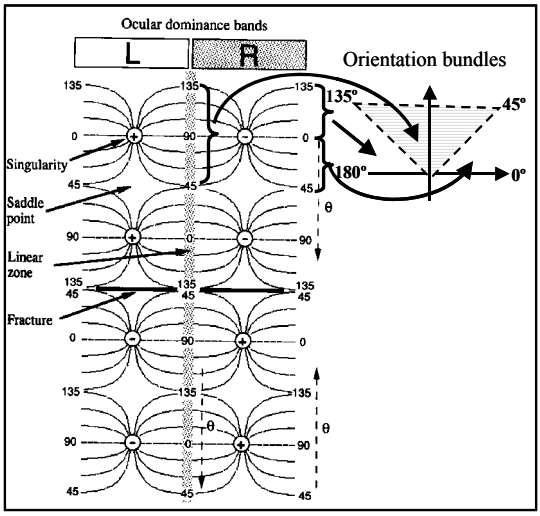


Figure 3. Schematic presentation of relationship between ocular dominance bands and the organisation of orientation selectivity in the visual cortex (Obermayer and Blasdel [26]). Cells along "iso-oriented contour" have the same optimal orientation. Adjacent iso-contour bundles linked at points of singularity range within two complementary sectors of 90° each (shown in black arrows). Due to this arrangement for each cell in a given bundle there exist a counterpart cell of orthogonal orientation belonging to the adjacent bundle.

Optical imaging studies on patterns of activation across a region of monkey cortex [26] revealed a regular structure of "iso-orientation" contours radiating from points of singularity (Fig. 3). Along each of these contours the orientation preference of cell is constant, hence the name - "iso-contours". Cells at the singularities are not orientation selective. Orientations within each bundle of iso-contours range within the interval 90° with adjacent contours representing orientations 11.25° apart. A complete circle around each singularity represents a rotation from 0° to 180°. We suggest that these iso-oriented contours are the links serving the propagation of inhibitory signals to retinotopically close cells of orthogonal orientation.

The activity of each simple cell is inhibited by four cells from the neighbouring orientation columns that are tuned to orthogonal orientation (A5). This mechanism of cross-orientation suppression affects the activity of simple cells in two ways. On one hand, the activity of a simple cell is cancelled out by the activity of retinotopically close cells of orthogonal orientation if these are strongly activated. On the other hand, the response of a strongly activated cell would only be slightly suppressed by retinotopically close cells of orthogonal orientation if these are weakly activated. This cross-oriented inhibition eliminates weak responses of simple cells while sharpening their strong responses.

4 Perception of luminance changes by the two models

We investigated the behaviour of the two models through a set of computer simulations. Each model is probed with several test stimuli, illustrating typical instances of contrast variations in images. All test stimuli are two-dimensional functions. Each model proceeds through 5 iterative cycles before edge responses of simple cells are generated. Edge response S is computed by rectifying the sum of activities of ON and OFF cells minus their difference:

$$S = \left[S_{on} + S_{off} - \left| S_{on} - S_{off} \right| \right]^{+} \tag{1}$$

Each plot, illustrating responses of simple cells, is a one-dimensional slice through stimuli, activity levels of simple cells at selected iterations and edge responses (1). All plots, displaying the activity level of the ON and OFF cells for the model 1 (see Fig. 4a, 5a, 10a), share a common feature: regions of strong activity levels spread onto nearby areas at later iterations. The spreading is less pronounced for the model 2 (see Fig. 4c, 5c, 10b) because the cross-oriented suppression (A5) wipes out the low-level activity at tale regions of the Gaussian-shaped response of ON and OFF cells.

Ramp transition. We conducted two series of simulations investigating the strength of responses depending on the transition range and width. On average, responses of the ON and OFF cells generated by the model 2 grow less with iterations than responses generated by the model 1 (Fig. 4 a, c; and Fig. 5 a, c). This behavioural difference occurs due to the orthogonal suppression. However, both models exhibit almost identical edge response (1) to the ramp transition regardless of its range (Fig. 4, b, d and Fig.5 b, d). Our investigation shows that the perception of ramp profile by the two models depends largely on the profile's width and much less on the range of ramp

transition. The strength of edge responses decreases approximately linearly with the increasing width of the ramp transition (Fig. 6). When the ramp width exceeds 28 pixels, responses of simple cells decay completely even though the transition range is high. This behaviour is similar for both models although the trend is quicker for the model 2: it becomes insensitive to the transition ramp wider than 22 pixels.

The dependency of responses on the ramp range is strongly non-linear for both models. Fig. 7 shows, that the increase of the ramp range by a factor of 9 causes about 25% growth in the strength of respective edge response. This non-linear dependency is aggregated during the iterative processing due to advantageous enhancement of initially weaker responses. This result supports a great part of the contrast invariance of orientation selectivity observed experimentally.

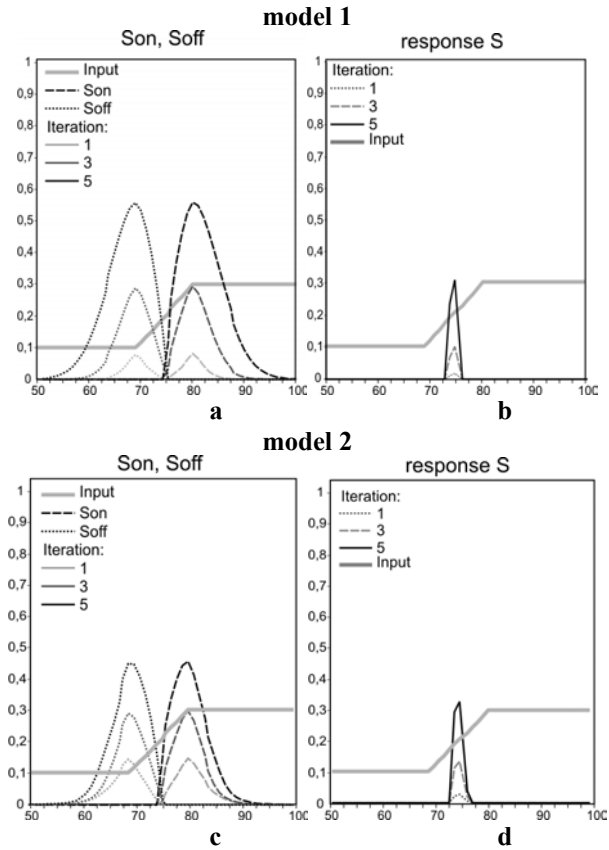


Figure 4. Responses to the ramp transition generated by the two models. Ramp range is equal to 0.2. The responses of ON and OFF cells, S_{on} and S_{off}, generated by the model 1 (**a**), grow stronger and spread wider with iterations as compared to the corresponding responses generated by the model 2 (**c**). Due to a steep fall off in the ON and OFF response for model 2 (**c**), it generates as strong edge response at 5th iteration (**d**) as that one of model 1 (**b**).

Further analysis of curves in Fig. 7 shows that a subtle difference in the behaviour of two models appears at later iterations. Edge responses generated by the model 2 stabilise at 6th iteration. The convergence rate is slower for the model 1. The explanation for this comes from the analysis of mechanisms stabilising the two models. There exist two such mechanisms. The first one, the cross-channel inhibition (A4), is common for both models. The second one, the cross-oriented suppression (A5) which is only present in the model 2, imposes an additional constraint on the propagation of excitation onto cells which do not receive salient oriented excitation from the visual input. Also, the cross-oriented suppression accelerates the convergence of the iterative processing of model 2 when compared to the model 1.

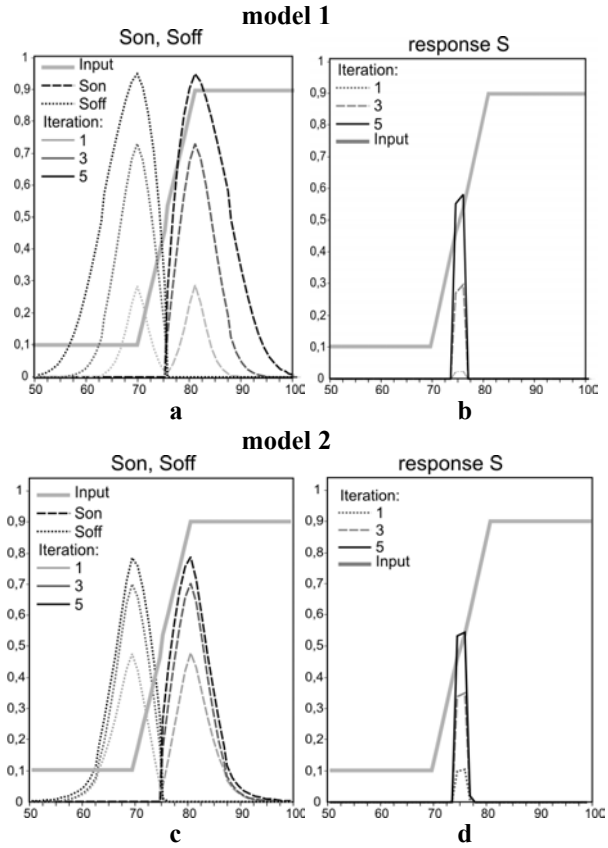


Figure 5. Responses to the ramp profile of range 0.8. Edge responses display similar tendencies in the shape and rate of growth as responses to the ramp transition of 0.2 (Fig. 4). Also the strength of edge response at 5th iteration of model 2 (**d**) is pretty the same as that one of model 1 (**b**).

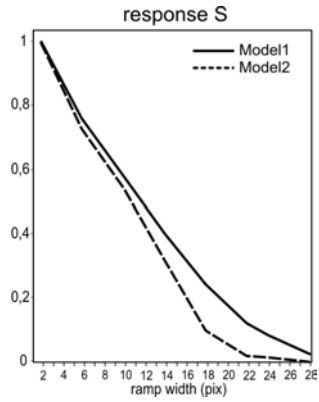


Figure 6. Edge response to ramp profiles of different widths. The strength of edge responses generated by the two models is plotted against the width of ramp transition with range 0.8. For both models edge response drops almost linearly with the increase of the ramp width. The fall-off is quicker for model 2, which does not perceive the luminance ramp above 22 pixels width.

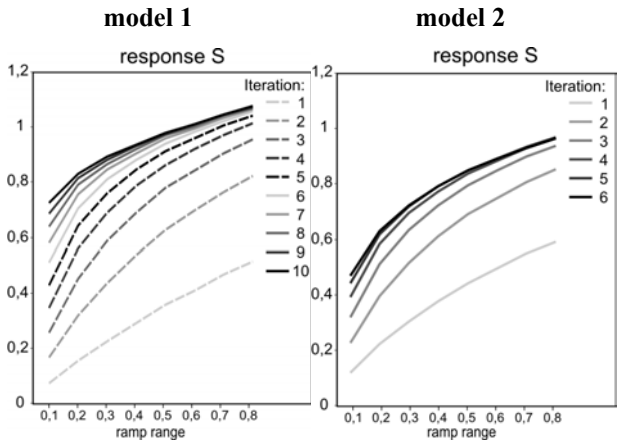


Figure 7. Convergence of the models on a stable solution occurs quicker for stronger responses. Whereas responses to weaker stimuli continue to grow at later iterations, strong responses do not rise any more. This feature, common for both models, is responsible for the partial contrast invariance of orientation selectivity.

Bar profile. Both models capture a large narrow variation in brightness in the form of two sharp responses (Fig. 8). The responses are associated with two sides of the bar profile, which are perceived as “edges”. The same double response pattern is induced by a narrowest possible bar profile with the width of 1 pixel. Note that the iterative amplification of edge responses by the model 2 is stronger than that one of the model 1.

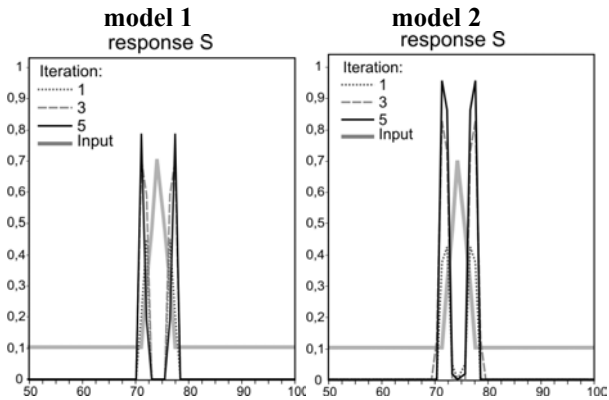


Figure 8**.** Responses to the bar profile generated by the two models at 1st, 3rd, and 5th iterations. Bar width is equal to 3 pixels.

Grating. Simulation of responses induced by a grating composed of four equal contrast bars produces eight sharp responses at "edge" positions, each one associated with particular bar side (Fig. 9). However, boundary responses evoked by the two external sides of the grating undergo stronger amplification at iterations than responses to any of the internal sides of grating bars. It seems that more isolated responses tend to override nearby responses of smaller or comparable magnitude. This behavioural aspect is particularly useful for the processing of noisy images. Weak responses to spontaneous luminance variations caused by noise get eliminated after several iterations. The elimination process is especially efficient in the vicinity of strong luminance changes.

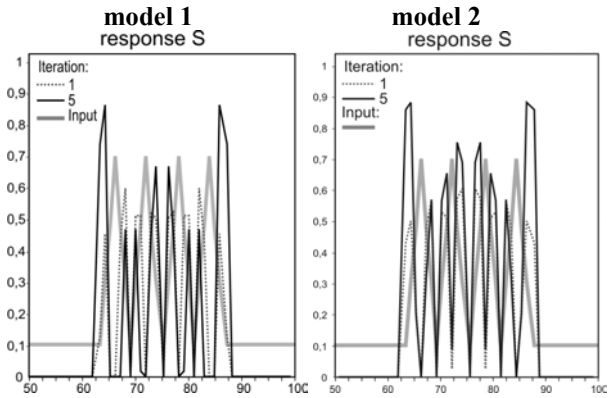


Figure 9**.** Responses to the grating generated by the two models at 1st, and 5th iterations. The width of bars in the grating is 3 pixels. The iterative processing advantageously amplifies responses to the external sides of the grating followed by the amplification of two responses induced by the two sides in the middle. Note that responses, neighbouring to the external bars on each side of the grating are not amplified.

