# Using Reinforcement Learning to Guide the Development of Self-organised Feature Maps for Visual Orienting

Kevin Brohan[1], Kevin Gurney[2], and Piotr Dudek[1]

[1] The University of Manchester,
School of Electrical and Electronic Engineering,
Manchester M13 9PL, United Kingdom
kevin.brohan@postgrad.manchester.ac.uk,
p.dudek@manchester.ac.uk
[2] University of Sheffield,
Department of Psychology,
Sheffield S10 2TP, United Kingdom
k.gurney@sheffield.ac.uk

**Abstract.** We present a biologically inspired neural network model of visual orienting (using saccadic eye movements) in which targets are preferentially selected according to their reward value. Internal representations of visual features that guide saccades are developed in a self-organised map whose plasticity is modulated under reward. In this way, only those features relevant for acquiring rewarding targets are generated. As well as guiding the formation of feature representations, rewarding stimuli are stored in a working memory and bias future saccade generation. In addition, a reward prediction error is used to initiate retraining of the self-organised map to generate more efficient representations of the features when necessary.

**Keywords:** saccade, oculomotor system, visual search, action selection, system model, self-organised map, internal representation, saliency.

## 1 Introduction

Artificial behaving systems must address a number of problems related to action selection and the organisation of knowledge. In particular, we are interested in establishing how an agent can develop useful internal representations of world-related information acquired via sensory inputs and potential rewards associated with various actions; how it can use this information to decide what action to perform next and how can it alter its strategies to adapt to changing circumstances in the outside world? Previous models, which address the learning of internal representations of the behavioural context in which actions take place, include the models of Dominey [1], Cisek [2] and Wilimzig et al. [3].

We address these issues within the context of a visual search task. A virtual eye explores a 2D scene containing a large number of cues by foveating to one cue at a time (Fig. 1). At any point in time, the retinal image contains many cues, and the system must determine which of these cues will become the next target. Some cues are

K. Diamantaras, W. Duch, L.S. Iliadis (Eds.): ICANN 2010, Part II, LNCS 6353, pp. 180–189, 2010.

associated with a reward, though the rewarding cues may change over time, and the task is to learn which categories of cue are rewarding and to foveate to them as often as possible.

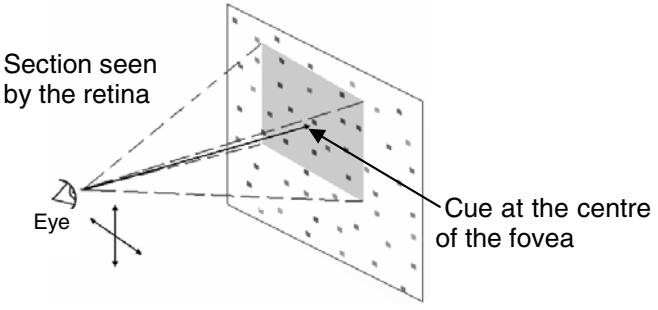


**Fig. 1.** Schematic of the experiment: the retina views a section of the world image at any point in time, making saccades so that a cue always falls on the centre of the fovea (solid black arrow). The target of the saccade may be rewarding, which trains a short term memory and biases the eye to search for similar cues.

The system model is broadly inspired by the visual pathways for saccade generation in the mammalian brain. In the model, low-level feature extraction is followed by a split in processing between two streams: a colour-sensitive 'what' pathway which subserves feature and object detection and a monochromatic 'where' pathway which subserves spatial processing. This scheme is related to that for biological visual processing, in which simple features are first extracted in visual area 1 (V1): these signals are then passed to a dorsal stream devoted largely to spatial (and motion) processing, and a ventral stream which largely subserves feature recognition (which may include a colour dimension) culminating in object identification in inferotemporal cortex (IT) [4]. In the primate brain, neuronal tunings in the dorsal stream show a retinotopic organisation [5][6] and there is evidence for topographic organisation of neuronal responses in IT such that neurons which respond to similar features are located close together in cortex [7]. We deploy similar signal representations in our model. There is also evidence that information from both ventral and dorsal streams is combined into a variety of salience maps (for example, in frontal eye fields) which represent candidate targets for saccades. The competition between these possible targets is resolved in looped circuits through the basal ganglia before the signal is expressed in the superior colliculus which, in turn, drives saccades via the saccadic generator in brainstem [8]. Behaviourally relevant information is maintained in prefrontal working memory [9] and there are several processes devoted to biasing saccades from previous targets under so-called inhibition of return (IOR) [10].

