Modeling grouping through interactions between top-down and bottom-up processes: The Grouping and Selective Attention for Identification Model (G-SAIM)

Dietmar Heinke[1], Yaoru Sun[1] and Glyn W. Humphreys[1]

School of Psychology
University of Birmingham
Birmingham B15
2TT
United Kingdom

Abstract. We preset a new approach to modelling grouping in a highly-parallel and flexible system. The system is is based on the Selective Attention for Identification model (SAIM) [1], but extends it by incorporating feature extraction and grouping processes: the Grouping and Selective for Identification for Identification model (G-SAIM). The main grouping mechanism is implemented in a layered grouping-selection network. In this network activation spreads across similar adjacent pixels in a bottom-up manner based on similarity-modulated excitatory connections. This spread of activation is controlled by top-down connections from stored knowledge. These top-down connections assign different groups within a known object to different layers of the grouping-selection network in a way that the spatial relationship between the groups is maintained. In addition the top-down connections allow multiple stances of the same objects to be selected from an image. In contrast, selection operates on single objects when the multiple stimuli present are different. This implementation of grouping within and between objects matches a broad range of experimental data on human visual attention. Moreover, as G-SAIM maintains crucial features of SAIM, earlier modeling successes are expected to be repeated.

1 Introduction

SAIM (Selective Attention for Identification Model) is a connectionist model of human visual attention[1]. SAIM's behaviour is controlled by interactions between processing units within and between modules that compete to control access to stored representations for translation invariant object recognition to take place. SAIM gives a qualitative account of a range of psychological phenomena on both normal and disordered attention. Simulations on normal attention are consistent with psychological data on: two-object costs on selection, effects of object familiarity on selection, global precedence, spatial cueing both within

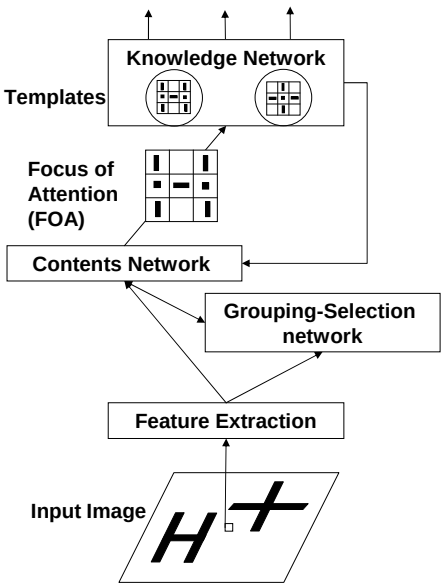

Fig. 1. Architecture of G-SAIM. The depiction of the contents of FOA and the templates of the knowledge network illustrates feature values arranged in a spatial grid. Each grid element in the FOA represents the average feature values of a group, as computed by the contents network.

and between objects, and inhibition of return. When SAIM was lesioned by distorting weights within the selection module, it also demonstrated both unilateral neglect and spatial extinction [2], depending on the type and extent of the lesion. Different lesions also produced view-centred and object-centred neglect, capturing the finding that both forms of neglect can occur not only in different patients but also within a single patient with multiple lesions. In essence, SAIM suggested that attentional effects on human behaviour result from competitive interactions in visual selection for object recognition, whilst neurological disorders of selection are due to imbalanced competition following damage to areas of the brain modulating access to stored knowledge. In [3] we compared SAIM with the most important models on human selective attention (e.g. MORSEL [4], SERR [5], saliency-based models e.g. [6] and biased-competition models, e.g. [7]) and showed that SAIM covers widest range of experimental evidence.
Interestingly, this comparison also highlighted the fact that few models of human visual selection incorporate grouping processes, despite the fact that grouping plays an important role in the processing of visual information in humans (see [8] for a review). SERR [5] implemented a simple grouping process only based on the identify of objects, not taking into account other forms of grouping, e.g. similarity-based grouping. In its original form SAIM too employed very simplistic grouping processes. In particular, units in the selection network supported activity in neighbouring units that could contain a proximal element, leading