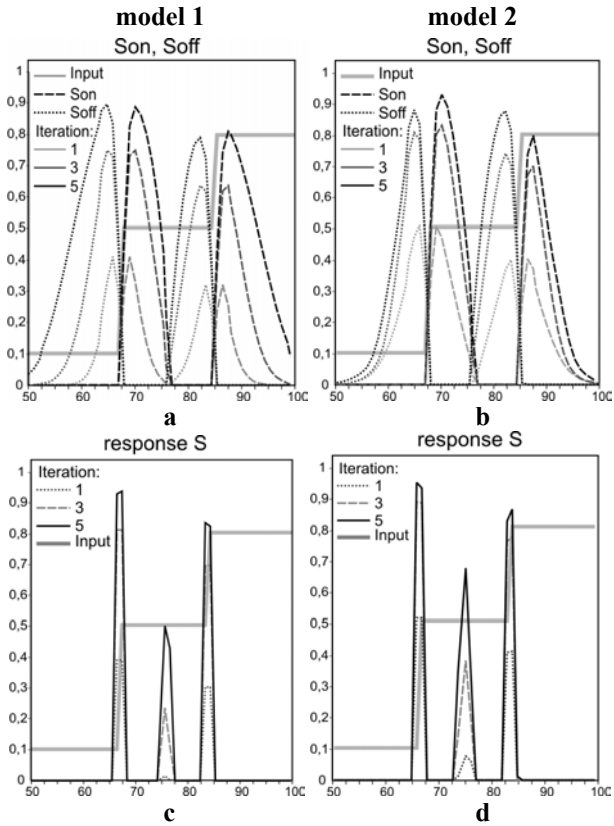


Figure 10. Responses to the staircase profile generated by the two models at 1st, and 5th iterations. The models perceive the illusory line in the middle of the step. Due to a sharper shape of the ON and OFF responses generated by model 2 (b), the perception of illusory line by the model 2 is stronger.

Staircase profile. The processing of a luminance staircase reveals that both models perceive an illusory line right in the middle of step's plateau. The origin of this curious phenomena becomes clear when we analyse the shape of responses of the ON and OFF cells (Fig. 10, a, b). The OFF cell response to the abrupt luminance change between two subsequent steps has overlay with the ON cell response evoked by the adjacent luminance change in the staircase. Computation of edge responses (1) results into a sharp edge response in the middle of the step's plateau (Fig. 10, c, d). The illusory line vanishes when the step's plateau width exceeds 26 pixels. Perception of the illusory line might be related to the well known Chevreul illusion, in which a regular luminance staircase is perceived as not perfectly uniform along the luminance plateaus.

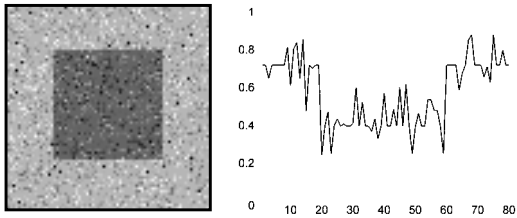


Figure 11. Noisy input image (left) and the cross section taken at the centre of the image (right).

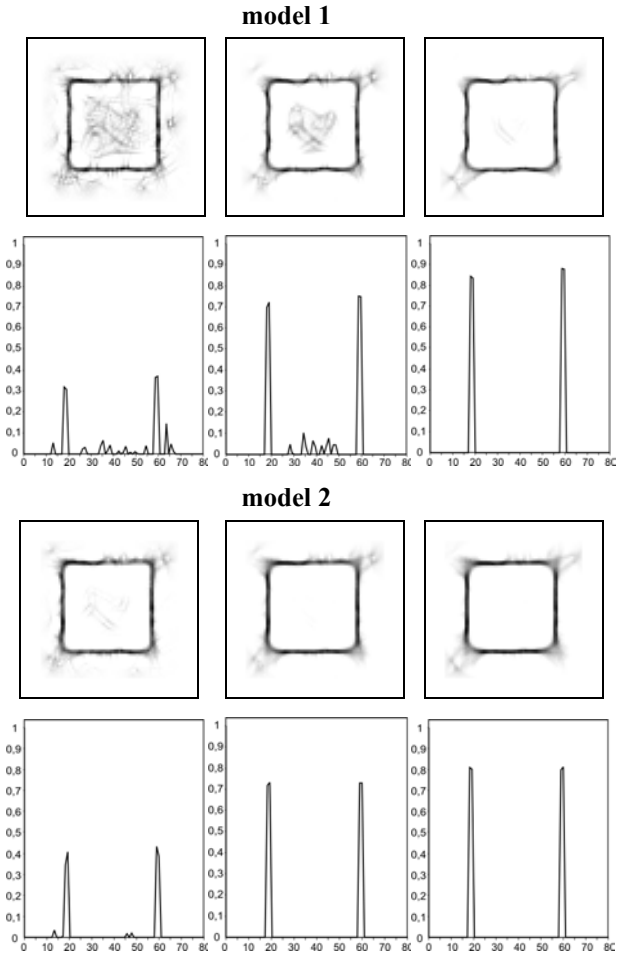


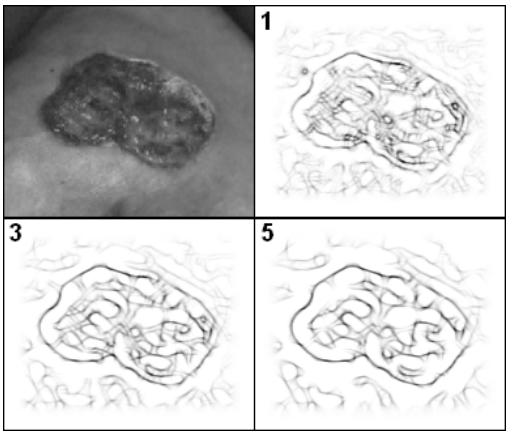
Figure 12. Edge responses to the noisy image (Fig.11) after 1st, 3rd, and 5th iteration generated by the model 1 (two upper rows) and model 2 (two bottom rows). Images of edge responses are inverted. Corresponding cross sections are taken at the centre of the images. Responses to noise weaken noticeably as the models proceed through iterative cycles. The cross-oriented inhibition introduces an additional mechanism of noise suppression for the model 2 due to suppression of the activity of cells at non-optimal orientations.

Noisy input. The processing of a synthetic image of a dark rectangle on a lighter background corrupted with 50% Gaussian noise (Fig. 11) exhibits that responses to weak contrast variations caused by noise are significantly diminished after several iterations (Fig. 12). Noise reduction is especially pronounced in the vicinity of rectangle edges, where small responses are "cancelled out" by stronger responses which spread in the process of iterative tuning.

5 Processing of natural images

Edge responses to natural images clearly illustrate a common feature exhibited by the two models, namely the enhancement of isolated weak edges at subsequent iterative cycles (Fig. 13). The reason for this behaviour is a non-linear normalization of responses due to the cross-channel inhibition: the divisive normalisation tends to boost weaker responses over the stronger ones while normalising an overall magnitude of cell responses to the range [0,1].

model 1



model 2

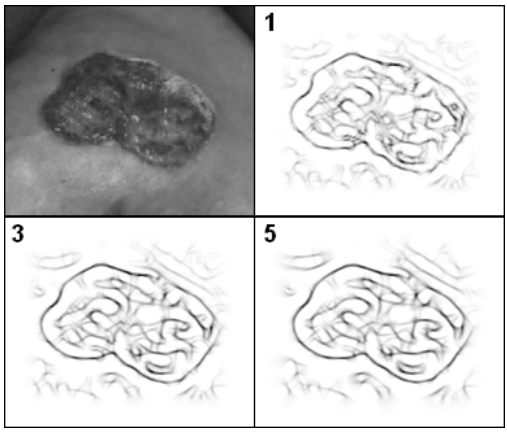


Figure 13**.** Input image of a wound and edge responses (inverted) generated by the two models at 1st, 3rd, and 5th iteration (clockwise from top left). Weak edges disappear and strong edges are amplified as the models proceed through iterative cycles. The iterative tuning diminishes spurious responses both for skin and the wound region. Edge enhancement is more pronounced for the model 2.

6 Summary and conclusions

The motivation of this work is the development of a biologically justified model for contrast detection in images, the model, which has the potential to outperform purely computational approaches to edge detection. The orientation selectivity of simple cells in V1 gives us an exciting example of a biological system, which is capable of responding to oriented visual stimuli with great efficiency. We have proposed a model for the iterative orientation tuning with cross-oriented suppression. The model is an extension of the iterative orientation tuning model introduced in [23]. We have compared the performance of the two models by probing them with synthetic stimuli, synthetic noisy images and natural images.

The two major and common features for both models are the iterative processing of visual input and the local intracortical amplification of proximate cells belonging to a same channel. The local amplification explicitly exploits the morphology of simple cells in V1. The contribution of local amplification to the responses of simple cells grows with iterations. This iterative amplification greatly enhances responses of both retinotopically proximate cells and cells tuned to similar orientations. Consequently, the local amplification activates a process of selective orientation tuning enhancing responses of cells of "proper" orientations at retinotopically close positions. The cross-orientation inhibition, incorporated in the model 2, is the mechanism of selective suppression, which affects locally the activity of cells receiving stimuli with no distinguished orientation.

Three processes play a major role in the generation of the contrast invariance of orientation selectivity: 1) the iterative processing of video input, 2) the local amplification, and 3) the cross-channel inhibition of activities of simple cells. These processes are responsible for a very similar behavioural pattern exhibited by the two models. We note that the incorporation of cross-oriented suppression into the model 2 only slightly improves the model's performance in terms of contrast invariance, which, on average, remains at a level of 75%. The orthogonal suppression does not seem to play a crucial role in the generation of contrast invariance: it does not significantly affect the magnitude of edge responses. Our investigation indicates that a major contribution to the contrast invariance comes from the local intracortical amplification of responses at subsequent iterative cycles. The cross-oriented inhibition did not account for a significant part of the contrast invariance in our simulations. This conclusion contradicts earlier suggestions that intracortical inhibition tuned to the orthogonal orientation plays a major role in the generation of cortical orientation selectivity [13], [24], [25]. However, our simulations do suggest that the cross-oriented suppression sharpens responses of simple cells. It appears that this sharpening accelerates the convergence of edge responses on a stable solution.

We conclude that the cross-oriented inhibition introduces an important stabilising factor into the process of orientation tuning, but cannot preserve the contrast invariance of orientation selectivity. Additional mechanisms may therefore be involved in the generation of the contrast invariance observed experimentally in monkeys and cats.

Although neither model includes any additional mechanism for the suppression of noise, both of them have demonstrated high robustness to noisy input. It seems that a good resistance to noise is an inherent feature of the functionality of simple cells taken over by the models for free.

One final conclusion of this study is as follows: however efficient the functionality of simple cells is, neither the iterative amplification of activities of simple cells nor the cross-oriented suppression can provide a selective extraction of object edges. It seems that additional mechanisms such as object recognition linked with memory association feedback should play a decisive role in the selective extraction of object contours.

Appendix

The processing of visual input undergoes several iterative cycles, each one containing either four (model 1) or five (model 2) subsequent stages. Stages #1, #2, #3, and #5 are common for both models. Stage #4 stands for the model 2. Below we list all processing stages noting explicitly the differences between the two models when these are present**.**

1. Retina-LGN. Responses of retinal ganglion ON and OFF cells, u_{ij}^{+} and u_{ij}^{-}, are given by:

$$\begin{aligned} X_{ij} &= I_{ij} - G_{\sigma} * I_{ij}; \\ u_{ij}^{+} &= \left[X_{ij}\right]^{+}; \quad u_{ij}^{-} = \left[-X_{ij}\right]^{+} \end{aligned} \tag{A1}$$

where $*$ is the spatial convolution operator and $[x]^{+} := max\{x,0\}$ denotes half-wave-rectification. The G_{σ} is a centre Gaussian with standard deviation σ=3 for the model 1 and σ=5 for the model 2. The Gaussian is sampled within a filter mask of 35x35 and 45x45 pixels for the model 1 and model 2, respectively. Visual input I is normalised to the range [0,1].

2. Simple cell subfields. Simple cells are modelled for twelve discrete orientations θ = 0°, 15°,…, 165°, and two opposite contrast polarities $p=1, 2$:

$$\begin{aligned} D_{\theta\sigma_M\sigma_m\boldsymbol{\tau}} &= G_{\theta\sigma_M\sigma_m\boldsymbol{\tau}} - G_{\theta\sigma_M\sigma_m-\boldsymbol{\tau}}\ , \\ G_{\theta\sigma_M\sigma_m\boldsymbol{\tau}} &= \frac{1}{2\pi}\exp\left[-\frac{1}{2}\left((\mathbf{x}-\boldsymbol{\tau})^T\mathbf{R}^T\mathbf{C}\mathbf{R}(\mathbf{x}-\boldsymbol{\tau})\right)\right] \\ \mathbf{C} &= \begin{pmatrix} 1/\sigma_m^2 & 0 \\ 0 & 1/\sigma_M^2 \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \end{aligned} \tag{A2}$$

where $\mathbf{x}^T = (i, j)$ denotes the position *(i,j)*, $\boldsymbol{\tau}^T = (\cos\theta, \sin\theta)$ is relative offset for two Gaussian lobes from their central position *(i,j)*, and the space constants $\sigma_m=1$ and $\sigma_M= 4$ define the degree of filter's elongation. The filter mask of the simple cell subfield is 19x19 pixels.