In our model, the action selection problem is addressed through a saliency map that combines bottom-up processing of sensory cues with an IOR mechanism and top-down memory signals that bias saccades towards features that are expected to be rewarding. The model has two novel features: firstly, sensory cues are internally represented in a self-organised map (SOM) of feature space and a working memory of rewarding cues is topographically projected onto this map. Secondly, we bias the development of the

SOM with a reward prediction error signal. The development of the SOM is thus modulated to facilitate the efficient allocation of limited computational resources towards resolving uncertainty in the predicted reward associated with cues.

In the next section, the operation of the model will be overviewed. Implementation details will be given in Section 3. Section 4 contains results of simulations and discussion of these results. Conclusions are presented in Section 5.

## 2 Model Overview

The overview of the model is presented in Fig. 2. The world image $\mathbf{S}^{\mathbf{W}}$ is composed of a large number of colour cues on a black background. The retina contains a subwindow of the world image, as determined by the gaze coordinates. The retinal image is split into the luminance channel $\mathbf{R}^{\mathbf{L}}$ which is processed in the retinotopic space ('where' pathway) and three colour channels $\mathbf{R}^{\mathbf{R}}$, $\mathbf{R}^{\mathbf{G}}$, $\mathbf{R}^{\mathbf{B}}$, which together form a retinotopic feature vectors $\mathbf{C}$ and are further processed in the feature space ('what' pathway). It should be noted that we chose to operate with colour cues for simplicity, but retinotopic maps of other features (e.g. orientations etc.) could be considered.

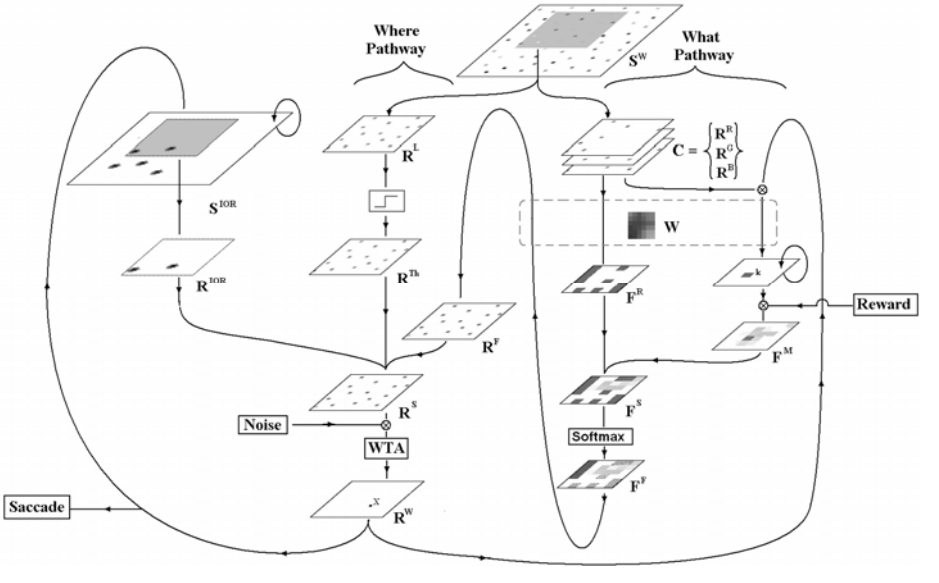


**Fig. 2.** Schematic of the neural network: the retinal image is divided into feature-based and retinotopic channels. See text for details.

The feature vectors $\mathbf{C}$ from each retinal location are classified using a self-organised map $\mathbf{W}$, and results are pooled into a feature map $\mathbf{F}^{\mathbf{R}}$ which develops an internal representation of the observed stimuli. A small number of nodes are used in the SOM in order to force the inputs to compete for representation on the map. The SOM can reorganise and its plasticity is modulated by the reward signals, so that behaviourally relevant features are given more precise representations. The balance between the stability

and adaptability of the representation is determined by the overall success of the system in predicting the reward associated with various cues.

