Contextual Learning in the Selective Attention for Identification model (CL-SAIM): Modeling contextual cueing in visual search tasks

Andreas Backhaus, Dietmar Heinke, Glyn W. Humphreys
Behavioural Brain Sciences Centre,
School of Psychology,
University of Birmingham,
Birmingham B15 2TT, United Kingdom,
E-mail: axb388@bham.ac.uk

Abstract

Visual search is a commonly-used paradigm in psychological studies of attention. It is well-known that search efficiency is influenced by a broad range of factors, e.g. the featural similarity between targets and distractors [4] or the featural configuration (see [16] for a review). Recently, a series of paper by Chun and colleagues (see [1] for a review) has established a new factor the influences search termed 'contextual cueing': visual search is more efficient when targets and distractors are repeated in the same locations across trials, compared with when they fall in new locations. In order to simulate this effect we extended the Selective Attention for Identification model (SAIM [5, 7]) with a mechanism for contextual learning (CL-SAIM). The learning mechanism is based on a Hopfield pattern memory with asymmetric weights. This memory module integrates two functions: On one hand it stores the spatial configuration of search displays, and on the other it improves target detection for already seen displays. In this paper we will demonstrate that this relatively simple extension of SAIM is cable of simulating the experimental findings by [2].

Keywords: Contextual Cueing, perceptual learning, implicit memory, Hopfield memory

1. Introduction

Visual search is a commonly-used paradigm in psychological studies of attention. In this task, a participant has to search for a target item among distracting items and report its presence or absence. The reaction time (RT) is measured and often shows a linear dependency with the number of items in the scene. The intercept and the slope of this linear function are interpreted as indicators of the underlying search mechanism. Shallow search slopes (0-10ms/items) are frequently taken as evidence of a nearly parallel search among objects. In contrast a steep search slope (20-80ms/items) is often interpreted as indicating a serial search through the objects in the field and is named as a inefficient search (See [16] for a review; though see [7, 9] for an alternative view). Several models for visual search have been put forward, for instance, the Guided Search Model [15], the saliency model of Itti and Koch [10], MORSEL [14], SEarch via Recursive Rejection [SERR] [9] and the biased-competition model by Deco [3]. Originally, the Selective Attention for Identification Model (SAIM) set out to model different attentional effects including object-based and space-based attention along with attentional deficits such as visual neglect and extinction. Within this class of findings SAIM covers provides an extensive account of human performance (see [6] for a discussion). Recently, SAIM has been extended by [7] to simulate visual search tasks. In that paper the interaction between top-down and bottom-up factors was examined. In the present paper we focus access the success of the model when applied to a further factor that has recently been shown to affect human search: the spatial context provided by search displays. Chun and Jiang [1, 2] contrasted search for two sets of displays. In one set, all spatial configurations were completely random at any time of the ongoing experiment (the 'new' condition); in the other set, the spatial configurations were randomly chosen at the beginning of the experiment but they were repeated throughout the experiment (the 'old' condition). New and old displays appeared at random. Chun and Jiang [2] reported a difference in RTs between the 'new' and 'old' conditions. Participants were faster on old configurations which is seen as evidence of a memory process which retrieves learned configurations and shifts attention towards the target position more quickly

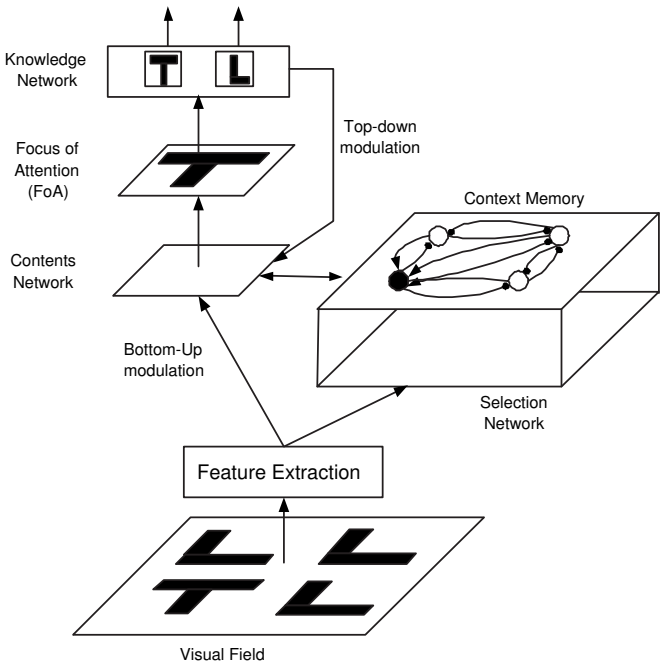

Figure 1. The new architecture of CL-SAIM with a contextual memory built-in. The pattern in the selection network illustrates the result of the learning process. Units at the target (T) position (back circle) receive excitatory activations from the units at distractor (L) positions (white circles), whereas units at target locations inhibit units at distractor position.