At each position, *(i,j)*, and for each orientation, θ, and polarity, p, the model has an even symmetric simple cell with two parallel elongated parts: an ON subfield, $R_{i,j,\theta,p}$, which receives excitation from LGN ON cells beneath it, u_{ij}^{+}, and is inhibited by input from the LGN OFF cells at the same position, u_{ij}^{-}; and an OFF subfield, $L_{i,j,\theta,p}$, for which the reverse relation to the LGN channels holds true. This physiology is embodied in the equation for the ON subfield by subtracting the half-wave rectified LGN OFF channel, u^{-}, from the rectified ON channel, u^{+}, and convolving the result with the positive lob of the oriented filter, $[\,D_{\theta\sigma_M\sigma_m\boldsymbol{\tau}}\,]^{+}$, [27]. The OFF subfield, $L_{i,j,\theta,p}$, is constructed similarly.

In addition, each ON subfield, $R_{i,j,\theta,p}$, receives excitatory input, $A_{i,j,\theta,p}$, from all simple ON cells that are spatially close to position *(i,j)* in the *Icecube* layout, and is inhibited by input $B_{i,j,\theta,p}$ from all close OFF cells. The reverse arrangement holds true for the computation of the activation level of each OFF subfield, $L_{i,j,\theta,p}$. The mutual amplification-inhibition of neighbouring cells is a time varying function which is updated iteratively. The above considerations give rise to the following expressions for $R^n_{i,j,\theta,p}$ and, $L^n_{i,j,\theta,p}$, at iteration *n:*

$$\begin{aligned} R^n_{i,j,\theta,p} &= \left[\left(u^+_{ij} + A^n_{i,j,\theta,p} - u^-_{ij} - B^n_{i,j,\theta,p} \right) \ * \ \left[D^p_{\theta\sigma_M\sigma_m\tau} \right]^+ \right]^+ \\ L^n_{i,j,\theta,p} &= \left[\left(u^-_{ij} + B^n_{i,j,\theta,p} - u^+_{ij} - A^n_{i,j,\theta,p} \right) \ * \ \left[- D^p_{\theta\sigma_M\sigma_m\tau} \right]^+ \right]^+ \end{aligned} \tag{A3}$$

The strength of local interaction for both models, $A_{\theta,p}, B_{\theta,p}$, that we call amplification functions, varies over time. The values of amplification functions at initial iteration *n=0*, are set to $A_{\theta,p} = B_{\theta,p} = 0$, for all orientations and polarities. Note, that due to the offset of the positive and negative lobes of $D_{\theta\sigma_M\sigma_m\tau}$, subfield responses are shifted from their central positions. To compensate, both subfields, $R_{i,j,\theta,p}$ and $L_{i,j,\theta,p}$, are shifted in the opposite directions, τ and $-\tau$, respectively.

3. Cross-channel inhibition. The activation of simple ON cell, S^n_{on}, at iteration *n,* is obtained as the steady-state solution of inhibitory shunting interaction:

$$S^n_{on} = \left[(R^n - L^n) / (1 + R^n + L^n) \right]^+ \tag{A4}$$

Here variables occur for all positions, orientations and polarities; indexes *i, j, θ,* and *p* are omitted to simplify notations. Activation of simple OFF cell is obtained by interchanging R^n and L^n.

4. Cross-oriented suppression.

<u>*Model 2:*</u> Simple cells are engaged in the cross-orientation inhibition, so that the cells' activity at position *(i,j,θ),* is inhibited by four neighbouring cells in the 3-D array (see caption to Fig.2) that are tuned for the orthogonal direction*:*

$$\begin{aligned} &\hat{S}^n_{i,j,\theta} = \left[S^n_{i,j,\vartheta} - (S^n_{i+1,j,\vartheta} + S^n_{i-1,j,\vartheta} + S^n_{i,j+1,\vartheta} + S^n_{i,j-1,\vartheta}) / 4 \right]^+, \\ &\begin{cases} \vartheta = \theta + \pi/2 & if \ \ 0 \le \theta < \pi/2 \\ \vartheta = \theta - \pi/2 & if \ \ \pi/2 \le \theta < \pi \end{cases} \end{aligned} \tag{A5}$$

5. Local amplification. At each at position *(i,j,θ)* in the 3-D array, the excitatory input, $A^n_{i,j,\theta}$, from proximate ON cells, is an inverse function of squared distance:

$$\underline{Model\ 1}:$$

$$A^n_{i,j,\theta} = \mu \sum_{l,m,\vartheta} \frac{S^n_{(on),l,m,\vartheta}}{D^2\left[(l,m,\vartheta),(i,j,\theta)\right]} \tag{A6}$$

$$\underline{Model\ 2}:$$

$$A^n_{i,j,\theta} = \mu \sum_{l,m,\vartheta} \frac{\hat{S}^n_{(on),l,m,\vartheta}}{D^2\left[(l,m,\vartheta),(i,j,\theta)\right]}$$

$$D^2\left[(l,m,k),(i,j,\theta)\right] = \omega\left((l-i)^2 + (m-j)^2\right) + (\theta - \vartheta)^2$$

where μ - is a scaling factor set to: $\mu = 0.18$, and ω - is a weighting factor set to $\omega = 16$. Above computations are repeated twice for both polarities. Excitatory input B to an OFF-cell is obtained by substituting S_{off} for S_{on}.

References

1. Marr, D., Hildreth, E. Theory of edge detection. *Proceedings of the Royal Society*, B 207: 187-217, 1980.
2. Canny, J., F. Finding edges and lines in images. *Technical Report AI-TR-720*, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, June 1983.
3. Haralick, R., M. Digital step edges from zero crossing of second directional derivatives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(1): 58-68, 1984.
4. Deriche, R. Using Canny's criteria to derive an optimal edge detector recursively implemented. *The International Journal of Computer Vision*, 2: 167-187, April 1987.
5. Hubel, D., H., Wiesel, T., N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Psychology*, 160:106-154, 1962.
6. Hubel, D., H., Wiesel, T., N.: Integrative action in the cat's lateral geniculate body. *Journal of Psychology*, 155: 385-398, 1961.
7. Kuffler, S., W.: Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16: 37-68, 1953.
8. Movshon, J., A., Thompson, I., D., Tolhurst, D., J. Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *Journal of Physiology,* (London), 283: 53-77, 1978.
9. Carandini, M., Ferster, D. A tonic hyperpolarization underlying contrast adaptation in the cat visual cortex. *Science*, 276:949-952. 1997.
10. Sclar, G., Freeman, R.: Orientation selectivity in the cat's striate cortex is invariant with stimulus contrast. *Experimental Brain Research*, 46: 457-461, 1982.
11. Albrecht, D., G., Geisler, W., S. Motion sensitivity and the contrast response function of simple cells in the visual cortex. Visual Neuroscience. 7:531-546.

12. Heeger, D., J. Nonlinear model of neural responses in cat visual cortex. In M. Landy, J. A. Movshon (eds.): Computational models of visual processing. Cambridge, MIT: 119-133, 1991.

13. DeAngelis G., C., Robson, J., G., Ohzawa, I., Freeman, R., D. The organization of suppression in receptive fields of neurons in the cat's visual cortex. *J. Neurophysiology*. 68:144-163, 1992.

14. Carandini, M., Heeger, D., J. Summation and division by neurons in visual cortex. *Science*, 264:1333-1336, 1994.

15. Pessoa, L., Mingolla, E., Neumann, H.: A contrast- and luminance-driven multiscale network model of brightness perception. *Vision Research*, 35:2201-2223, 1995.

16. Neumann, H., Pessoa, L., Hansen, Th.: Interaction of ON and OFF pathways for visual contrast measurement. *Biological Cybernetics*, 81:515-532, 1999.

17. Hansen, Th., Neumann, H. A model of V1 visual contrast processing utilizing long-range connections and recurrent interactions. In Proc. of the International Conference on *Artificial Neural Networks*, Edinburgh, UK, September 7-10: 61-66, 1999.

18. Ferster, D.: The synaptic inputs to simple cells in the cat visual cortex. In: D. Lam and G. Gilbert (eds.): Neural mechanisms of visual perception, Ch. 3, Portfolio Publ. Co, The Woodlands, Texas: 63-85, 1989.

19. Hubel, D., H., Wiesel, T., N.: Sequence regularity and geometry of orientation columns in the monkey striate cortex. Journal of Comparative Neurology, 158, (1974) 267-294.

20. Hubel, D., H., Wiesel, T., N.: Functional architecture of macaque monkey visual cortex. Proceedings of the Royal Cosiety of London, B, 198, (1977) 1-59.

21. Braitenberg, V., Schüz, A. Cortex: Statistics and Geometry of Neuronal Connectivity. Berlin: Springer-Verlag, 1998.

22. Koulakov, A., A., Chklovskii, D., B. Orientation Preference Patterns in Mammalian Visual Cortex: A Wire Length Minimization Approach. *Neuron*, 29:519-527, 2001.

23. Kolesnik, M., Barlit, A., Zubkov, E. Iterative Tuning of Simple Cells for Contrast Invariant Edge Enhancement. Proc. of the 2nd International Workshop on *Biologically Motivated Computer Vision* (BMCV'2002): 27-37, 2002.

24. Morrone, M., C., Burr, D., C., Maffei, L. Functional implications of cross-orientation inhibition of cortical visual cell. 1. Neurophysiological evidence. Proc. Royal Society London [Biol.], 216:335-354, 1982.

25. Bonds, A., B. Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Visual Neuroscience*, 2:41-55, 1989.

26. Obermayer, K., Blasdel, G., G. Geometry of orientation and ocular dominance columns in monkey striate cortex. *Journal of Neuroscience*. 13:4114-4129. 1993.

27. Grossberg, S., Raizada, R., D., S.: Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. CAS/CNS TR-99-008, Boston University: 1-35, 1999.

28. Borg-Graham, L., J., Monier, C., Fregnac, Y.: Visual input invokes transient and strong shunting inhibition in visual cortical neurons. Nature, 393, (1998) 369-373.

29. Douglas, R.,J., Koch, C., Mahowald, M., Martin, K.,A.,C., Suarez, H.,H. Recurrent excitation in neurocortical circuits. *Science*, 269: 981-985, 1995.

30. Ben-Yishai, R., Bar-Or, R., L., Sampolinsky, H. Theory of orientation tuning in visual cortex. *Proc. National Academy of Science,* USA92: 3844-3848, 1995.

31. Somers, D., C., Nelson, S., B., Sur, M. An emergent model of orientation selectivity in cat visual cortical simple cells. Jo*urnal of Neuroscience.* 15: 5448-5465, 1995.

32. Maex, R., Orban, G., A. Model circuit of spiking neurons generating directional selectivity in simple cells. *Journal of Neurophysiology.* 75: 1515-1545, 1996.

33. Chung, S., Ferster, D. Strength and orientation Tuning of the Thalamic input to Simple Cells Revealed by Electrically Evoked Cortical Suppression. *Neuron.*20: 1177-1189, 1998.

AISB Journal **1**(5), © SSAISB, 2004
http://www.aisb.org.uk

Adaptive Agents and Multi-Agent Systems

Dimitar Kazakov*, Daniel Kudenko* and Eduardo Alonso†

* Department of Computer Science, University of York, Heslington, York YO10 5DD, *kazakov@cs.york.ac.uk ; kudenko@cs.york.ac.uk*
† Department of Computing, City University, London, Northampton Square, London EC1V 0HB, *eduardo@soi.city.ac.uk*

Editor's Introduction

We met the Year 2001 – magical milestone to the future – without being surrounded by either Arthur C. Clark's intelligent computers or their moody cousins of Douglas Adams's cut. We wish to believe though that one of the many steps needed in this direction was made when the First Symposium on Adaptive Agents and Multi-Agent Systems (AAMAS) was organised in this year. The past two years have seen an increasing interest, and the beginning of consolidation of the European research community in the field. The first book on the eponymous subject, largely based on contributions to AAMAS and AAMAS-2 has now been published.

This volume contains two articles from the Third AAMAS Symposium, which persisted in the goals set in 2001, namely, to increase awareness and interest in adaptive agent research, encourage collaboration between machine learning and agent system experts, and give a representative overview of current research in the area of adaptive agents. The paper by Kapetanakis *et al.* focuses on the learning of co-ordination for two-player co-operative single-stage games. The authors propose a method that enables agents to learn to co-ordinate without explicit communication and without the need to observe each other's actions. The second paper by Strens investigates a search technique for optimal agent behaviour with partially observable states (and specifically a hidden goal state). The algorithm is based on particle filtering and enables the agents to evaluate a policy against all possible hidden states at the same time. The system is evaluated in a hunter-prey scenario, where the hunters do not know the position of the prey.

AISB Journal **1**(5), © SSAISB, 2004
http://www.aisb.org.uk

Learning to coordinate using commitment sequences in cooperative multi-agent systems

Spiros Kapetanakis*, Daniel Kudenko* and Malcolm Strens†

* Department of Computer Science, University of York, Heslington, York, YO10 5DD, England, *spiros@cs.york.ac.uk ; kudenko@cs.york.ac.uk*
† Future Systems Technology Division, QinetiQ, Ively Road, Farnborough, Hampshire, GU14 0LX, England, *mjstrens@qinetiq.com*

Abstract

We report on an investigation of the learning of coordination in cooperative multi-agent systems. Specifically, we study solutions that are applicable to *independent* agents, i.e., agents that do not observe one another's actions and do not explicitly communicate with each other. In previously published work (Kapetanakis and Kudenko, 2002) we have presented a reinforcement learning approach that converges to the optimal joint action even in scenarios with high miscoordination costs. However, this approach failed in fully stochastic environments. In this paper, we present a novel approach based on reward estimation with a shared action-selection protocol. The new technique is applicable in fully stochastic environments where mutual observation of actions is not possible. We demonstrate empirically that our approach causes the agents to converge almost always to the optimal joint action even in difficult stochastic scenarios with high miscoordination penalties.