In the task, a reward is given for making saccades to cues of some determined colours. A working memory ($\mathbf{F}^{\mathbf{M}}$) is topographically mapped onto $\mathbf{F}^{\mathbf{R}}$ and activity in this layer represents reward expectation for a feature at the corresponding location in the $\mathbf{F}^{\mathbf{R}}$. The feature saliency map $\mathbf{F}^{\mathbf{S}}$ is developed combining bottom-up classification of features in the scene by $\mathbf{F}^{\mathbf{R}}$ and top-down memory of rewarding features $\mathbf{F}^{\mathbf{M}}$. The $\mathbf{F}^{\mathbf{F}}$ layer "sharpens" the activity of the $\mathbf{F}^{\mathbf{S}}$ layer such that the activity of the most active units is increased. The "sharpened" feature saliency $\mathbf{F}^{\mathbf{F}}$ is then transformed back to the retinotopic coordinate frame, providing a map $\mathbf{R}^{\mathbf{F}}$, which associates the reward expectation assigned to the features with their locations.

The visual saliency map $\mathbf{R}^{\mathbf{S}}$ is formed by combining the feature-based saliency map $\mathbf{R}^{\mathbf{F}}$, a bottom-up cue-background separation map $\mathbf{R}^{\mathbf{Th}}$ (obtained in the 'what' pathway from the luminance map $\mathbf{R}^{\mathbf{L}}$) and inhibition of return map $\mathbf{R}^{\mathbf{IOR}}$. The activity in the visual saliency map $\mathbf{R}^{\mathbf{S}}$ represents the reward anticipation at a given cue location. Competition between locations is resolved stochastically and the next saccadic target ($\mathbf{R}^{\mathbf{W}}$) is determined. The saccadic target is classified by the self-organised map ($\mathbf{F}^{\mathbf{W}}$), and if a saccade is made to a rewarding target, the activity in the corresponding location in $\mathbf{F}^{\mathbf{M}}$ is increased. If a target is unrewarding, then the activity in that region of $\mathbf{F}^{\mathbf{M}}$ is reduced.

After a saccade has been made, the target location is activated on a world-centric $\mathbf{S}^{\mathbf{IOR}}$ map which prevents the eye from returning to that position for a short period of time.

## 3 Implementation Details

In the first stage of processing, previously visited locations are inhibited by using current gaze coordinate information to map the spatial inhibition field ($\mathbf{S}^{\mathbf{IOR}}$) onto retinotopic coordinates ($\mathbf{R}^{\mathbf{IOR}}$). The inhibition of return world map $\mathbf{S}^{\mathbf{IOR}}$ contains traces of previously visited locations. After a saccade is made, the activity in $\mathbf{S}^{\mathbf{IOR}}$ is increased with a Gaussian profile centred at the target location. $\mathbf{S}^{\mathbf{IOR}}$ is implemented as a leaky integrator and the activity decays over time.

The feature map $\mathbf{F}^{\mathbf{R}}$ represents the retinotopic RGB vectors of $\mathbf{C}$ on a self-organising 2D map. The function of the $\mathbf{F}^{\mathbf{R}}$ layer is to pool the feature classification results across a retinotopic array of shared-weight SOM classifiers. As in the classical SOM [12], the neurons of the $\mathbf{F}^{\mathbf{R}}$ layer are tuned to preferred input vectors ($\mathbf{W}$). For each retinal location $i$, the best matching unit (BMU) is calculated as the location $j$ in the SOM with the shortest distance between its preferred vector $\mathbf{W}_j$ and the feature vector $\mathbf{C}_i$. The results are pooled across the retina so that the activity of neuron at location $j$ in $\mathbf{F}^{\mathbf{R}}$ is equal to 1 if a feature is present and 0 if a feature is not present in the visual scene, regardless of how many retinotopic locations activate the point $j$ in the SOM:

$$\mathbf{F}_j^{\mathbf{M}} = \begin{cases} 1 & \text{,if } \exists\, i \ : j = \arg\min_m \left\{ \left\| \mathbf{C}_i - \mathbf{W}_m \right\| \right\} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The activity of the feature saliency map $\mathbf{F}^{\mathbf{S}}$ is calculated as the summation of the activities of the feature memory $\mathbf{F}^{\mathbf{M}}$ and the $\mathbf{F}^{\mathbf{R}}$ layer:

$$\mathbf{F^S} = \mathbf{F^M} + \mathbf{F^R} \tag{2}$$

The $\mathbf{F^S}$ activity is then mapped back to retinotopic space as $\mathbf{R^F}$, such that the strength of a feature in $\mathbf{F^S}$ is reflected in the retinotopic locations at which this feature is present:

$$R_i^F = F_j^S : j = \arg\min_m \left\{ \left\| \mathbf{C}_i - \mathbf{W}_m \right\| \right\} \tag{3}$$

The top-down modulatory activity ($\mathbf{R^F}$) is combined with the bottom-up cue segmentation (thresholded luminance activity $\mathbf{R^{Th}}$) and the previously inhibited locations ($\mathbf{R^{IOR}}$) to form the combined retinotopic saliency map $\mathbf{R^S}$:

$$\mathbf{R^{Th}} = H\left(\mathbf{R^L} - T\right) \tag{4}$$

$$\mathbf{R^S} = \mathbf{R^{Th}}(1 + \frac{1}{2}\mathbf{R^F}) - \mathbf{R^{IOR}} \tag{5}$$

where $T$ is the luminance threshold and H is the Heaviside step function. To implement stochastic competition, the saliency ($\mathbf{R^S}$) is multiplied by an array of white noise ($\mathbf{N}$) and the maximum activity is selected as the target location $x$ for the next saccade:

$$x = \arg\max_i \left\{ R_i^S N_i \right\} \tag{6}$$

In the next stage, the weights $\mathbf{W}$ of the SOM layer are trained. The winning unit $k$ corresponding to the feature vector at saccade target $x$ is located in the map space by calculating the BMU:

$$k = \arg\min_j \left\{ \left\| \mathbf{C}_x - \mathbf{W}_j \right\| \right\} \tag{7}$$

If a cue is rewarding ($r = 1$), some activity is introduced to the location surrounding $k$ in the learning layer $\mathbf{F^M}$ while if a cue is unrewarding, ($r = 0$) the activity at the point $k$ is reduced:

$$\Delta F_j^L = \begin{cases} G(j,k,A,\sigma_L) & ,r = 1 \\ -B & ,r = 0,\ j = k \\ 0 & ,r = 0,\ j \neq k \end{cases} \tag{8}$$

where $G$ is a Gaussian function in map space centred on neuron $k$ with amplitude A = 1, spread $\sigma_L = 0.25$ and B = 0.25.

The SOM weights are modified such that the distance between each weight $\mathbf{W}_j$ and the input vector of the winning unit $\mathbf{C}_x$ is reduced. The degree to which a weight is modified $\Delta\mathbf{W}_j$ decreases as a Gaussian function of the topographic distance from the BMU in the map space. The strength of the change is modulated by the learning rate $\alpha_t$:

$$\Delta\mathbf{W}_j = \alpha_t \left\| \mathbf{C}_x - \mathbf{W}_j \right\| G(j,k,\sigma_t) \tag{9}$$

where G is a Gaussian function of constant area. The SOM weights are initialised with very small values. In the classical SOM global organisation is generated across

the map by constantly decreasing values of $\alpha_t$ and $\sigma_t$ over time. This slowly 'freezes' the weights across the map and allows increasingly finer details to be represented on the SOM [12].

Since the stimuli and rewards may change over time, we introduce in our model a feature that allows the SOM weights to become plastic again. If the network is unable to predict rewards with sufficient success, we should assume that the internal representation of that stimulus is not sufficiently resolved on the SOM and that it may be useful to 'unfreeze' the SOM weights and to attempt to generate a new representation of the stimulus. The values of $\alpha_t$ and $\sigma_t$ are thus modulated by the reward history, in the following way:

The reward prediction error $\delta$ is a measure of the surprise when an expected reward is undelivered or a reward is received from an unexpected source. It is defined as the absolute value of the difference between the received reward *r* and the expected reward, represented by the activity of the winning neuron in the working memory layer $F_k^M$:

$$\delta(r) = \left| r - F_k^M \right| \tag{10}$$

A non-linear function of the reward prediction error is integrated over time with a leaky neuron of output $\xi$:

$$\xi_t = (1-\gamma)\xi_{t-1} + \gamma \frac{\delta(r)}{1-\delta(r)+\theta} \tag{11}$$

where $\theta$ was equal to 0.1, and g is the leak rate, which was equal to 0.05.