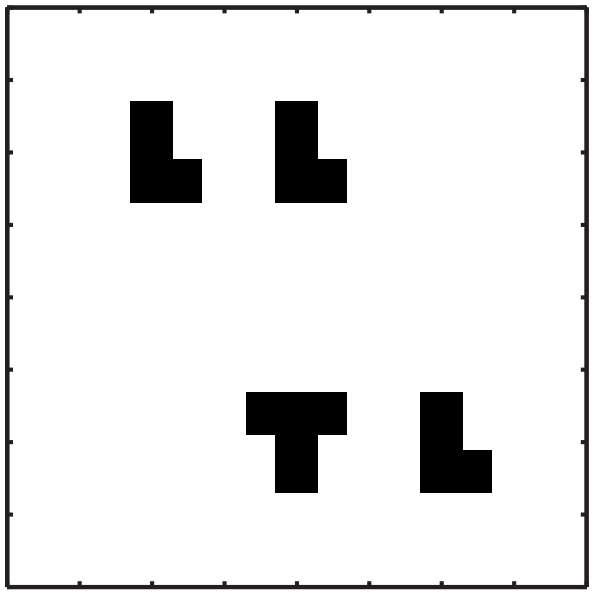

Figure 2. Example search display with four items. Items are arranged in a fixed 3x3 grid

than would otherwise be the case (the ”contextual cueing effect”). This memory process decreased the slope of the search function while it did not affect the intercept. Interestingly, participants were not instructed to memorize displays explicitly and further tests revealed that they could not explicitly discriminate old and new displays. Thus, contextual cueing represents an implicit learning process which allows the memorization of complex information without awareness or intention.

In the present paper we evaluated whether SAIM is able to capture this pattern of reduced search slopes, if a learning mechanism is added to the paper. Such a learning process may provide an important addition to the model, enabling it to generate efficient search when presented with familiar environments. The architecture of this contextual learning version of SAIM (CL-SAIM) is illustrated in Figure 1. The details of the contextual memory will be explained in the following section.

1.1. Overview

Figure 1 gives an overview of SAIM’s architecture and highlights the different influences on the selection process. In a bottom-up path, features (pixel intensity and orientation) are extracted from the visual field and a section is mapped into the Focus of Attention (FOA). This mapping process takes place through the ’contents network’, which itself is modulated by the selection network. By enabling a mapping into the FOA to be achieved from any region where a stimulus falls, this interaction between the content and selection networks enables SAIM to perform translation invariant object recognition. Multiple soft-constrains within the selection network lead to a consistent object mapping, where just one object is selected and mapped in a way that it keeps its original shape. These soft constraints are implemented through competitive and co-operative interactions between units in the selection network. In addition to being activated in a bottom-up manner, SAIM is also sensitive to top-down activity generated via activity in memory templates held in a ’knowledge network’. This top-down activity biases selection towards ’known’ over ’unknown’ objects. These bottom-up and top-down mechanisms were incorporated into the an earlier version of model [5]. In the present version we extended the model by incorporating a learning mechanism into the selection network that could be sensitive to the relative locations of targets and distractors in the displays. This ”contextual memory” learns activa-