to grouping by proximity. Here, we present the results from new simulations showing that SAIM's architecture can be extended to incorporate more sophisticated forms of grouping sensitive to the similarity of the feature present in the display (bottom-up). In addition, experimental evidence shows that grouping is also influenced by top-down factors (e.g. [9]). Therefore, grouping-SAIM does not have a hard-wired coding of specific conjunctions of elements, but rather uses a flexible grouping and selection procedure which operates in interaction between image-based similarity constraints and top-down knowledge imposed by an object recognition system. Moreover, this form of grouping does not only operate within objects but it can also occur across multiple objects, linking separate objects into representations encompassing the whole display. This matches data on human search, where people can respond at the level of the whole display to multiple, homogenous stimuli (e.g. [10]). Interestingly, the approach to grouping and selection in the revised model produce a form of size-invariant object representation as an emergent property.

2 SAIM

2.1 Overview

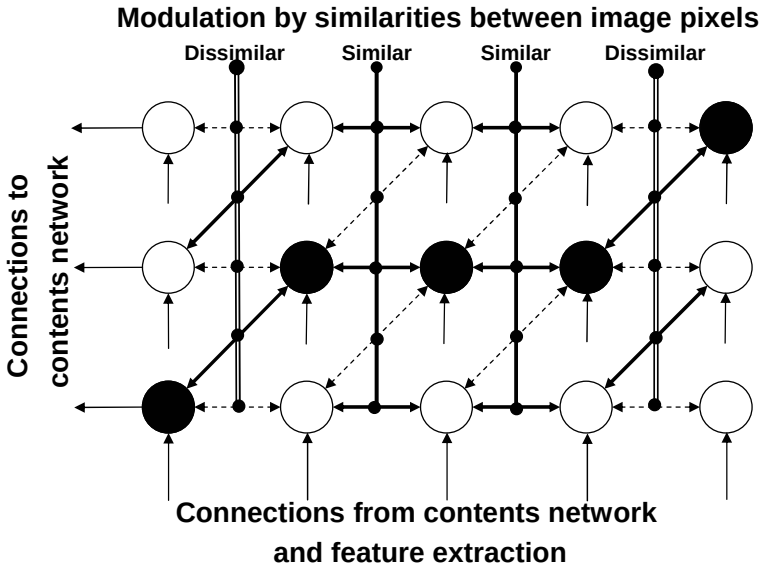

Fig. 2. One-dimensional illustration of the structure and functioning of the grouping-selection network (see text for details).

Figure 1 gives an overview of G-SAIM's architecture and highlights the modular structure of the model. In the first stage features are extracted from the

input image. The contents network maps a section of grouped features into a smaller Focus of Attention (FOA), a process modulated by the grouping-selection network. In addition the mapping of the contents network into the FOA is translation-invariant, enabling G-SAIM to perform translation-invariant object recognition. The grouping-selection network has a multilayered structure where each layer corresponds to a particular location in the FOA. The operation of the grouping-selection network is controlled by competitive and cooperative interactions. These interactions ensure that adjacent locations with similar features are grouped together, whilst adjacent locations with dissimilar features are separated into different groups. Grouped items are represented by conjoint activity within a layer of the grouping-selection network and different groups are represented in different layers. At the top end of the model, the knowledge network identifies the contents of the FOA using template matching. Importantly, in addition to these feedforward functions there are feedback connections between each module. The feedback connection from the knowledge network to the contents network aims at activating only known patterns in the FOA. The connection from the contents network into the grouping-selection network modulates the grouping and selection process so that patterns matching the content of the FOA are preferred over unknown objects.
The design of G-SAIM's network follows the idea of soft constraint satisfaction in neural networks based on "energy minimization" through gradient descent [11]:

$$\tau \dot{x_i} = -\frac{\partial E(y_i)}{\partial y_i} \tag{1}$$

with x_i being the internal activation of the unit i and y_i the output activation of the unit i. Both activations are linked through a non-linear function.
In G-SAIM the "energy minimization" approach is applied in the following way: Each module in SAIM performs a pre-defined task (e.g. the knowledge network has to identify the pattern in the FOA). In turn, each task describes allowed states of activations in the network after the network has converged. These states then define the minima in an energy function. To ensure that the model as a whole satisfies each constraint, set by each network, the energy functions of each network are added together to form a global energy function for the whole system. The minima in the energy function is found via gradient descent, as proposed by [11], leading to a recurrent network structure between the modules. At this point it should be noted that the gradient descent is only one of many possible algorithms that could be applied to find the minimal energy. In our earlier work this approach turned out to be sufficient for modeling psychological data. For technical applications of G-SAIM alternative approaches might need to be considered. In the following sections the energy functions for each network are stated. The global energy function and the gradient descent mechanism are omitted, since they are clearly defined by the subcomponents of the energy function.