If the value of $\xi$ exceeds a predefined threshold, which occurs at a time $t_{th}$, the variables $\xi$, $\alpha_t$ and $\sigma_t$ are reset to their initial values of 0, $\alpha_0$ and $\sigma_0$. The values of $\alpha_t$ and $\sigma_t$ decrease linearly with decay rate $\lambda$ after reset time $t_{th}$.

$$\alpha_t = \max\{-\lambda(t-t_{th})+\alpha_0, 0\} \quad , \quad \sigma_t = \max\{-\lambda(t-t_{th})+\sigma_0, \varepsilon\} \tag{12}$$

and $\varepsilon$ is a small positive number (1E-5 in this paper) which prevents $\sigma_t$ from decreasing to zero. The time course of the variables associated with the modulation of the SOM during a typical experiment is shown Figure 3.

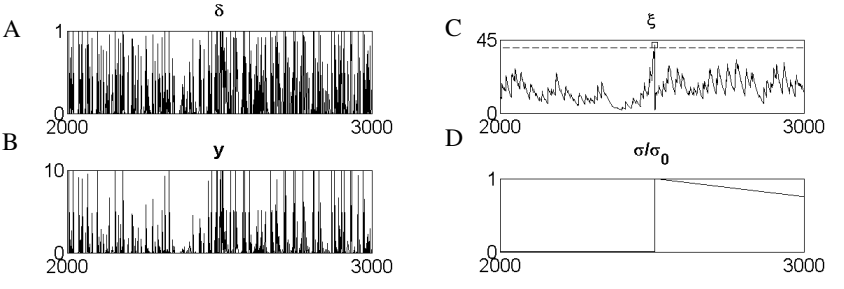


**Fig. 3.** Representation of the modulation of $\alpha_t$ and $\sigma_t$ in the neural network. A) $\delta$ is the reward prediction error; B) y is a non-linear function of the error, which filters out small errors. C) $\xi$ is the leaky integration of *y*, in which the dashed line marks the threshold. D) $\sigma_t/\sigma_0$, $\alpha_t/\alpha_0$ are the modulated learning parameters.

# 4 Results

The model was implemented on a desktop PC in APRON software, a tool for implementing neural simulations on massively parallel processor arrays [13]. In the first experiment, we test the performance of the working memory during a rewarding search task with static SOMs. In the second experiment we examine the effectiveness of the reward-modulated SOM in adapting to a changing reward scenario. In both cases, experiments ran for 10000 time steps, with 10 trials per experiment. The world image is of dimensions 256 x 256 pixels and contains 256 2 x 2 pixel colour cues. The squares were chosen from a set of 64 distinct colours and each colour appears in the world image at four locations. The positions of the cues are random and repeated across trials, and the colours are randomly assigned to the cue positions at the beginning of each trial. The retina views a 128 x 128 pixel region of the world, in which the mean number of cues on the retina at any time is 46.0, with a standard deviation of 14.3.

## 4.1 Experiment I

In the first experiment, SOMs were trained for 200000 steps with constant linear decay throughout, and $\alpha_0$ and $\sigma_0$ values of 0.02 and 0.05 respectively. A new SOM was generated for each trial and no rewards were given in the control experiment. The same SOMs were used during the experiment and the control stages and the SOM did not develop during either the experiment or the control.

There were four different reward epochs in the experiment, each lasting for 2500 time steps. For the first 2500 steps, a reward was presented for foveating to a single orange cue (RGB = {1, .85, 0}). For the second epoch, a reward was presented for foveating to one of two green cues (RGB = {0, 1, 0.07} or RGB = {0, 1, 0.17}). For the third epoch, a reward was present for foveations to one blue cue (RGB = {0, 0.24, 1}). Finally, for the fourth epoch, rewards were given for foveating to one of two magenta cues (RGB = {1, 0, 0.73} or RGB = {1, 0, 0.64}).