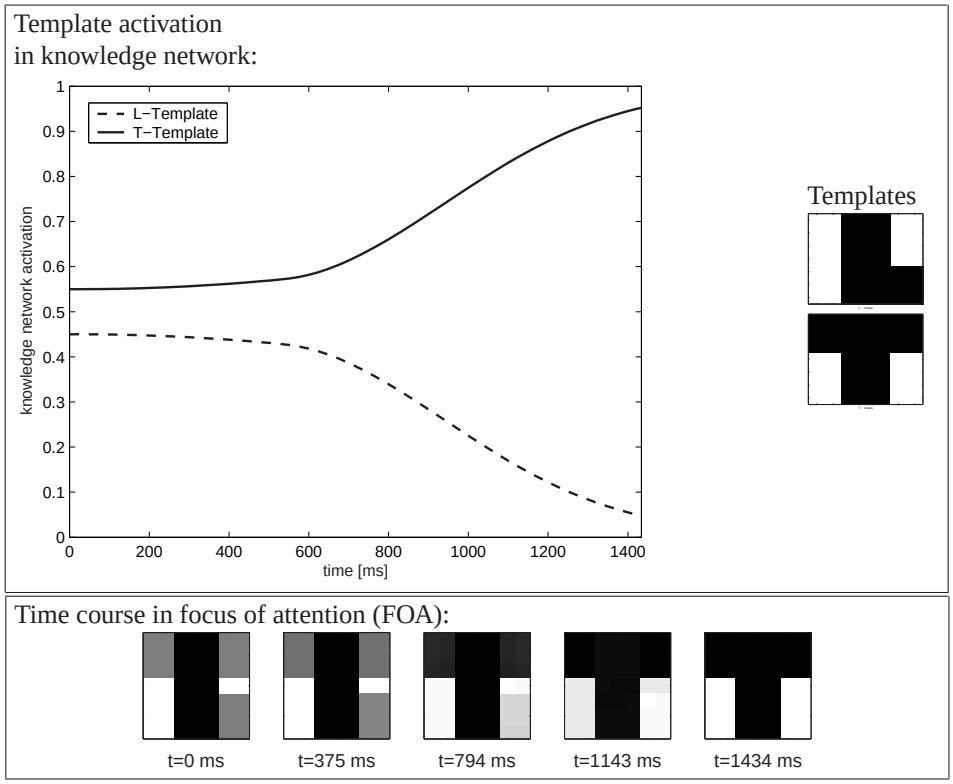

Figure 3. Example of the time course of activation in the knowledge network and the FoA. The T target is selected, based on its template being activated.

tion pattern in the selection network and can retrieve these patterns when similar or identical displays are processed again guiding the selection process. These three modulations (bottom-up input, top-down knowledge of particular objects, learned configural knowledge) influence the competition for selection within the selection network and affect the model's reaction time.

The design of SAIM's architecture follows the idea of soft constraint satisfaction in neural networks based on energy minimization [8]. This minimization approach is developed in the following way: Each module in SAIM aims at a certain end-state of neuronal activation (e.g. just one knowledge network neuron is activated indicating one object to be recognized). These end-states form points of minimal energy in a highdimensional energyfield. These states are also often referred to as attractor states. To ensure that the model as a whole satisfies each constrain expressed by the attractor states, all module energy functions are added together to form a global energy function. A gradient descent methods like proposed in [8] is used to find the minima in the global energy function. In the following section, the energy functions of each module are descried in detail.

1.2. Feature extraction

The feature extraction stage results in a three dimensional feature vector per pixel. One dimension corresponds to the pixel intensity (e.g. the grey value) information, the other two dimensions consist of vertical and horizontal line detectors. The convolution filter mask for a vertical detector was:

$$K_{vert} = \begin{pmatrix} -1 & 2 & -1 \\ -1 & 2 & -1 \\ -1 & 2 & -1 \end{pmatrix} \quad (1)$$

The horizontal filter mask was the transposed version of the vertical filter mask. The feature vector is noted as f_{ij}^n, with indices i and j referring to the locations within the input display and n to the feature dimension. This feature extraction process is an approximation of simple cells responses in V1.

1.3. Contents Network

The energy function for the contents network is

$$E^{CN}(y^{SN}, y^{CN}) = \sum_{ijlmn} (y_{lmn}^{CN} - f_{ij}^n)^2 y_{lmij}^{SN} \quad (2)$$

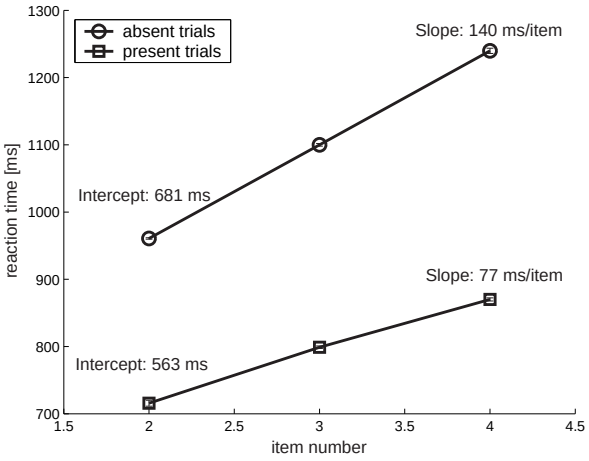

Figure 4. CL-SAIM's reaction time (RT) behaviour in a visual search task for a T as target and L as distractors. The RT slope suggests a "serial search" and shows a RT slope ratio between absent and present displays of 2:1 which is consistent with psychological findings.