1 Introduction

Learning to coordinate in cooperative multi-agent systems is a central and widely studied problem, see, for example (Lauer and Riedmiller, 2000; Boutilier, 1999; Claus and Boutilier, 1998; Sen et al., 1994; Weiss, 1993; Nowé et al., 2001). In this context, coordination is defined as *the ability of two or more agents to jointly reach a consensus over which actions to perform in an environment*. We investigate the case of *independent* agents that cannot observe one another's actions and do not explicitly communicate with each other, which often is a more realistic assumption. This generality distinguishes our approach from alternatives (Wang and Sandholm, 2002; Chalkiadakis and Boutilier, 2003, *inter alia*) that require complete mutual observation of actions.

In this investigation, we focus on scenarios where the agents must learn to coordinate their actions through environmental feedback. In previous research (Kapetanakis and Kudenko, 2002) we have presented a reinforcement learning technique (called FMQ) for independent agents, that converged to the *optimal* joint action in scenarios where miscoordination is associated with high penalties. However, the FMQ approach failed in fully stochastic environments, where the rewards associated with joint actions are non-deterministic.

In this paper, we present a novel technique that is based on a shared action-selection protocol (called *commitment sequence*) that enables the agents to estimate the rewards for

specific joint actions. We evaluate this approach experimentally on a number of stochastic versions of two especially difficult coordination problems that were first introduced in 1998 (Claus and Boutilier, 1998): the *climbing game* and the *penalty game*. The empirical results show that the convergence probability to the optimal joint action is very high, in fact reaching almost 100%.

Our paper is structured as follows: we initially illustrate coordination through examples extracted from human, animal and artificial agent societies. We then introduce a common testbed for the study of learning coordination in cooperative multi-agent systems: stochastic cooperative games. We continue to introduce a basic version of the novel commitment sequence technique that uses simple averaging and discuss the experimental results. Finally, we present an extension using Gaussian estimation that improves both the probability of convergence to the optimal joint action and the convergence speed. We finish with an outlook on future work.

2 Coordination without communication: examples

Coordination is often naturally associated with explicit communication. However, there are many examples of agents coordinating without explicitly communicating in animal, human and artificial agent societies.

Take, for example, the coordination required by a group of five lionesses who hunt wildebeest, as reported By Donald Griffin (Griffin, 1984). Two of the five lionesses mount ant hills so as to be clearly visible. These two hunters make themselves seen but stay at a safe distance from the two bands of wildebeest. A third lioness creeps along a ditch parallelling the road until she positions herself between the two herds in a covered position. A fourth lioness charges out of a nearby wooded area to one band of prey driving them across the road towards the other band. The lioness in the ditch has an easy kill to make with wildebeest jumping over the ditch that she occupies. They then apparently all feed on the same prey.

The need for coordination without communication does also arise in human societies. Take, for example, any team sports game such as football, basketball, baseball, cricket, volleyball, or hockey. The existence of two opposing teams that consist of multiple cooperating players makes communication a costly action in the sense that a player is much less likely to (audibly or otherwise) signal to his/her teammates when he/she knows that such signalling normally gives away the intention to the opposition. Communication does occur, of course, when its expected benefit outweighs the disadvantage of opponent discovery. However, what typically happens is that players have this type of untold coordination built in and engage in correctly coordinated behaviour without communicating. So, much in the same sense as the five lionesses, the players can achieve coordination without explicit communication.

Finally, in artificial agent societies, the achievement of coordination is not always a consequence of communicative acts. Instead, it often is the case that agents coordinate their activities without communication. Take the following network routing control system, for instance: in static conditions, the optimal routing tables, i.e., the choice of appropriate neighbour to route a packet that arrives at a network node, can be calculated by various means e.g., ant-based routing (Legge and Baxendale, 2002; Legge and Baxendale, 2003). This provides the system with the optimal way to route packets along the network in static conditions. However, in the presence of congestion and potential downtime for links and nodes, it is essential to be able to calculate new routing tables for the network. This is done by placing a single agent on each node to handle the routing. The

joint decision of all the agents/nodes to change their routing tables so as to alleviate traffic congestion on the network can be seen as a massive coordination step. For each agent, the decision to make every n time steps is which way to route a packet that arrives at this node and is destined for some other node in the network. All the nodes maintain a routing table that determines, for each possible destination on the network, which nearest neighbour node to forward the packet to. For example, if a node is adjacent to nodes A, B and C, its routing table may look like 1:

	A	B	C
A	1	0	0
B	0	1	0
C	0	0	1
D	0	1	0
E	0	1	0
⋮	⋮	⋮	⋮
S	0	0	1

Table 1: Example routing table

If a packet for node S arrives at this node, it is first forwarded to node C. In other words, the decision to make every n timesteps is how to allocate a '1' in any of three places for $m-1$ destination nodes, where m is the total number of nodes in the network.

This problem can be solved by measuring the traffic on the network and rewarding the agents every n timesteps depending on how well the network is working. When there is no congestion, we expect the reward to be positive whereas lots of congestion leads to severe negative rewards or penalties. These nodes are able to communicate with one another but should choose not to do so when each communicative act will only worsen the state of the network. Under such circumstances it is desirable to achieve better routing without explicit communication.

In the rest of this paper, we will not deal with the aforementioned real-world problems, but rather focus on a more abstract formalism to describe coordination problems: stochastic cooperative games.

3 Stochastic cooperative games

Cooperative single-stage games (Fudenberg and Levine, 1998) are a means to illustrate the complex interactions between two or more agents when they learn to coordinate their actions. As such, they provide a common platform on which to evaluate new machine learning algorithms, statistical approaches and coordination protocols.

A single-stage game defines one or more outcomes for every possible joint action the agents may undertake. Single-stage games cannot possibly describe a series of coordination actions in an environment but they can accurately model one interaction step. Finally, these games are unsuitable for problems where coordination must be achieved in complicated state spaces but they are able to model the static dynamics of coordination in one of these states. Therefore, single-stage games are indeed the ideal choice for the study of coordination in static single-step multi-agent problems and a first step towards understanding the true complex dynamics of agent interaction, in which coordination is merely one layer.

In this work, we have concentrated only on cooperative games, i.e., games where the agents are rewarded based on their joint action and all agents receive the same reward. In these games, every agent chooses an action from its action space for every round of the game. These actions are executed simultaneously and the reward that corresponds to the joint action is broadcast to all agents.

Table 2 describes the reward function for a simple cooperative single-stage game. For example, if agent 1 executes action b and agent 2 executes action a, the reward they receive is 5. Obviously, the optimal joint-action in this simple game is (b, b) as it is associated with the highest reward of 10.

		Agent 1	
		a	b
Agent 2	a	3	5
	b	0	10

Table 2: A simple cooperative game reward function.

Our goal is to enable the agents to learn optimal coordination from repeated trials in cases where the game matrix is not known to the agents. To achieve this goal, one can use either *independent* or *joint-action* learners. The difference between the two types lies in the amount of information they can perceive in the game. Although both types of learners can perceive the reward that is associated with each joint action, the former are unaware of the existence of other agents whereas the latter can also perceive the actions of others. In this way, joint-action learners can maintain a model of the strategy of other agents and choose their actions based on the other participants' perceived strategy. In contrast, independent learners must estimate the value of their individual actions based solely on the rewards that they receive for their actions. In this paper, we focus on individual learners, these being more widely applicable.

In the present study, we analyse a number of coordination problems, all of which are descendants of the *climbing game* and the *penalty game*. We show why these problems are of interest to us and why they are hard problems. The climbing game is representative of problems with high miscoordination penalties and a single optimal joint action, whereas the penalty game is representative of problems with miscoordination penalties and multiple optimal joint actions. Both games are played between two agents and their reward functions are shown in Tables 3 and 4:

		Agent 1		
		a	b	c
	a	11	-30	0
Agent 2	b	-30	7	6
	c	0	0	5

Table 3: The climbing game.

In the climbing game, it is difficult for the agents to converge to the optimal joint action (a, a) because of the negative reward in the case of miscoordination. For example, if agent 1 plays a and agent 2 plays b, then both will receive a negative reward of -30. Incorporating this reward into the learning process can be so detrimental that both agents tend to avoid playing the same action again. In contrast, when choosing action c, miscoordination is not punished so severely. Therefore, in most cases, both agents are easily tempted by action c. The reason is as follows: if agent 1 plays c, then agent 2 can play either b or

		Agent 1		
		a	b	c
	a	10	0	-10
Agent 2	b	0	2	0
	c	-10	0	10

Table 4: The penalty game.

c to get a positive reward (6 and 5 respectively). Even if agent 2 plays a, the result is not catastrophic since the reward is 0. Similarly, if agent 2 plays c, whatever agent 1 plays, the resulting reward will be at least 0. From this analysis, we can see that the climbing game is a challenging problem for the study of learning coordination. It includes heavy miscoordination penalties and "safe" actions that are likely to tempt the agents away from the optimal joint action.

Similarly, the penalty game is a hard problem as it not only has potentially high penalties for miscoordination (depending on the choice of k) but also includes multiple optimal joint actions. If agent 1 plays a expecting agent 2 to also play a so they can receive the maximum reward of 10 but agent 2 plays c (perhaps expecting agent 1 to play c so that, again, they receive the maximum reward of 10) then the resulting penalty can be very detrimental to both agents' learning process. In this game, b is the "safe" action for both agents since playing b is guaranteed to result in a reward of 0 or 2, regardless of what the other agent plays, thus maximizing the worst-case reward.

Because of the difficulties discussed above, regular Q-learning agents failed to converge to the optimal joint action in both the climbing game and the penalty game. A Q-learning variant that solves these games for independent agents has been presented in the past (Kapetanakis and Kudenko, 2002) using an approach known as the FMQ heuristic. This heuristic works in a reinforcement learning setting and it allows two agents that have no communication capabilities or shared knowledge to jointly reach the optimal joint action in single-stage games.

However, the FMQ heuristic cannot distinguish adequately between miscoordination penalties and reward variance in stochastic games. Therefore, it fails to solve more complex games such as the *stochastic climbing game* which is shown in Table 5. The stochastic version of the climbing game differs from the original in that each joint action now corresponds to two rewards instead of just one. These two rewards are received with probability $1/2$. If the two agents were to commit to playing a specific joint action indefinitely, the reward they would accumulate over time would converge to the same value as in the original game. In this respect, the stochastic climbing game is equivalent to the original. This equivalence is maintained across all variations of the climbing and penalty game that we introduce in this work.

		Agent 1		
		a	b	c
	a	10/12	5/-65	8/-8
Agent 2	b	5/-65	14/0	12/0
	c	5/-5	5/-5	10/0

Table 5: The stochastic climbing game table (50%).

The difficulty in solving the stochastic climbing game with the FMQ heuristic stems

from the fact that the heuristic is designed to deal with one type of uncertainty, namely that which arises from sampling actions with associated mis-coordination penalties. This uncertainty is due to the inability of one agent to observe the other agent's actions. The FMQ heuristic filters out the impact of failed coordination attempts to help the agents reach the optimal joint action. The difference in the stochastic version of the climbing game is that there are now two sources of uncertainty in the game, the other agent's actions (as before) and the multiple rewards per joint action.

We will show two ways to tackle the stochastic climbing game using the idea of commitment sequences. We also presently introduce two new variants of the game so as to show the full potential of our methods. The first variant is the *three-valued climbing game* in which that there are three rewards corresponding to any joint action. This game is shown in Table 6.

		Agent 1		
		a	b	c
	a	16/22/-5	4/6/-100	10/20/-30
Agent 2	b	4/6/-100	25/0/-4	10/5/3
	c	8/12/-20	10/20/-30	4/5/6

Table 6: The three-valued stochastic climbing game.

The probability of receiving any one of the three rewards that correspond to each joint action is 1/3. If the two agents were to play the action profile (b, b) indefinitely, they would accrue an average reward of 7 as in the original game.

The next variant of the climbing game that we introduce is the *variable-probability climbing game*. There are again two rewards for each joint action. However, they are now received with non-uniform probabilities. Equivalence with the original game is again maintained. The variable-probability climbing game is included in Table 7. The notation $(\pi)n$ signifies that the probability of getting a reward of n for playing that joint action is π. For example, the probability of getting a reward of 5 for joint action (c, b) is 0.8 whereas the probability of getting a reward of 10 for the same joint action is 0.2 .

		Agent 1		
		a	b	c
	a	(0.4) 6.5 (0.6) 14	(0.25) -36 (0.75) -28	(0.6) 4 (0.4) -6
Agent 2	b	(0.25) -36 (0.75) -28	(0.8) 5 (0.2) 15	(0.8) 5 (0.2) 10
	c	(0.7) 3 (0.3) -7	(0.6) 4 (0.4) -6	(0.8) 4 (0.2) 9

Table 7: The variable-probability stochastic climbing game.

Finally, we introduce a stochastic variant of the penalty game. This game is called the *stochastic penalty game* and is shown in Table 8.

4 Reward estimation

In games with stochastic rewards, such as all the climbing game variants and the stochastic penalty game, it is difficult to distinguish between the two sources of variation in the

		Agent 1		
		a	b	c
	a	8/12	-3/3	-8/12
Agent 2	b	-3/3	0/4	-3/3
	c	-8/-12	-3/3	8/12

Table 8: The stochastic penalty game (50%).

observed reward for an action. It would be useful to have a protocol that allows 2 or more agents to select the same joint action repeatedly in order to build up a model for the stochastic reward distribution. This section describes a novel approach for achieving this.

The basic principle is that agents follow a common action selection policy that enables them to estimate the potential reward for each joint action. The action selection policy is based on the following idea: if an agent chooses an action at time i, then the agent is required to choose the same action at specific future time points. The only assumption that this approach makes is that all agents share the same global clock and that they follow a common protocol for defining sequences of time-slots.