Figure 4(A, B) shows the number of saccades made to each colour in this experiment. In the case with working memory (Fig. 4B), saccades were preferentially made to rewarding cues.

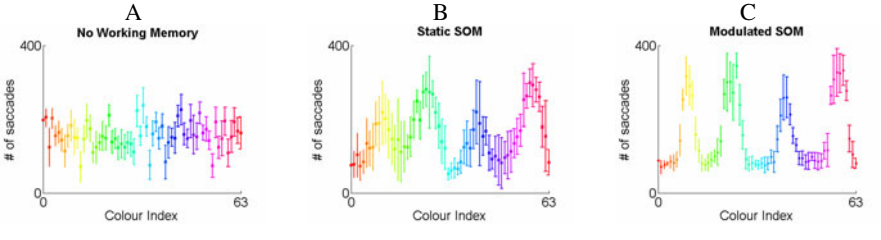


**Fig. 4.** The y axis represents the total number of saccades across 10000 time steps to each possible colour with A) no working memory, B) working memory & static SOM and C) working memory & modulated SOM

Figure 5 (A, B) shows the number of saccades made to each colour in time bins of 100 steps. In the case with working memory (Fig. 5B), saccades were biased towards the rewarding cues, and cues which were similar. In the case without working memory (Fig. 5A), the system had no significant preference for any cue colour.

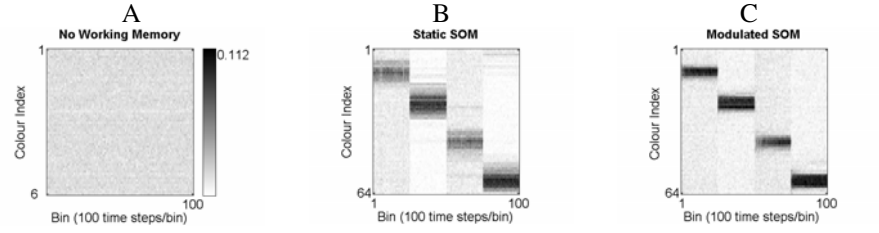


**Fig. 5.** The system learned to distinguish between the rewarding and the unrewarding cues. In the control experiment the rate of detection was approximately the same for all colours. The pixel values represent the average across 10 trials. A) No working memory, B) working memory & static SOM and C) working memory & modulated SOM.

### 4.2 Experiment II

In the second experiment, we investigated the effect of modulating the learning rate $\alpha_t$ and influence $\sigma_t$ online. The results from experiment I are used as the control for this experiment. Figure 4(C) shows the number of saccades made to each cue type in this experiment. The modulated SOM was far more successful at creating a useful representation of the rewarding cues than the static SOM and it shows much better discrimination between rewarding and unrewarding stimuli. Far fewer saccades were being made to unrewarding objects when compared with the control, as is clear from the broadness of the peaks in Fig. 4(B, C). This result is to be expected as the modulated SOM is able to re-train when the rewarding colour changes, allowing the network to re-deploy its limited resources to address the new scenario.

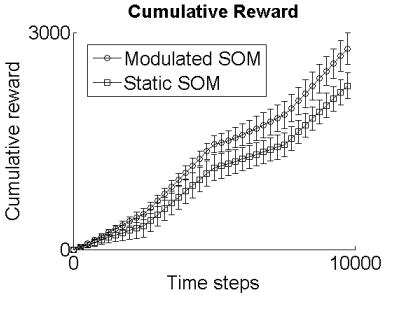


**Fig. 6.** The total number of rewards received over time. The error bars mark one standard deviation across 10 trials. For clarity, every 250th point is plotted.

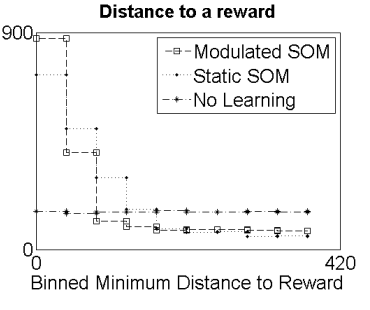


**Fig. 7.** Number of saccades to each colour plotted against the distance between a cue and a reward in RGB space normalised by the number of cues per bin

Figure 5(C) shows the number of saccades made to each colour in time bins of 100 steps. Again, the peaks are more intense and more tightly tuned around the rewarding cues.