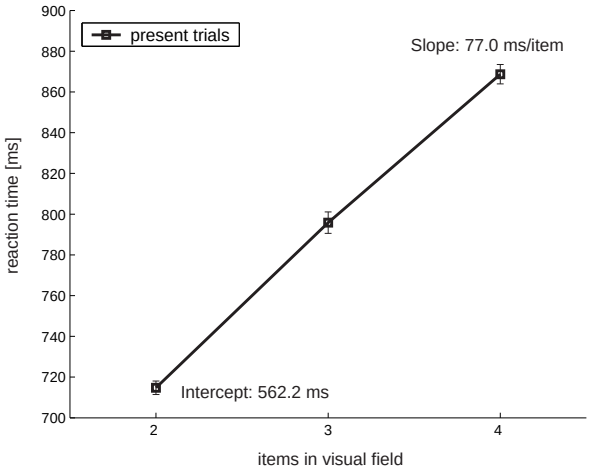

Figure 5. RT behaviour of a search for a T amongst Ls and all displays were 'new'.

where y^{SN}_{lmij} is the activation of a selection network unit connecting the location (i, j) within the visual field with the FOA location at (l, m) and y^{CN}_{lmn} is the activation of units in the contents network. The term $(y^{CN}_{lmn} - f^n_{ij})^2$ forbids states where the contents network does not match the feature values in the input. This term is multiplied with y^{SN}_{lmij} to ensure that the FOA just reflects the region selected by the selection network. Since any location can be routed to the FOA, an object can appear in the FOA regardless of its position in the display; the mapping is therefore translation invariant.

1.4. Selection Network

The mapping process between the input display and the FOA is controlled by the selection network. Here, three constrains must be met to achieve a consistent object mapping:

1. Units in the FOA should receive activity from just one unit in the visual field.

2. Activity from one location in the visual field should be mapped only once into the FOA

3. Neighbourhood relations between units in the input display should be preserved throughout the mapping process.

The first and second constraints are modelled with the equations

$$E^{SN1}_{WTA}(y^{SN}) = \sum_{ij} \left(\sum_{lm} y^{SN}_{lmij} - 1 \right)^2 \tag{3}$$

$$E^{SN2}_{WTA}(y^{SN}) = \sum_{lm} \left(\sum_{ij} y^{SN}_{lmij} - 1 \right)^2 \tag{4}$$

These capture Winner-Take-All behavior, suggested in [13]. The third constraint is implemented via the neighbourhood function:

$$E^{SN3}_{neightbour}(y^{SN}) = -\sum_{ijlm} \sum_{\substack{s=-S \\ s\neq 0}}^{S} \sum_{\substack{r=-R \\ r\neq 0}}^{R} g_{sr} y^{SN}_{lmij} y^{SN}_{\substack{i+r,j+s \\ l+r,m+s}} \tag{5}$$

with g_{sr} being defined by a Gauss function

$$g_{sr} = \frac{1}{A} e^{-\frac{s^2+r^2}{\sigma^2}} \tag{6}$$

where A is a normalization factor. This function leads to the activation of units which map neighbouring locations units the one currently selected into the FOA and thereby support a consistent mapping where object shape is preserved.

1.5. Knowledge Network

The energy function for the knowledge network is defined as following

$$E^{KN}(y^{KN}, y^{CN}) \quad = \quad a^{KN} \left(\sum_k y^{KN}_k - 1 \right)^2$$

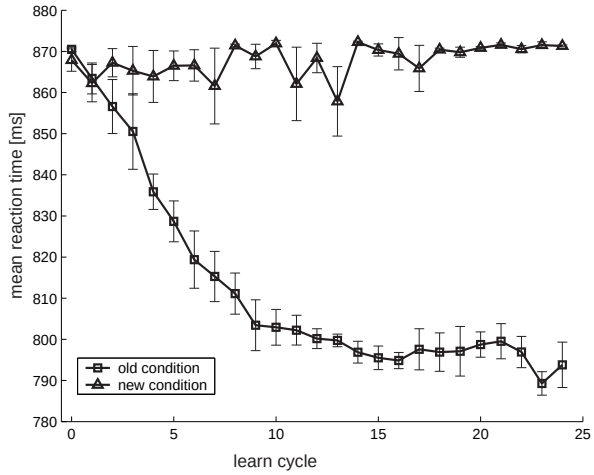

Figure 6. RT behaviour for "old" and "new" displays with a fixed number of items.