4.1 Commitment sequences

A *commitment sequence* is a list of *time slots* $(t_1, t_2, \ldots)$ for which an agent is committed to taking the same action. If two or more agents have the same protocol for defining these sequences, then the ensemble of agents is committed to selecting a single joint-action for every time point in the sequence. Although each agent does not know the action choices of the other agents, it can be certain that the observed rewards will be statistically stationary and represent unbiased samples for the reward distribution of *some* joint action. In order to allow an arbitrarily high number of joint actions and consequently commitment sequences to be considered as the agent learns, it is necessary that the sequences have an increasing time interval $\delta_i \equiv t_{i+1} - t_i$ between successive time slots. A sufficient condition is that $\delta_{i+1} \geq \gamma\delta_i$ where $\gamma > 1$ for all $i > i_0$ (for some pre-defined constant i_0). In the results given here, sequences are infinite with $\gamma = 5/4$.

Here, the successive increments are generated by the function:

$$\delta_{i+1} = \left\lfloor \frac{c\delta_i + c - 2}{c - 1} \right\rfloor$$

where $c > 1$ is the increment factor and $\lfloor \cdot \rfloor$ indicates rounding down to an integer value. For example, if the increment factor is 5 (i.e., the increment ratio γ is $5/4$) the rule becomes: $\delta_{i+1} = \lfloor (5\delta_i + 3)/4 \rfloor$. The first such sequence starts with $(1, 3, 6, 10, 15, 22, \ldots)$. The second sequence therefore starts at time slot 2. To prevent any 2 sequences from selecting the same time slot, each sequence excludes the time slots in the existing ones. Hence the second sequence starts with $(2, 5, 9, 14, 20, 28, \ldots)$.

For time i suppose the agents chose actions $(a_1^i, a_2^i, \ldots, a_m^i)$ (where m is the number of agents). Then an estimate of the value of this joint action is available as the average reward received during the part of the sequence that has been completed so far. Longer sequences provide more reliable estimates. Initially, we evaluate the simplest approach possible where the agents maintain a running reward average for all the active commitment sequences. Later, we will explore a method that takes into account the stochasticity in the rewards and the length of sequences within the framework of Gaussian estimation.

4.2 Action selection policy

Each agent must choose an action at the start of each sequence. A new sequence starts whenever no existing sequence is active in the current time slot. There are two obvious ways to select the new action: either explore (select the action randomly and uniformly) or exploit (select the action currently considered optimal). The simple approach used here is to choose randomly between exploration and exploitation for each sequence. In order to prefer longer sequences (more reliable estimates), we maintain statistics about the commitment sequences that are active. One such statistic is the length of the sequence until the current time point. Agents only consider a commitment sequence to yield a reliable reward estimate if its length becomes greater than a threshold value, N_{min}. In these experiments, N_{min} was set to 10.

For a 2-agent system, we chose the exploration probability p to be 0.9. As an exception, the first N_{init} sequences (where $N_{init} \geq 1$) must be exploratory to ensure that an exploitative action can be calculated. The algorithm followed is shown in Figure 1. In the results below, $N_{init} = 10$.

if number of sequences $< N_{init}$ **then**
  explore randomly and uniformly
else
  with probability p : explore randomly and uniformly
  with probability $1 - p$: exploit
end if

Figure 1: The average reward estimation algorithm.

The $\mathrm{exploit}$ function simply returns the action that corresponds to the commitment sequence with the highest observed average reward among those whose current length is at least N_{min}.

4.3 Parameter analysis

In this section, we analyse the influence of the algorithm's parameters on the learning. These parameters are: the minimum number of commitment sequences that must have been started before the agents can exploit (N_{init}), the minimum length that a sequence must reach before it is considered for exploitation (N_{min}), the increment factor (c) and the exploration probability (p).

N_{init} was set to 10 in all our experiments. Since none of our experiments have less than 10 commitment sequences, N_{init} has no role in the learning other than to make sure that, upon selection of an exploitative action, there are *some* sequences among which to choose.

N_{min} affects the learning as follows. In experiments where only short commitment sequences are created (either short experiments or long ones with a small increment factor), learning performance improves by setting N_{min} to a reasonably high value since our confidence is higher for longer commitment sequences. In longer experiments, N_{min} plays no part in the learning other than to decrease learning performance if set too high. This is because it is possible that an agent never explores as no commitment sequences reach the required length. Figure 2 illustrates this effect. We have plotted the convergence probability of the learners in the stochastic climbing game after 1000 moves. Since no commitment sequences of length greater than 20 have been created, the learning performance decreases for values of $N_{min} > 20$.

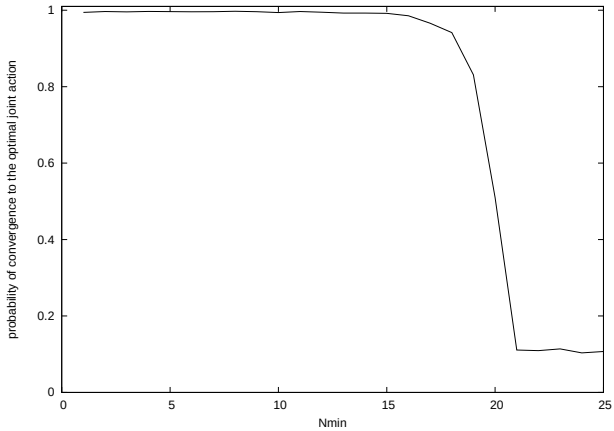


Figure 2: Number of successful experiments out 1000 runs for $N_{min} \in [1, 25]$.

The increment factor c defines how quickly a commitment sequence is visited again. Informally, the higher the value of the increment factor, the greater the time between two successive updates of a commitment sequence. Consequently, the total number of commitment sequences reduces monotonically with the increment factor for any length of the experiment. To illustrate this relationship, Figure 3 shows the total number of sequences for an experiment of 1000 moves with increment factor $c \in [2, 29]$.

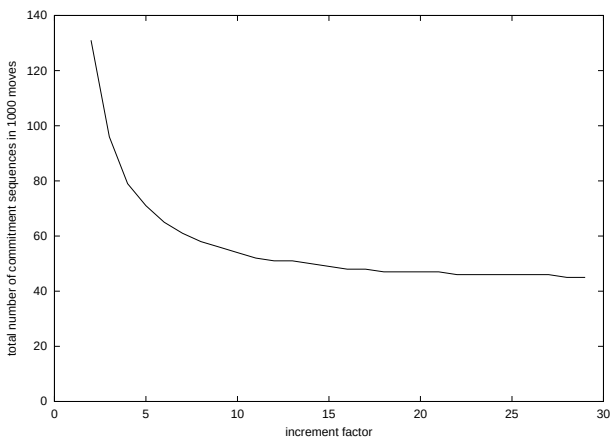


Figure 3: The relationship between the increment factor and the total number of sequences.

The performance of the learners for different increment factors is also plotted in Figure 4. This plot was generated for the stochastic climbing game with $N_{min} = 10$ and experiment length 1000.

Finally, in order to understand the influence of the exploration probability p on the learning, we have plotted its effect on the learning performance in Figure 5. The explo-

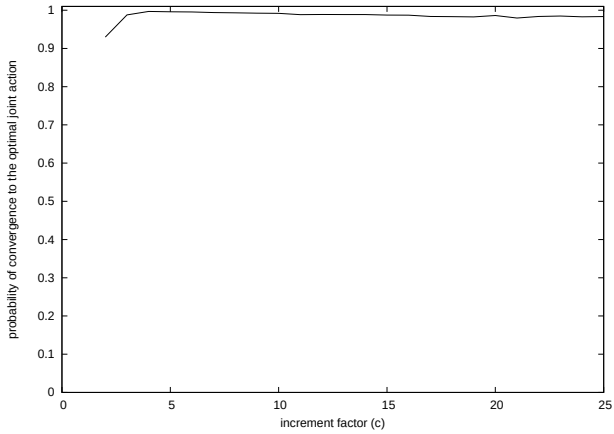


Figure 4: The relationship between the increment factor and the probability of convergence to the optimal joint action.

ration probability has been varied from 0 (always exploit) to 1 (always explore), N_{min} and N_{init} were set to 10 and the experiment was 1000 moves long. This plot was created for the stochastic climbing game.

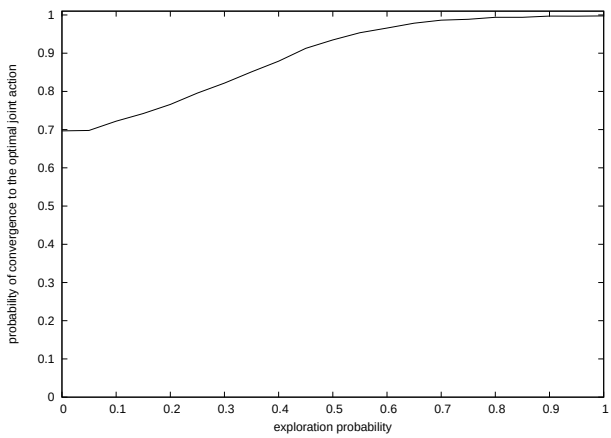


Figure 5: The relationship between the exploration probability and the probability of convergence to the optimal joint action.

From Figure 5, we can see that 100% exploration is optimal in this case. However, we have identified some games where one agent's exploitative behaviour can help the other agent to learn. If performance during learning is an issue, a low exploration probability is also desirable. In our experimental evaluation, the exploration probability will be 0.9.

4.4 Experimental results

This section contains the experimental results for the basic approach in all three versions of the climbing game and in the stochastic penalty game. As before, we repeated each experiment 1000 times to obtain high confidence in the results. The number of moves was varied from 500 to 3000 and the parameters N_{init} and N_{min} were both set to 10. In all experiments, we chose $\gamma = 5/4$. The results for the climbing game variants are plotted in Figure 6.

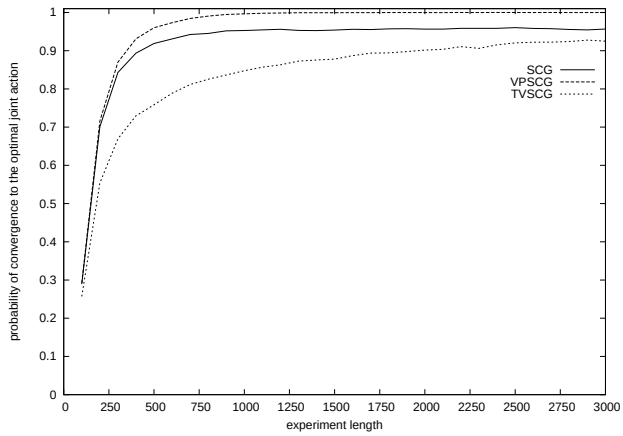


Figure 6: Convergence of the basic approach on the three games.

In Figure 6, SCG stands for Stochastic Climbing Game, TVSCG stands for Three-Valued Stochastic Climbing game and VPSCG stands for Variable-Probability Stochastic Climbing Game. From Figure 6, we can see that the probability of convergence to optimal eventually reaches over 90% for all cases, with some reaching over 95% even for relatively short experiments. The most difficult game is the three-valued climbing game as there is more variance in the rewards that correspond to each joint action.

The stochastic penalty game is solved much quicker by this approach. The probability of convergence to the optimal joint action reaches over 95% even for very short experiments (500 moves). This is because the variance in the rewards is small and the method is impervious to the existence of multiple optimal joint actions. Regardless of how many optimal joint actions there are in the game, one will always have a higher estimate than others[1] so that choosing an action for exploitation is not affected by the number of optimal joint actions.

5 Variance-sensitive approach

The basic approach using averaging performs fairly well. However, there are cases (e.g. the three-valued stochastic climbing game) where convergence only reaches above 90% for very long experiments. This section outlines an extended approach using a Gaussian

[1]If multiple estimates are equal, the agents will choose the one that corresponds to a longer commitment sequence.

estimator that improves both the convergence speed and the probability of convergence to the optimal joint action.

5.1 Finding the exploitative action

As in the basic approach, if the agents chose actions $(a_1^i, a_2^i, \ldots, a_m^i)$ (where m is the number of agents) at time i then an estimate of the value of this joint action is available as the average reward received so far for this sequence. For the extended approach, we attempt to reason about the *true* expected reward. To do this, we must make some assumptions about the possible form of the reward for each joint action, e.g. that it must have finite variance.

Here we use a Gaussian model and estimate its mean and variance from the observations. If n rewards are observed with empirical average m and sum of squares S, we obtain estimates for the population mean μ and *its* variance σ_μ (estimates of a quantity x are denoted by $\hat{x}$):

$$\hat{\mu} = m$$

$$\hat{\sigma}_\mu^2 = \frac{S + \sigma_0^2}{n^2} - \frac{m^2}{n}$$

σ_0 is a parameter to the algorithm and should be based on the expected variance of rewards in the game. In order to prefer longer sequences (more reliable estimates), a pessimistic estimate $\hat{\mu} - N_\sigma \hat{\sigma}_\mu$ is used to provide a lower bound on the expected return for each sequence. At any given time, the exploitative behaviour for an agent is to choose the action corresponding to the sequence with the greatest lower bound. Large values of N_σ reduce the risk that an optimistic bias in the reward estimate from a short sequence will affect the choice of action. However, smaller values may give faster initial learning. In the results below, $N_\sigma = 4$.

5.2 Exploration policy

In the variance-sensitive approach, the agents choose randomly between exploration and exploitation for each sequence. For a 2-agent system, we chose the exploration probability to be 0.9 as before. We have maintained the N_{init} parameter ($N_{init} >= 1$) but have eliminated N_{min}. We simply allow N_{init} commitment sequences to start and then only use the variance-sensitive estimator to find the exploitative action. In the results below, $N_{init} = 10$.

5.3 Experimental results

Figure 7 depicts the convergence performance of the variance-sensitive approach for the three variants of the climbing game. In these experiments, σ_0 was set to 50 for the three-valued stochastic climbing game and to 10 for the rest. In Figure 7, SCG, VP-SCG and TVSCG stand for Stochastic Climbing Game and its Variable-Probability and Three-Valued variants.