When rewarding cues were present on the retinal image, saccades were preferentially made to the rewarding cues during each epoch. Results from the static SOM were similar to those from the modulated SOM, though the static SOM had more difficulty in resolving between rewarding and unrewarding cues. Figure 6 shows the cumulative number of rewards received over time. The modulated model proved far more effective at finding rewarding cues than the static model.

For both experiments, we investigated the mean number of times a saccade was made to a cue as a function of cue distance from the reward in RGB space (Fig. 7). Distances between cue and reward were assigned to 10 bins and the number of cues per bin was counted. The number of saccades per bin was normalised by the number of different cues that were assigned to each distance bin. In the modulated SOM, saccades are more tightly tuned towards rewarding units.

## 5 Conclusions

We have presented a neural network model of working memory on a self-organised feature map as a solution to an action selection problem. The model contains the novel features of a) generating internal representations of stimuli and reward expectation values for different cues and b) using a reward prediction error to remap the SOM when the rewarding stimuli are not sufficiently resolved.

This model is an example of a self-organised map being implemented as part of a larger dynamic neural network system. Rewarding features are learned as activity in the working memory, which is topographically projected onto a self-organised feature map. The working memory biases the system towards saccadic targets which are expected to be rewarding.

By purposely limiting computational resources (restricting the number of nodes in the SOM) we addressed the issue of efficient internal representations of the feature space. The feature maps adapt and self-reorganise to make the model more effective at recognising behaviourally relevant stimuli.

We have demonstrated that the model successfully learns to search for rewarding stimuli, and that the search performance is improved through the action of the modulated SOM.

## Acknowledgment

This work was supported by EPSRC Grant no. EP/C516303.

## References

1. Dominey, P.F.: Complex sensory-motor learning based on recurrent state representation and reinforcement learning. Biol. Cybern. 73, 265–274 (1995)
2. Wilimzig, C., Schneider, S., Schöner, G.: The time course of saccadic decision making: Dynamic field theory. Neural Networks (19), 1059–1074 (2006)

3. Cisek, P.: Integrated Neural Processes for Defining Potential Actions and Deciding between Them: A Computational Model. J. Neurosci. 26(38), 9761–9770 (2006)
4. Mishkin, M., Ungerleider, L.G.: Contribution of Striate Inputs to the Visuospatial Functions of Parieto-Preoccipital Cortex in Monkeys. Behavioural Brain Research 6, 57–77 (1982)
5. Colby, C.L., Goldberg, M.E.: Space and Attention in Parietal Cortex. Annu. Rev. Neurosci., 319–349 (1999)
6. Larsson, J., Heeger, D.J.: Two Retinotopic Visual Areas in Human Lateral Occipital Cortex. J. Neurosci., 13128–13142 (2006)
7. Tanaka, K.: Inferotemporal Cortex and Object Vision. Annu. Rev. Neurosci. 19, 109–139 (1996)
8. Gurney, K., Prescott, T., Redgrave, P.: A computational model of action selection in the basal ganglia. i. a new functional anatomy. Biol. Cybern. 84(6), 401–410 (2001a)
9. Rao, S.C., Rainer, G., Miller, E.K.: Integration of What and Where in the Primate Prefrontal Cortex. Science 276, 821–824 (1997)
10. Posner, M.I., Cohen, Y.: Components of visual orienting. In: Bouma, H., Bouwhuis, D. (eds.) Attention and Performance, vol. X, pp. 552–556. Erlbaum, Mahwah (1984)
11. Sapir, A., Soroker, N., Berger, A., Henik, A.: Inhibition of return in spatial attention: direct evidence for collicular generation. Nature Neuroscience 2, 1053–1054 (1999)
12. Kohonen, T.: Self-Organized Formation of Topologically Correct Feature Maps. Biol. Cybern. 43, 59–69 (1982)
13. Barr, D.R.W., Dudek, P.: APRON: A Cellular Processor Array Simulation and Hardware Design Tool. EURASIP J. Adv. Sig. Pr, Article ID: 751687 (2009)