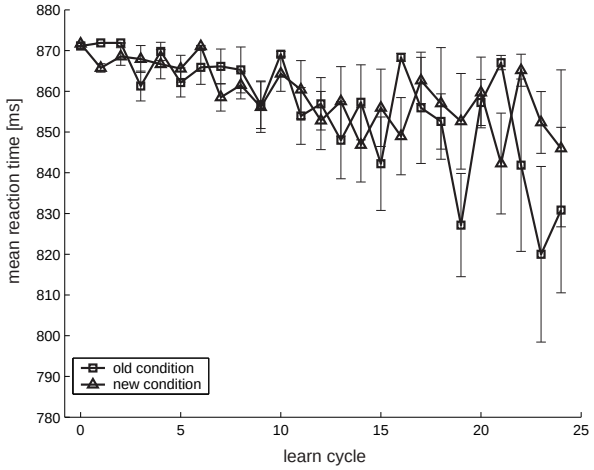

Figure 7. RT behaviour for the "old" and "new" displays when target and distractor positions are swaped in repeated "old" displays.

$$+ \quad b^{KN} \sum_k \left(I_k - \overline{I}\right) y_k^{KN} \quad (7)$$

$$\overline{I} = \frac{1}{K-1} \sum_{\substack{k' \\ k' \neq k}}^{K} I_{k'}$$

$$I_k = \sum_{lm} \left(y_{lmn}^{CN} - w_{lmn}^k\right)^2$$

where the index k refers to template units whose templates are stored in the weights (w_{lmn}^k). K is the total number of templates in the model. The WTA term $\left(\sum_k y_k^{KN} - 1\right)^2$ restricts the knowledge network to activate only one template unit. The term $\sum_k \left(I_k - \overline{I}\right)$ ensures that the sum of the knowledge network output is always one and $I_k = \sum_{klmn} \left(y_{lmn}^{CN} - w_{lmn}^k\right)^2$ ensures that the best-matching template unit is activated. a^{KN} and b^{KN} weight the constraints against each other.

1.6. Contextual Memory

The contextual memory aims at mimicking the experimental effects of contextual cueing. The key finding in contextual cueing shows that the search for known spatial configurations of items is faster than search for unknown configurations. To store spatial configurations in SAIM the selection network of SAIM was extended by a Hopfield pattern memory for spatial configurations of activations

$$E^{CM}(y^{SN}) = -\sum_{uvop} w_{uvop} y_{ccuv}^{SN} y_{ccop}^{SN} \quad (8)$$

where w_{uvop} are the weights storing the spatial patterns and y_{ccuv}^{SN} the activation of the selection network's centre layer. The pattern memory is only connected to the centre layer and not to the whole selection network, as the time course of activation in the center layer is sufficient to determine the spatial configuration of items in a display.
The weight update rule is a modified covariance learning rule [12]

$$w_{uvop}(k+1) = w_{uvop}(k) + \Delta w_{uvop}(k) \quad (9)$$

$$\Delta w_{uvop}(k) = \eta \left(y_{ccuv}^{SM} - \overline{y}_{ccuv}^{SM}\right) \left|y_{ccop}^{SM} - \overline{y}_{ccop}^{SM}\right| \quad (10)$$

where k are the iteration steps during the selection process. At the beginning of the simulations weights are initialized to zero and the weights are up-dated through batch learning. Following this learning method the weight changes are accumulated during the selection process and come only into effect after the selection process is terminated. The averages over time $\overline{y}_{ccij}^{SM}$ and $\overline{y}_{ccij}^{SM}$ are computed by a 'sliding average':

$$\overline{y}_{ccij}^{SM}(k+1) = \frac{k}{k-1} \overline{y}_{ccij}^{SM}(k) + \frac{1}{k} y_{ccij}^{SM}(k) \quad (11)$$

It is important to note that the expression $\left(y_{ccij}^{SM} - \overline{y}_{ccij}^{SM}\right)$ is mainly positive at selected (target) positions and mainly negative at deselected (distractor) positions, because at the target position activation increases and at distractor positions activation decreases. Now if during the selection processes a weight connects a distractor position to a target position, the weight changes are positive. If the weight connects a distractor position to a target position, the weight