2.2 Feature extraction

The feature extraction results in a three-dimensional feature vector: horizontal and vertical lines and the image itself. The lines are detected by filtering the image with 3x3 filters ($\begin{matrix} -2 & +1 & -2 \\ -2 & +1 & -2 \\ -2 & +1 & -2 \end{matrix}$ for vertical lines and its transposed version for horizontal lines). The feature vector is noted as f_{ij}^n hereafter, with indices i and j refereing to image locations and n to the feature dimension. This feature extraction process provides an approximation of simple cell responses in V1. As becomes obvious in the following sections, the use of this simple feature extraction mechanism is not of theoretical value in its own right and arises primilary from practical consideration (e.g., the duration of any simulations). In principle, a more biologically realistic feature extraction process can be substituted (e.g. using Gabor filter).

2.3 Contents network

The energy function for the contents network is:

$$E^{CN}(\mathbf{y^{GN}}, \mathbf{y^{CN}}) = \sum_{ijlm} (y_{lmn}^{CN} - f_{ij}^n)^2 \cdot y_{lmij}^{GN} \tag{2}$$

y_{lmij}^{GN} is the activation of units in the grouping-selection network and y_{lmn}^{CN} is the activation of units in the contents network. Here and in all the following equations the indices l and m refer to locations in the FOA. The term $(y_{lmn}^{CN} - f_{ij}^n)^2$ ensures that the units in the contents network match the feature values in the input image. The term y_{lmij}^{GN} ensures that the contents of the FOA only reflect the average feature values of a region selected by the grouping-selection network ($y_{lmij}^{GN} = 1$). Additionally, since setting an arbitrary choice of y_{lmij}^{GN}s to 1 allows any location to be routed from the feature level to the FOA level, the contents network enables a translation-invariant mapping. It should be noted that the energy function of the contents network results in a feedback connection to the grouping-selection network. This connection provides the grouping-selection network with featural information and its relative positions within the FOA.

2.4 Grouping-Selection network

The mapping from the retina to the FOA is mediated by the grouping-selection network. In order to achieve successful grouping and selection, the grouping-selection network has to fulfill certain constraints when it modulates the mapping process. These constraints are that: (i) the content of a image location should be mapped only once into the FOA; (ii) units whose related image pixels are similar appear in the same layer (implementing similarity grouping) (iii) dissimilar features are routed into separate layers; (iv) neighbouring groups should appear adjacent in the FOA. Figure 2 illustrates the functioning of the grouping-selection

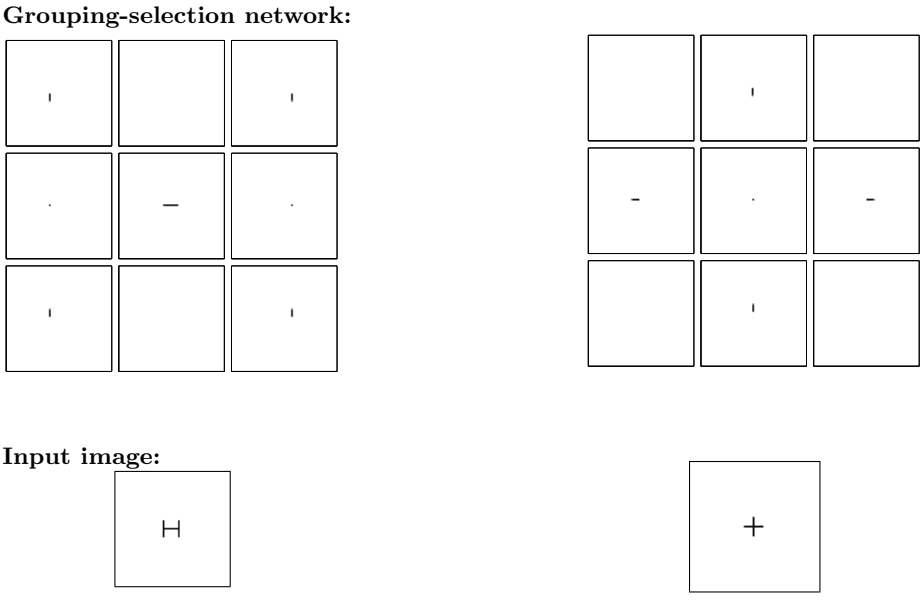

Fig. 3. Simulation results with a single object in the input image. As a result of the grouping mechanism implemented in the grouping-selection network, each layer represents a different group of the H.cross, indicating a grouping within an object.