As we can see from Figure 7, the variance-sensitive approach outperforms the averaging approach in all cases. In fact, the probability of convergence to the optimal joint action consistently reaches over 98%. For the three-valued game, we chose $\sigma_0 = 50$ because the variance in the stochastic rewards is higher. The probability of convergence to the optimal joint action in this game reached over 90% for longer experiments i.e. after 2000 moves.

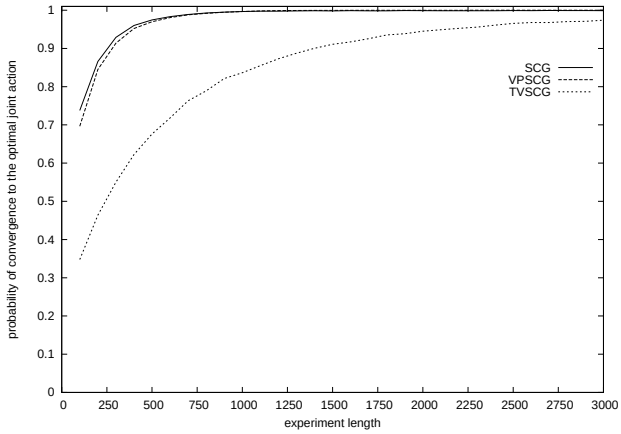


Figure 7: Convergence of the variance-sensitive (Gaussian) approach on the three games.

Finally, the stochastic penalty game was once again much easier to solve. Here, the probability of convergence to the optimal joint action exceeded 95% for 200 moves and reached 100% after only 800 moves. This clearly illustrates the ability of the variance-sensitive approach to resolve the problem of choosing between multiple optimal joint actions.

6 Related work

There are two main paradigms for the learning of coordination, one using independent agents and another using joint-action learners. While joint-action learners are able to observe one another's actions, their independent counterparts can not. Therefore, approaches using independent learners (such as ours) are more general and more universally applicable.

Claus and Boutilier (Claus and Boutilier, 1998) used joint-action learners and fictitious play in their approach to learning coordination in cooperative multi-agent systems but reported a failure to solve problems where miscoordination is heavily penalised. Later, Boutilier (Boutilier, 1999) developed an extension to the value iteration technique that allowed each agent to reason explicitly about the state of coordination, i.e., whether the group of agents are in a coordinated or non-coordinated state. More recently, Chalkiadakis and Boutilier described a Bayesian approach to multi-agent reinforcement learning problems but, again, made the limiting assumption that each agent can observe the actions and rewards of all other agents at each round of the game. Also, Wang and Sandholm (Wang and Sandholm, 2002) developed a learning algorithm for joint-action learners that provably converges towards an optimal Nash equilibrium.

Sen, Sekaran and Hale (Sen et al., 1994) argued convincingly that shared knowledge or the ability to communicate is not a necessary condition for multi-agent coordination. They implemented a system where two independent agents learnt to coordinate their actions so as to push a block to a specific location without even being aware of one another. However, their agents did only converge to suboptimal policies.

Similarly, Peshkin, Kee-Eung, Meuleau and Kaelbling) (Peshkin et al., 2000) devel-

oped a gradient-descent policy-search algorithm for cooperative multi-agent domains that is guaranteed to find a local optimum in the space of factored policies but may not always find an optimal Nash equilibrium.

Finally, Nowé, Parent and Verbeeck (Nowé et al., 2001) used social agents that employ a periodical policy to tackle learning in single-stage 2-player games. In contrast to our approach, these agents rely on communication to achieve coordination.

7 Concluding remarks and outlook

We have presented a novel learning technique based on commitment sequences that enables independent agents to converge to the optimal joint action even in difficult scenarios with miscoordination penalties and stochastic rewards. Such scenarios were previously approached with Q-learning techniques but remained unsolved using independent agents.

The ability to achieve coordination with independent agents is a major advantage of the commitment sequence approach and makes the technique much wider applicable in the real world. Specifically, by not relying on explicit communication and mutual observation, the technique is insensitive to disruptions in the communication channel and does not require the spatial proximity of agents.

A potential disadvantage of commitment sequences is the dependence on a shared global clock in order to achieve coordination. This is a limiting assumption in the sense that such a clock may not always be available. In such cases, the reward signal may sometimes be used to emulate the gloabl clock. However, the clock needs to be fail-safe and keep the agents synchronised at all times. Potential glitches can be overcome by permitting limited communication at regular time intervals to check synchronisation.

We are currently investigating the scaling-up performance of the commitment sequence approach, both in terms of the number of agents and in terms of the size of the game (i.e., the number of actions available to each agent). Early results show that our technique scales up well.

While agent coordination in the real world is typically much more complex than what can be described by a single-stage cooperative game, these games offer a useful abstraction. To study the essential dynamics of coordination in a real-world scenario, it is helpful to simplify the problem so as to isolate the issues involved and study them separately. In this way, the effect of each issue can be highlighted and its contribution to the overall problem of coordinating actions can be extracted. That is the main reason for using single-stage games as the testbed for learning coordination. It is their simplicity and expressional power that is well suited to this type of research.

To extend our system to multi-stage games will require a combination of the commitment sequence approach (applied separately in each state to evaluate the expected immediate reward) with state-action value functions that estimate expected discounted returns taking into account the state-transition function. Value functions of this kind may be associated with sequence-action pairs and/or state-action pairs. A hybrid method that makes use of reinforcement learning heuristics, such as FMQ (Kapetanakis and Kudenko, 2002), will be considered first.

References

Boutilier, C. (1999). Sequential optimality and coordination in multiagent systems. In *Proceedings of the Sixteenth International Joint Conference on Articial Intelligence*

(IJCAI-99), pages 478–485.

Chalkiadakis, G. and Boutilier, C. (2003). Coordination in multiagent reinforcement learning: A Bayesian approach. In *Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems*, Melbourne, Australia.

Claus, C. and Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Articial Intelligence*, pages 746–752.

Fudenberg, D. and Levine, D. K. (1998). *The Theory of Learning in Games*. MIT Press, Cambridge, MA.

Griffin, D. (1984). *Animal Thinking*. Harvard University Press, Cambridge, MA.

Kapetanakis, S. and Kudenko, D. (2002). Reinforcement learning of coordination in cooperative multi-agent systems. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI'02)*.

Lauer, M. and Riedmiller, M. (2000). An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proceedings of the Seventeenth International Conference in Machine Learning*.

Legge, D. and Baxendale, P. (2002). An agent-based network management system. In *Proceedings of the AISB'02 symposium on Adaptive Agents and Multi-Agent Systems*, pages 125–130, London, UK.

Legge, D. and Baxendale, P. (2003). An agent-managed ant-based network control system. In *Proceedings of the AISB'03 symposium on Adaptive Agents and Multi-Agent Systems*, pages 23–30, Aberystwyth, UK.

Nowé, A., Parent, J., and Verbeeck, K. (2001). Social agents playing a periodical policy. In *Proceedings of the 12th European Conference on Machine Learning*, Freiburg, Germany.

Peshkin, L., Kim, K.-E., Meuleau, N., and Kaelbling, L. (2000). Learning to cooperate via policy search. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*.

Sen, S., Sekaran, M., and Hale, J. (1994). Learning to coordinate without sharing information. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 426–431, Seattle, WA.

Wang, X. and Sandholm, T. (2002). Reinforcement learning to play an optimal Nash equilibrium in Team Markov Games. In *Proceedings of the 16th Neural Information Processing Systems: Natural and Synthetic (NIPS) conference*, Vancouver, Canada.

Weiss, G. (1993). Learning to coordinate actions in multi-agent systems. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, volume 1, pages 311–316. Morgan Kaufmann Publ.

AISB Journal **1**(5), © SSAISB, 2004
http://www.aisb.org.uk

Learning Multi-agent Search Strategies

Malcolm J A Strens

Future Systems Technology Division, QinetiQ Ltd.
G020/A9, Cody Technology Park, Farnborough, Hants. GU14 0LX. U.K.
mjstrens@QinetiQ.com

Abstract

We identify a specialised class of reinforcement learning problem in which the agent(s) have the goal of gathering information (identifying the hidden state). The gathered information can affect rewards but not optimal behaviour. Exploiting this characteristic, an algorithm is developed for evaluating an agent's policy against *all possible* hidden state histories at the same time. Experimental results show the method is effective in a two-dimensional multi-pursuer evader searching task. A comparison is made between identical policies, joint policies and "relational" policies that exploit relativistic information about the pursuers' positions.[1]

1 Introduction

We address the reinforcement learning problem (Sutton and Barto, 1998) for episodic tasks in partially observable environments. These tasks are characterised by the presence of 'hidden state' which is not observable by the agent, although it may be revealed over the course of a trial.

For example, suppose the task is to search an area of the ground using a robotic vehicle (the *pursuer*, under the learning agent's control). Initially, the pursuer may only know that the evader (another vehicle perhaps) is within some given uncertainty area. The pursuer may also have available a dynamics model that determines how the evader could move over the course of a trial. The goal is simply to bring the evader within detection range of the pursuer's sensors (and at this point the trial ends). Therefore the only use that the pursuer can make of its sensor measurements during the course of the trial is to eliminate regions of the search area: *i.e.* the pursuer must reason about *where the evader could be*. This is an extreme example of a *dual control* problem: an agent must take actions to gather information on the way to its goal.

If the means available to the agent for observing the environment (*e.g.* sensors) do not allow the agent to gather instantaneously all the information that is relevant to its decision-making, then the task is a *partially observable* (PO) one. In PO tasks, reactive policies based on only the instantaneous observations are rarely effective: it is necessary for the agent to fuse information over time to make best use of the observations (and initial information). The searching task is a good example: the instantaneous observations are identical at every time step until the evader is found, and so no useful memoryless policy could be found. However, if the agent uses the observations to modify a longer-term

[1] This paper is ©QinetiQ 2004.

memory representing where the evader could be at the current time, it can use this memory as the foundation for decision-making.

A standard model for reasoning about PO environments is the Partially Observable Markov Decision Process (POMDP). This models the transition dynamics of the (full) environment state, the expected return for every (state, action) pair, and a stochastic function mapping states to observations. Most success for large POMDP problems has come from restricting the complexity of learning by choosing the agent's policy from a parameterised family (our approach), rather than estimating state-action values for every information state. The learning problem becomes one of finding appropriate policy parameters. However, even with parameterised policies there remains a difficulty in how to define the 'inputs' to the policy. In fully observable problems, these inputs should be a compact representation of the state (a *feature vector*) that carries with it all information relevant to decision-making. Similarly, in PO problems the inputs can be *information features* that encode aspects of the belief relevant to decision-making. For example, in the detection task where the belief is an evader uncertainty area (EUA), candidate sets of features might include informative points in the EUA (center of mass, extrema of boundaries) or spatial moments of the distribution.

Section 2 gives formal descriptions for a POMDP and the proposed specialisation, an *information gathering problem*. Section 3 introduces recursive Bayesian filters for tracking beliefs, including the particle filter used in the experiments. Section 4 describes a set of appropriate information features for the searching task, and three types of policy parameterisation (individual, joint, and relational). Section 5 describes a fast way of learning in information gathering problems that exploits conditional independence between the true hidden state and the optimal behaviour, given the belief. Section 6 describes direct search methods for finding effective policy parameters. Section 7 describes evaluation of the approach using a multi-pursuer evader task, and section 8 concludes.

2 A special class of POMDP

Before describing a specialised version of the POMDP for information gathering problems, we give a formal description of the existing MDP and POMDP models. An MDP is a discrete-time model for the stochastic evolution of a system's state, under control of an external input (the agent's actions). It also models a stochastic reward that depends on the state and action.

> *Definition* A Markov Decision Process is given by $\langle X, A, T, R\rangle$ where X is a set of states and A a set of actions. T is a stochastic transition function defining the likelihood the next state will be $x' \in X$ given current state $x \in X$ and action $a \in A$: $P_T(x'|x, a)$. R is a stochastic reward function defining the likelihood the immediate reward will be $r \in \mathbf{R}$ given current state $x \in X$ and action $a \in A$: $P_R(r|x, a)$.

A POMDP is a general-purpose discrete-time mathematical model for reasoning about single-agent interaction in the presence of partial observability. The POMDP assumes that the agent receives observations that do not necessarily convey the full hidden state of the environment.

> *Definition* A POMDP $\langle X, A, T, R, Y, O\rangle$ builds upon a MDP $\langle X, A, T, R\rangle$ by adding a set of observations Y and an observation function O that generates stochastic observations from the hidden state according to $P_O(y|x)$.

2.1 Optimal behaviour in POMDPs

Reinforcement learning for partially observable problems is generally more difficult than for fully-observable ones. The reason for this can be seen by analysing optimal behaviour in a POMDP (assuming all parts of the POMDP are known[2]). Optimal behaviour is given not by a policy that maps observations to action probabilities, but by a policy that maps *beliefs* to action probabilities. The agent's belief or *information state* is its representation for uncertainty in the hidden state at the current time. This belief is a probability distribution $P(x|H)$ over X, where H is the complete interaction history. It is possible to find the agent's optimal policy by constructing a MDP from the POMDP but with a much larger state space corresponding to the set of possible beliefs. This larger MDP can, in theory, be solved by standard RL methods. In practice, however, the belief space is usually so large that the methods are intractable without some kind of approximation. For example Thrun estimated state-action values for a discrete set of belief 'exemplars', each corresponding to a particle filter estimate of the belief (Thrun, 2000). The KL divergence was used as a distance-measure between information states, to allow the nearest neighbor exemplar to be found for each information state encountered during learning. In the worst case, the belief space has size exponential in the number of time steps of interaction.