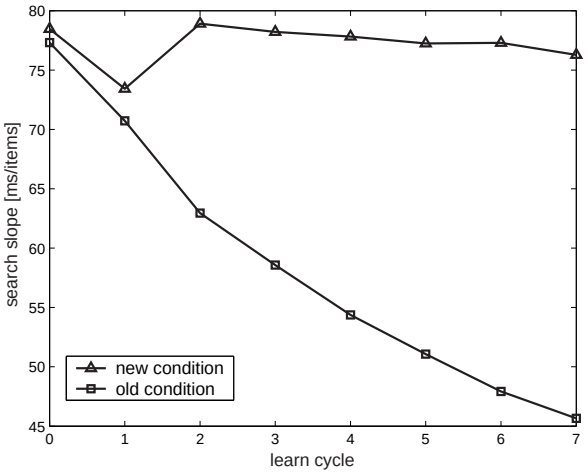

Figure 8. Search slopes for "old" and "new" displays with 2, 3 or 4 items.

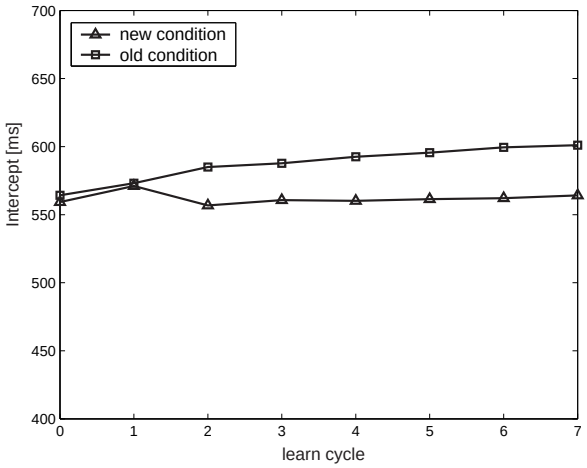

Figure 9. Search intercepts for "old" and "new" displays with 2, 3 or 4 items.

changes are negative. Negative weight changes also apply to links between distractors. As a consequence of the learning rule, the stored target-distractor configurations support the target's selection and a suppression of the distractor selection. Also distractor-distractor relations are suppressed. Fig. 1 depicts the positive and negative weight changes through arrows and line endings with bullet points respectively.

This learning rule is potentially capable of reproducing the basic findings of [2], as the selection was found to improved when known target-distractor relations were presented. However, it is unclear at that stage if all experimental findings in [2] (see introduction) can be simulated with this simple extension. This question will be explored in the following section.

2. Simulations

Before we attempted to simulate the contextual cueing experiments by [2], we chose input display and parameters for SAIM, so that it was able to simulate the basic results for the search task used by [2]. These parameters were kept the same throughout the following simulations. The results on setting up this baseline behaviour are reported in the first section. In the following section we assess four experiments from Chun et al.'s paper and will demonstrate that these findings can be simulated. In the discussion section we will examine the explanation for the simulation results and give an outlook on further simulations.

2.1. Visual Search

In their experiments, Chun et.al [2] used a visual search task where participants had to search for a T among rotated L-shaped items. Items were randomly colored and the L-shaped distractors were rotated with four different angles. This visual search task is considered as inefficient at a slope around 70ms/item and an intercept around 580 ms. Only target present trials were used. However, search through such a display is considered to be highly inefficient. Therefore, it would be expected to find a reaction time ratio between absent and present trials of approx. 1:2. These values (the slope, intercept and present-absent ratios) constrained our choice of parameters. As SAIM's processing is neither rotation-invariant nor does it include colour, we used non-rotated black L and black T items. Also to reduce computer time the displays contained only two, three or four items. For each display type ($items \times absent/present$), ten random spatial configurations were created and simulated, 60 in total. A display had a fixed grid of 3x3 possible items positions (see Fig. 2 for an four item example). Figure 3 gives an example for the time course of activations in the knowledge network and the FOA with a four item display. The complete simulation results are shown in Figure 4. The search slope was approx. 77 ms/object and the intercept was approx. 563 ms. The RT slope was similar to the experimental findings by [2] as well as the slope ratio of absent/present trials indicating a an inefficient search. For the target present trials, Figure 5 shows an additional set of simulations with learning in the contextual memory. The reaction time behaviour did not change substantially. These simulations mimicked the situation in a standard vi-

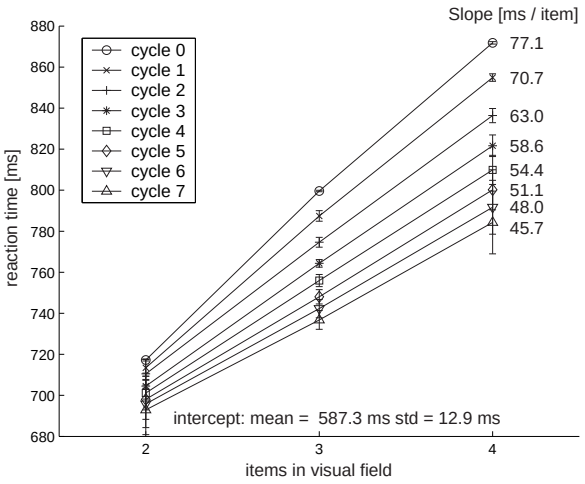

Figure 10. The efficiency in the visual search task increases with learning, as indicated by a decrease in search slope for "old" displays.

sual search task with its complete random display configurations, where the context in a display is not predictive of the target location. This simulation used the same learning parameters as in all following simulations.

2.2. Simulation 1: Displays with fixed item number

In their first experiment, Chun and Jiang [2] investigated the learning effect when a set of displays (the 'old' condition) were repeated among other random displays (the 'new' condition). All displays in this experiments consisted of 12 items. One learning cycle consisted of 12 different 'old' and 12 'new' displays. There were 30 learn cycles in total (later arranged in six epochs with five cycles in each epoch to increase statistical power). The target positions for the repeated and the novel display groups were mutually exclusive and there were no restrictions for distractor positions. The results revealed a significant difference between reaction times to 'old' and 'new' displays with 'old' displays becoming increasingly fast as the number of epochs increased, until RTs converged over time.
For the simulation of these experimental findings, four displays for the 'old' condition were generated and repeated for 15 learning cycles. In every cycle, every 'old' display was paired with a 'new' display. Figure 6 shows how the RT for the 'new' and 'old' stimuli evolved over the learning cycles. With repetition the old displays gained a RT benefit over the new displays. There was also a saturation effect in the RT as the epochs increased.

2.3. Simulation 2: Swapping target and distractor positions

In their second experiment, Chun and Jiang examined whether that the benefit in the 'old' condition is a form of a low-level repetition priming, rather than associative learning between target location and the distractor context. According to the notion of low-level repetition priming the global spatial configurations of the displays may be learned, and the perceptual processing of all terms in repeated displays is facilitated. According to the associative learning hypothesis, search is facilitated by participants, more specifically learning the spatial relations between the target and the distractors. Through learning, the positions of the distractors become predictive of the positions of targets. In their experiment, Chun and Jiang altered the predictive nature of the context in the 'old' displays. Throughout the repeat-trials, the target position in a display was swapped with one of the distractors. Which distractor was chosen was equally likely, so there was no longer a consistent context between the target and the distractors. In other aspects the procedure was the same as in the previous experiment. Chun and Jian found that target-distractor swapping eliminated the contextual learning.
In our simulation of this result the display settings and arrangements were the same as in the previous simulation, however target and distractor positions were swapped to match the psychological experiment. Figure 7 shows how the RTs for the 'old' and 'new' conditions evolved over time. In contrast to the previous simulation, the old displays no longer benefitted, and there was no evidence for contextual learning. These findings are consistent with the psychological experiment. The small overall learning effect stems from the limited number of 'old' displays with swapped target-distractor positions that can be generated for the given display size. Due to this limitation CL-SAIM begins detecting regularities in the statistics of the displays leading to an overall learning effect.

2.4. Simulation 3: Displays with variable item numbers

In their last experiment, Chun and Jiang looked for more support for the assumption that context guides spatial attention towards the target and that the learning effect is not due to perceputal priming or response priming. They argued that, if perceptual or motor priming was involved the RT intercept should change significantly between the old and new conditions. If on the other hand, the learning effect was due to cueing attention towards the target location, there should only be a change on search slopes. The same methods as in the first experiment were used except different numbers of items in the search display: 8, 12, or 16 items. Chun and