2.2 Information gathering problems

Now we investigate special cases where part of the state is *fully* hidden. By *fully* hidden we mean that observations that depend on this state do not affect optimal behaviour (and are therefore excluded from the model). Firstly, suppose that the fully hidden state represents the location of a physical entity (*e.g.* a vehicle) and that the agent's *only goal* is to discover this information. When the information is discovered the trial ends. In this case the agent knows its own physical state, and observations convey no information about the hidden state until the end of the trial. Therefore optimal searching behaviour is not dependent on these observations. More generally, consider an agent that aims to discover some part of the hidden state, but passes the information on to another agent for action. The trial does not end, but from that point on the agent has "no interest" in that part of the state. An example would be a remote surveillance system (e.g. a relocatable satellite) that must plan its own course to gather imagery at a set of locations, and report this information for interpretation by another system (e.g. a ground station). This is an example of a *continuing* information gathering problem (IGP) for which the following definition is proposed:

> *Definition* An IGP is given by $< X, S, A, T^h, T^s, R >$. The fully hidden part of the state $x \in X$ evolves according to a transition function $P_{T^h}(x'|x, s, a)$. The observable part of the state $s \in S$ evolves separately according to a transition function $P_{T^s}(s'|s, a)$. R defines the stochastic immediate reward $P_R(r|x, s, a)$ which may depend on both the hidden part and the observable part of the state.

Note that there is no observation function because the observations are identical to the observable part of the state (s). The observable state could incorporate the physical state of the agent(s) and any other observable scenario information. A stationary policy for the

[2]If the observation, transition and reward functions of the POMDP are not known *a priori* an even more difficult learning problem is encountered.

agent is now a stochastic function of (i) the belief $b \equiv P(x|H)$ over the fully hidden state and (ii) the fully observable state:

$$\pi(b, s, a) \equiv P(a|b, s)$$

The benefit of this new formulation is that the belief b evolves independently of the actual fully hidden state x. (There are no observations to convey information about it.) The belief b is obtained simply by a recursive Bayesian filter that has no information update step.

3 Recursive Bayesian filtering

Bayesian filtering is the process of estimating the distribution of possible values (belief) for the state of a dynamic system, given a sequence of noisy measurements. A *recursive* Bayesian filter implements this process by using only the current measurement to update the belief at each time step; it never refers back to previous measurements. Consider first the non-interactive case where the agent is simply observing the dynamic system. Implementation requires a way to represent the current belief $P(x_t|Y_t)$ where x_t is the state vector and Y_t is the sequence of measurements so far. There are many possible implementations. The Kalman filter approximates the belief by a multivariate Gaussian distribution. Particle filters use a 'cloud' of samples as the representation, allowing multi-modal distributions to be represented. Grid-based filters discretise the state space and store a probability at each grid point.

The filter has two updates at each step: *prediction* applies the known dynamics $P(x_{t+1}|x_t)$ to obtain a prior distribution $P(x_{t+1}|Y_t)$ for the state at next time step; *information update* uses Bayes' rule to account for the new observation y_{t+1}, yielding a posterior distribution (the belief at $t + 1$):

$$P(x_{t+1}|Y_{t+1}) \propto P(y_{t+1}|x_{t+1})P(x_{t+1}|Y_t)$$

This states that the likelihood of a new state (x_{t+1}) is proportional to ($\propto$) the likelihood of the observation (y_{t+1}) given that state, weighted by its prior probability given previous observations (Y_t).

For example, a recursive Bayesian filter can be applied to the problem of tracking the location of an aircraft using noisy sensor measurements from a radar. The state is the aircraft's location, pose, and speed. The dynamics is determined by the aircraft's acceleration capability. The measurement model $P(y_t|x_t)$ describes the radar's performance (e.g. typical error) and must be known. Applying the filter yields a belief that represents the uncertainty in the aircraft's state at each time step, and allows decisions to be made.

3.1 Formulation for interactive systems

In many sequential decision problems it is feasible to estimate the belief (even when it is not feasible to enumerate the space to represent a value function). Using the POMDP notation, the vector x_t represents the full state of the environment (e.g. the position and motion of one or more objects). The belief is a probability density for x_t given the interaction history $H_t \equiv (A_{t-1}, Y_t)$ where $A_{t-1} \equiv (a_1, \ldots, a_{t-1})$ is the action history and Y_t is the observation history. The initial information is $P(x_0)$. The recursive Bayesian filter makes use of a dynamics model $P_T(x'|x, a)$ and an observation model $P_O(y|x)$ to recursively estimate the belief:

$$\underbrace{P(x_{t+1}|A_t, Y_t)} = \int_{x_t} P(x_t|H_t) P_T(x_{t+1}|x_t, a_t) dx_t$$

$$P(x_{t+1}|H_{t+1}) \propto P_O(y_{t+1}|x_{t+1}) \overbrace{P(x_{t+1}|A_t, Y_t)}$$

3.2 The particle filter

Our experiments make use of the particle filter (Doucet et al., 2001) because it is very easy and convenient to work with: it represents the belief as a sum of weighted hypotheses $\{(x_t^i, w_t^i)\}$:

$$P(x_t|H_t) = \sum_{i=1}^{N} w_t^i \delta(x_t^i, x_t)$$

The update can be implemented (separately for each particle) by importance sampling, using some proposal density $q(x_{t+1}^i|x_t^i, H_{t+1})$ then re-weighting the particles according to:

$$w_{t+1}^i \propto w_t^i \frac{P_O(y_t|x_t^i) P_T(x_{t+1}^i|x_t^i, a_t)}{q(x_{t+1}^i|x_t^i, H_{t+1})}$$

Over the course of time, the weights of the particles may become imbalanced causing the set of particles to be a poor representation of the true belief. To overcome this problem, *resampling* is usually necessary. Resampling makes all weights equal, but there tend to be more copies in the new population of the particles that had the largest weights.

3.3 Particle filter for fully hidden state

The full recursive Bayesian filter is not required in information gathering problems. In particular, there are no observations conveying information about the hidden state, and so the belief depends only on the initial distribution $P(x_0)$ and on the dynamics expressed as a transition model $P_T^h(x_{t+1}|x_t, a_t)$ (now using IGP notation). The update becomes a single step:

$$P(x_{t+1}|H_{t+1}) = \int_{x_t} P(x_t|H_t) P(x_{t+1}|x_t, s_t, a_t) dx_t$$

where $H_t \equiv (A_{t-1}, S_t)$ and $S_t \equiv (s_1, \ldots, s_t)$ is the observable state history. We will be use this "history filter" for two different purposes: representing an agent's beliefs and representing an ensemble of scenarios for evaluation.

A detection event in the pursuer-evader problem will lead to the weight of the corresponding hypothesis being set to 0. If resampling were then to take place, these zero-weight hypotheses would be replaced in the filter by duplicates of other hypotheses; i.e. the computational effort of the whole particle filter would be focussed on cases where the evader would not yet have been intercepted. However, there is a strong argument for not resampling[3]. Given that no computational effort is expended on zero-weight particles, the processing requirement will be proportional to the likelihood that the evader has not yet been detected. This implies that, without resampling, computational load is matched

[3]However, resampling *is* required if the transition function does not assign similar probabilities to all feasible outcomes.

to the maximum change in total return from the trial. Therefore, in the pursuer-evader problem, the particle filter has been reduced to a very simple update for each hypothesis, implementing the evader's instantaneous motion.

4 Policy parameterisation

Our goal is for the learning agent to acquire effective control policies (searching behaviours). A policy is a means for selecting an appropriate action, given the current situation. For large problems, there are two main approaches: policy search and value function approximation. In policy search, the designer provides the agent with a parameterised control policy, and the parameters are the objective for learning. In contrast, value function approximation approaches (common in reinforcement learning) estimate a mapping from states to values. This allows an agent to exploit a known (or sampled) transition function to reason about the values of future states (using Bellman's "backup" operator within the learning rule), and to *derive* a control policy. The advantage of representing and exploiting the state information in this way is faster learning. The disadvantage is that it is difficult to meet the assumptions required by the Bellman operator (especially in the presence of large state spaces or partial observability).

The method used here is best described as "direct policy search" because it does not exploit Bellman's operator in learning (even though it does represent the policy intrinsically as a parameterised value function). The uncertainty in the evader's location is given at any instance in time by a belief consisting of $N_H = 256$ hypotheses (particles in the filter). To pose the problem as one of direct policy search, a set of information features is required that summarises the set of particles adequately for robust decision-making. This process is sometimes called *belief compression* (Roy and Gordon, 2002). The features described here make use only of the position part of each particle's state.

4.1 Information features

Let (x_i, y_i) be the position vector for hypothesis i expressed in a coordinate system of a pursuer of interest[4]. To obtain a compact set of information features, we integrate over the particles using a small number of basis functions ϕ_{jk} for $j \in \{1, 2\}$ and $k \in \{1, 2, 3, 4\}$. A suitable set is given by the regularised spatial derivatives:

$$\phi_{jk}(x, y) = H(j, 2k - 2)$$

where:

$$H(m, n) = \frac{\partial^m}{\partial x} \frac{\partial^n}{\partial y} \exp(-\frac{x^2 + y^2}{2})$$

These basis functions are all derivatives of a Gaussian distribution placed at the origin (of the pursuer's coordinate system). They are mutually orthogonal and (multiplicatively) separable in the two axis directions. Only even derivatives are selected in the y-direction;

[4]This coordinate system has its x-axis parallel to the pursuer's velocity vector. A nonlinear transformation is also applied: a displacement of (u, v) in the pursuer's coordinate system is mapped to $(x, y) \equiv (u/\{R_0 sqrt|u|\}, v/\{R_0 sqrt|v|\})$, in order to provide more spatial resolution close to the agent. R_0 is a constant that will determine the effective spatial extent of the basis functions, and is chosen equal to 512 in our evaluation.

the odd ones are of no interest as a result of symmetry in the problem. This yields a set of information features:

$$z_{jk} = \frac{1}{N_H} \sum_i \phi_{jk}(x_i, y_i)$$

We then define an information-state value function:

$$V_\alpha(x_i, y_i) = \sum_{j,k} \alpha_{jk} C_{jk} z_{jk}$$

C_{jk} is a constant that ensures the magnitudes of the information features are balanced:

$$C_{jk}^{-1} \equiv \int_x \int_y \phi_{jk}^2(x, y) \mathrm{d}y \mathrm{d}x$$

The 2x4 parameter matrix α will be the target for learning. We make use of the smooth dynamics to avoid computing V explicitly. Instead, each pursuer selects its action (turn left or turn right) according to the sign of the gradient of V with respect to the angle θ of its velocity vector. The gradient is given here without proof:

$$\frac{\partial V}{\partial \theta} \equiv \sum_i \sum_{j,k} C_{jk} \{x_i H(j, 2k-1) - y_i H(j+1, 2k-2)\}$$

Since only the sign of the gradient is used to determine the pursuer's action, an arbitrary scaling of α will not affect its behaviour. This redundancy is eliminated by requiring $\|\alpha\| = 1$.

4.2 Joint and relational policies

In general, every pursuer need not be given the same policy. Instead, the problem can be regarded as a search for the joint policy $(\alpha_1, \ldots, \alpha_{N_p})$ which has $8N_P$ dimensions. We will evaluate whether the extra representational freedom (and complexity) of a joint policy can be exploited.

In order to allow any number of pursuers to *cooperate* without the policy size increasing, a simple relational policy has also been designed. Every pursuer has a copy of the policy which consists of two components: α (as before) and β which contains information about the relative positions of the pursuers. β affects the value function in the same way as α except that the basis functions are summed over pursuer locations instead of particle locations:

$$z_{jk}^{rel} = \frac{1}{N_P - 1} \sum_l \phi_{jk}(x_l, y_l)$$

$$V_\beta(x_l, y_l) = \sum_{j,k} \beta_{jk} C_{jk} z_{jk}^{rel}$$

where (x_l, y_l) is the location of pursuer l expressed in the coordinate system of the pursuer for whom the decision is to be made. Using relational policies, it is the gradient of the combined value function $V_\alpha + V_\beta$ that determines each pursuer's action. The dimension of this policy is 16 nomatter how many pursuers are present. (Also, the number of pursuers taking part in the search could change dynamically during a trial.) This approach will be evaluated to determine whether pursuers are aided by taking account of each other's locations.

5 Dual particle filter method

One major benefit of identifying the problem as an "information gathering process" rather than (the more general) POMDP is that specialised solution methods can be derived. In the pursuer-evader problem, the pursuers' locations are the fully observable state (s) and the evader's location is the hidden state. When the evader comes within detection range of a pursuer, the trial ends. Therefore, the hidden state is never used to determine future actions: i.e. observation of the hidden state conveys no information about optimal behaviour (pursuer policy). The implication of this is that *any proposed policy can be tested against all possible hidden state histories in parallel*. The outcome of a parallel trial is not success or failure (evader detected or not) but instead the *proportion* of the histories in which the evader was detected.

The method proposed here uses one particle filter to represent the pursuers' beliefs about the location of the evader, and a separate particle filter to represent hidden state histories for parallel evaluation. Although it may not be immediately obvious, the two particle filters are estimating the same distribution: the possible state of the evader at each time step. Nevertheless, it is essential that they are kept separate because if the approximation errors (inevitable in a particle filter) were correlated, a grossly overoptimistic estimate of the performance of a policy would be obtained.

We found that 256 particles in each filter were sufficient, whereas using a single filter for both estimation tasks means that over 4000 particles are required for similar performance. Each trial does not end until the evader has been detected in every parallel scenario, or a maximum duration has expired. This approach is also very efficient compared with the naive alternative: running 256 separate trials; one for each possible trajectory taken by the evader.

6 Direct Policy Search

Suppose that the agent's policy has a parameter vector w of size m; for example $w \equiv \alpha$ for pursuers with identical policies; $w \equiv (\alpha_1 :: \ldots :: \alpha_{N_P})$ for joint policies (where :: indicates concatenation of vectors); or $w \equiv \alpha :: \beta$ for relational policies. A single simulation trial in which w defines the pursuers' behaviours will lead to a stochastic return. The aim of learning is to find a w that maximises the expected return. *Direct policy search* methods attempt to optimise expected return without reference to performance gradient information. Some methods use a large number of trials for each proposed w to obtain a near-deterministic evaluation, then apply a deterministic optimization procedure (e.g. downhill simplex method). However, here we use a method that can work with unreliable policy comparisons (Strens and Moore, 2001).

6.1 Differential evolution

Our approach is a variant of *differential evolution*, an evolutionary method that operates directly on a population of real-valued vectors (rather than binary strings) (Storn and Price, 1995). Proposals are obtained by linear combination of existing population members.

Initially, the population is chosen randomly from some prior density on w. To generate each proposal, one candidate in the population is chosen (systematically) for im-

provement. Then the vector difference between two more randomly[5] chosen members, weighted by a scalar parameter, F, is added to a third randomly chosen 'parent'. Crossover (see below) takes place between this and the candidate, to obtain a proposal point. The proposal is compared with the candidate by running a small set of new simulation trials for each. To reduce variance in this comparison, the same set of scenarios[6] is used for every comparison. According to the outcome *either* the proposal replaces the candidate *or* the population remains unchanged. In either case, a new candidate for replacement will be selected on the next iteration.

Crossover is implemented here by selecting each element of the proposal vector from either the candidate or the new vector with equal probability. Crossover helps to prevent the population from become trapped in a subspace. For each proposal, $\log_2 F$ was chosen uniformly from the range $[-10, 0]$ and the result was scaled by a value $F_{max}(t)$ that decreased with time. ($\log_2 F_{max}(t)$ was reduced uniformly from 0 to -10 during learning.) This ensures convergence of the population.

6.2 Illustration

Figure 1 illustrates this process in more detail: the weighted difference between two population members (1,2) is added to a third population point (3). The result (4) is subject to crossover with the candidate for replacement (5) to obtain a proposal (6). The proposal is evaluated (using a number of trials) and replaces the candidate if it is found to be better. Note that the proposal could be identical to (4) or (5), depending on the outcome of crossover.

This form of DE has a very useful property; replacing any one population member due to an occasional incorrect comparison is not catastrophic. It suffices that the comparison be unbiased, and correct with probability only slightly better than chance (0.5) in large populations (Strens and Moore, 2002). This means that it is possible to use only a small number of simulation trials (4 in this case) per proposal.

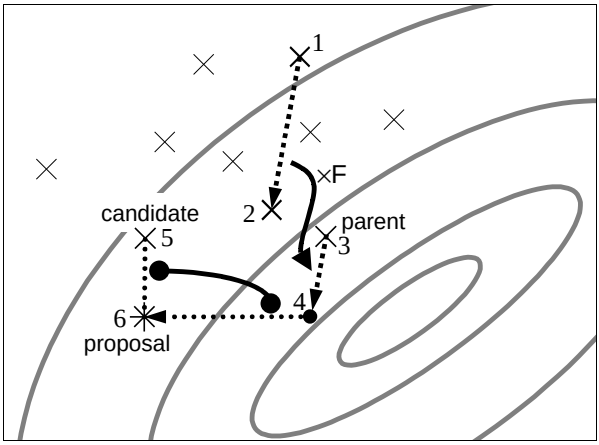


Figure 1: Obtaining a new proposal in differential evolution.

[5]All vectors used in the process of generating a proposal are mutually exclusive.

[6]A scenario is an initial configuration for the pursuers' locations.

7 Evaluation

We perform a comparison between individual, joint and relational policies for different numbers of pursuers. The aim is to demonstrate the effectiveness of the dual particle filter method and to collect some evidence about the best way to structure a policy for this type of task.

7.1 Multi-pursuer evader task

We consider a multi-pursuer evader task in which the observable state is the location and velocity of each pursuer, and the hidden state is the location and goal direction of an evader. The pursuers must cooperate to detect the evader before the uncertainty in its position becomes too large.

- **States.** Each pursuer's state is a location on the 2-dimensional plane and a motion direction. The evader's state is a location in the plane and a preferred direction; therefore each particle in the filters will be a 3-vector.

- **Scenarios.** The initial belief for the evader's location is an uncertainty area given by a Gaussian distribution with standard deviation 100, located at the origin. The pursuers all start at a distance from the origin chosen uniformly from $[2000, 3000]$. Their angular separation (subtended from the origin) is 30 degrees. Note that rotational and translational invariance of our formulation will mean that the policy obtained will work just as well for any rotated or translated versions of this set of scenarios.

- **Dynamics.** The pursuers can move at a speed of 24 (length units per time step) and the evader at 1. The evader can instantaneously change its direction whereas the pursuers have a maximum turn angle of $\pi/8$ at each time step. In our implementation, each pursuer selects either $+\pi/8$ or $-\pi/8$ according to the information-state value function's gradient. It remains possible for a pursuer to follow an almost straight course by alternating between these two actions. The evader's behaviour is given by a stochastic policy: with equal probability it either moves in its preferred direction or moves in a direction that takes it away from the nearest pursuer (ensuring the problem is non-trivial[7]). If the evader's position is within detection range (24 units) of a pursuer, it is deemed to have been found.

- **Termination.** The maximum trial duration is 256 steps. With a naive implementation in which only one evader trajectory (history) is used in each trial, detection would often be completed within this time limit. However, using the dual particle filter method, the trial will not end until detection takes place in *every* hypothesis for the evader's trajectory. Therefore, often the maximum duration will be reached (but with a diminished number of active particles in each filter.)

- **Returns.** The return at the end of a simulation trial is the proportion of the original probability mass within the evaluation filter that has been eliminated during the course of the trial. This represents the expected return over the 256 parallel scenarios described above, and can be interpreted simply as the probability that the evader

[7] In particular it ensures there is a dependency between the pursuers' actions and the evolution of the hidden state; otherwise it would be possible to pre-compute evader trajectories.

is found within the trial duration. The differential evolution method uses this return as the (stochastic) evaluation of the weight vector associated with that trial.

Figure 2 shows the trajectories of 3 pursuers during a trial as sequences of connected circles. The radius of each circle is the area within which the evader can be detected. A more detailed view of the central circle (radius 400 units) is also given, for a different learning trial. It shows a strategy has been learnt for the 3 pursuers that covers the central part of the space very well. By this stage in the trial the pursuers have moved towards the edge of this central circle, because the evader could have reached these areas. The plotted markers indicate the states of one particle filter at the end of a trial. The grey points indicate particles that have been detected. These are frozen in the locations at which they were detected, for illustration. The outer (black) points represent hypotheses for the current location of an evader that has not yet been captured.

7.2 Configuration of learning system

Differential evolution used a population size of 16 and F was chosen as described in section 1. 256 particles were used in each filter. 4 trials were performed for each policy evaluation (with pursuers at distances of 1125, 1375, 1625 and 1875 from the origin). Therefore 8 trials were needed for each policy comparison, and so 512 policy comparisons were possible in the total budget of 4096 trials. (This is a relatively small number of function evaluations for an evolutionary algorithm.) The best policy within the population was tested at the end of learning, using a set of 512 scenarios, in which the initial pursuer positions were uniformly distributed in the interval [2000,3000]. A baseline performance was obtained using a very simple policy in which $\alpha_{11} = 1$ and all other elements are 0. This causes each pursuer to turn towards the centre of probability mass of the evader's position distribution, taken under the weighting function $H(0, 0)$.

7.3 Results

Table 1: Success rate (%) for different policy types.

N_P	1	2	3	4
Baseline	46 ± 4	64 ± 3	75 ± 4	81 ± 3
Identical	47 ± 7	72 ± 3	80 ± 2	88 ± 2
Joint	NA	73 ± 6	79 ± 3	85 ± 2
Relational	NA	72 ± 6	86 ± 3	93 ± 1

Table 1 shows the performance of the baseline policy and three types of learnt policy as a percentage (likelihood of detecting evader), for different number of pursuers. The error bounds given are at one standard deviation: the standard errors ($n = 500$) are about 20 times smaller so a difference of 2% is significant, using Gaussian statistics.

The task is sufficiently difficult that no method is able to obtain 100% success, even with 4 pursuers. This is not surprising, because the uncertainty in the evader position expands at a rate that increases rapidly with time. The baseline strategy (no learning) provides performance that increases with the number of pursuers: this indicates that the pursuers do not all converge onto the same trajectories even though they are acting independently according to a simple strategy. Introducing learning improves performance

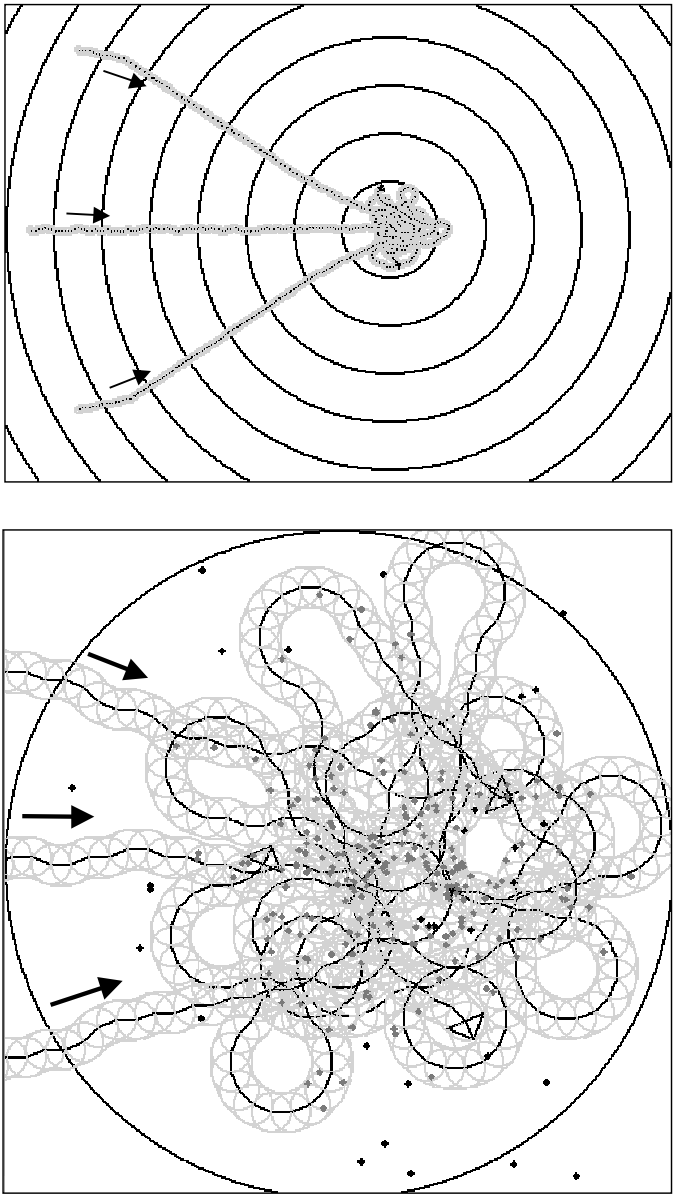


Figure 2: Pursuer trajectories and final particle states in a simulation trial. Full trial (top); detail from a different trial (below).

(significantly) for 2 or more pursuers even though the pursuers continue to ignore each others' positions in their decision-making. The learnt joint policy, in which each pursuer has a separate set of policy parameters performs very similarly: it seems there is no great advantage in allowing each to have a different strategy. There is actually a performance decrease with 4 pursuers, probably because the joint strategy has 32 parameters and so could be expected to take much longer to learn than the identical (8 parameter) strategies.

The learnt relational policies are equally effective (compared with the other learnt strategies) for 2 pursuers, but show major benefits with 3 or more pursuers. This indicates pursuers that are aware of each others' positions can "divide and conquer" the search problem in a more systematic way. The gains that have been obtained are very significant: with 4 pursuers the chance that the evader escapes has been halved (compared with the other learnt policies) and reduced by nearly 3 times compared with the baseline policy.

8 Conclusions

Some information gathering problems such as searching tasks have special structure. Although there is hidden state, and beliefs must be tracked, the actual hidden state does not affect optimal behaviour. In other words, optimal behaviour is conditionally independent of the true hidden state given the belief. Therefore it is possible to evaluate a policy against any number of hidden state trajectories in a single learning trial.

The implementation described here matches the computational burden of agent reasoning (belief revision) and simulation (generating hidden state histories) exactly: a particle filter can be used for each of these processes. Furthermore, these two particle filters are performing the same estimation task. They are kept separate in the learning algorithm only because there are benefits if errors are not correlated. When the learnt strategy is transferred into a real environment (not a simulation) only the particle filter representing agent beliefs would be retained: the second filter is part of the learning algorithm rather than the agent itself.

When learning policies for multiple homogeneous agents, it is not necessary to learn separate policies for each agent (the joint policy option); however it is important that the agents make use of relational information (e.g. relative position) in order to divide the workload effectively. The "relational policy" achieved this by applying the same set of basis functions (used for obtaining evader information features) to represent the relative positions of the other pursuers. The deterministic structure of the pursuers' dynamics was also exploited in the policy formulation, that might otherwise have required a state-action value function to be estimated.

Acknowledgements

This research was funded by the UK Ministry of Defence Corporate Research Programme in Energy, Guidance and Control.

References

Doucet, A., de Freitas, J. F. G., and (Eds.), N. J. G. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.

Roy, N. and Gordon, G. (2002). Exponential family PCA for belief compression in POMDPs. In *Advances in Neural Information Processing Systems*.

Storn, R. and Price, K. (1995). Differential evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces. Technical Report TR-95-012, International Computer Science Institute, Berkeley, CA.

Strens, M. J. A. and Moore, A. W. (2001). Direct policy search using paired statistical tests. In *Proceedings of the 18th International Conference on Machine Learning*, pages 545–552. Morgan Kaufmann, San Francisco, CA.

Strens, M. J. A. and Moore, A. W. (2002). Policy search using paired comparisons. *Journal of Machine Learning Research*, 3:921–950.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning*. MIT Press, Cambridge, MA.

Thrun, S. (2000). Monte carlo POMDPs. In Solla, S., Leen, T., and Müller, K., editors, *Advances in Neural Information Processing Systems 12*, pages 1064–1070. MIT Press.