Jiang reported a significant effect on the search slope for both 'old' and 'new' displays with search slopes for 'old' displays always being smaller. Furthermore, slopes became shallower as the learning process increased. There was no significant change of the intercept throughout the learning process, and also no significant difference in the intercept for 'new' and 'old' displays.
For our simulation, three item set sizes were used: 2, 3, or 4. For every set size, three different displays were created for the 'old' condition. The number of learning cycles were limited to eight due to the lack of enough two-item display configurations to fill the 'new' condition display set. Figure 8 shows the slopes for the RT-display size function for the 'old' and 'new' conditions. In the case of the 'old' condition, the search slope decreased and the search task became more efficient over time. In contrast, while search in the 'new' condition stayed at its level of efficiency. Figure 9 shows effects on the intercepts of the search functions. This suggests a slight difference between the old and new configurations but this was constant across time. Finally, Figure 10 shows how contextual cueing increases the efficiency of a visual search indicated by shallower search slopes emerging as the number of repetition cycles increased.

3. Discussion

We have demonstrated that the new version of SAIM becomes more efficient in a visual search task through learning the context contextual relations between distractors (CL-SAIM). It was shown that like in [2] with the destruction of this predictive context, the benefit for old over new displays vanished. Thus the effect appears due to associative learning of the target-distractor relation, and not perceptual priming. The learning of selection network activation pattern was achieved with a Hopfield-like memory with asymmetric weights. These weights represent observed covariations in the occurrence of two spatial locations in the selection layer rather than a complete spatial configuration pattern. Further work by Jiang [11] supports the more local nature of learned associations, local in terms of contextual learning in search. She found that, if a display is presented which consists of a mixture of distractor-target pairs from two other learned display, the contextual cueing effect is preserved. This suggests that CL-SAIM not only captures the general pattern of learning in human visual search but also aspects of learning at a microscopic level. The learning mechanism also provides a way for SAIM to become sensitive to spatial contexts in a manner that is independent of the individual examples occupying the context positions, and may have utility for developing selective attention in vision systems in a manner that is sensitive to the familiar positions of objects in scenes.

Acknowledgement

This work was supported by grants from the BBSRC, EPSRC and MRC to Dietmar Heinke and Glyn Humphreys.

References

[1] M. Chun. Contextual cueing of visual attention. *Trends in cognitive science*, 4(5):170–178, May 2000.
[2] M. Chun and Y. Jiang. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, 36(1):28–71, June 1998.
[3] G. Deco and J. Zihl. Top-down selective visual attention: A neurodynamical approach. *Visual Cognition*, 8(1):119–140, 2001.
[4] J. Duncan and G. W. Humphreys. Visual Search and Stimulus Similarity. *Psychological Review*, 96(3):433–458, 1989.
[5] D. Heinke and G. W. Humphreys. Attention, spatial representation and visual neglect: Simulating emergent attention and spatial memory in the Selective Attention for Identification Model (SAIM). *Psychological Review*, 110(1):29–87, 2003.
[6] D. Heinke and G. W. Humphreys. Computational Models of Visual Selective Attention: A review. In G. Houghton, editor, *Connectionist Models in Psychology*, pages 273–312. Psychology Press, 2005.
[7] D. Heinke, G. W. Humphreys, and C. L. Tweed. A New version of the Selective Attention for Identification Model: Top-down Guidance of visual search in Simulations and Experiment. *Visual Cognition*, in press.
[8] J. Hopfield and D. Tank. "neural" computation of decisions in optimization problems. *Biological Cybernetics*, 52:141–152, 1985.
[9] G. W. Humphreys and H. J. Müller. SEarch via Recursive Rejection (SERR): A Connectionist Model of Visual Search. *Cognitive Psychology*, 25:43–110, 1993.
[10] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vison Research*, 40:1489–1506, 2000.
[11] Y. Jiang and L. Wagner. What is learned in spatial contextual cueing: configuration or individual locations? *Perception & Psychophysics*, 66(3):454–463., 2004.
[12] J. Ma. The stability of asymmetric hopfield networks with nonnegative weights. In I. K. K.Y. M. Wong and D. Yeung, editors, *Theoretical Aspects of Neural Computation: A Multidisciplinary Perspective, Proceedings of Hong Kong International Workshop (TANC'97)*, 1997.
[13] E. Mjolsness and C. Garrett. Algebraic transformations of objective functions. *Neural Networks*, 3:651–669, 1990.
[14] M. Mozer and M. Sitton. Computational modeling of spatial attention. In H. Pashler, editor, *Attention*, pages 341–393. Attention, 1998.
[15] J. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1:202–238, 1994.
[16] J. Wolfe. Visual search. In H.Pashler, editor, *Attention*, pages 13–74. Psychology Press, 1998.