network. For the sake of clarity only excitatory connections are depicted. The grouping-selection network has a layered structure where each each layer is connected to one node in the contents network. Within each layer and between each layer there are excitatory connections depicted as dotted and bolt lines. These connections are modulated by the similarity between adjacent pixels. The more similar two pixels are the higher is the excitation of the within-layer connection (bolt lines in Fig. 2) and the lower the excitation of the between-layer connections, vice versa.

In the Figure 2 the first two pixels and the last two pixels are dissimilar from each other, whereas the pixels in the middle are similar to each other. Consequently, the first and the last excitatory connection is strong between layers (diagonal) whereas the excitation within the layer is strong in the middle pixels. During the process of energy minimization process in G-SAIM this particular connectivity pattern can lead to an activation of units on the diagonal for the first and the last unit and an activation of units within one layer in the middle part (black circles). Hence, the pixels in the middle are grouped together, whereas the dissimilar pixel are separated from the middle pixels by activating different layers. The similarity between pixels is determined in the following way:

$$s_{ijsr} = h(\sum_{n}(f_{ij}^{n} - f_{i+s,j+r}^{n})^2) \tag{3}$$

$$h(x) = \frac{1}{1 + e^{(a \cdot x + b)}}$$

The parameter for the nonlinearity $g(x)$ is set so that the similarity values range from 0 to 1 (0 = dissimilar; 1 = similar) for given a range feature values.
The energy function for activating similar pixels within a layer of the grouping-selection network is:

$$E^{exc\_sim}(\mathbf{y^{GN}}) = -\sum_{lmij} y^{GN}_{lmij} \sum_{\substack{s=-M \\ s\neq 0}}^{M} \sum_{\substack{r=-M \\ M\neq 0}}^{1} g_{rs} \cdot s_{ijrs} \cdot y^{GN}_{l,m,i+r,j+s} \tag{4}$$

To prevent a region from spreading across the whole image an "inhibitory" energy function was introduced in each layer of the grouping-selection network:

$$E^{inh\_sim}(\mathbf{y^{GN}}) = \sum_{lm} (\sum_{ij} y^{GN}_{lmij})^2 \tag{5}$$

The coefficients g_{rs} drop off in a Gaussian shape to order to reduce the influence of pixels that are further apart. In essence the two equations above represent an implementation of a similarity-modulated Amari-network [12]. Amari showed that under certain conditions gaussian-shaped excitatory connections within a layer lead to contiguous areas of activation. In case of the grouping-selection network the shape of these areas are influenced by similarity in order to implement grouping.
To ensure that dissimilar adjacent pixels are assigned to separate layers, the following energy function was introduced:

$$E^{exc\_dis} = \sum_{lmij} y^{GN}_{lmij} \sum_{\substack{s=-1 \\ s\neq 0}}^{1} \sum_{\substack{r=-1 \\ r\neq 0}}^{1} (1 - s_{ijrs}) \cdot y^{GN}_{l+r,m+s,i+r,j+s} \tag{6}$$

The term $(1 - s_{ijrs})$ is a measure for the dissimilarity between pixels.
The following term prevents units at the same image location to be activated in different layers (constraint (i)):

$$E^{inh}(\mathbf{y^{GN}}) = \sum_{ij} (\sum_{lm} y^{GN}_{lmij})^2 \tag{7}$$

The Constraint (iv) was implemented by connecting all adjacent units which are not connected by similarity modulated connections in an inhibitory way. The energy functions above all include a stable state at zero activation. To prevent G-SAIM from converging into this state an "offset"-term was added to the network:

$$E^{offset}(\mathbf{y^{GN}}) = -\sum_{ijlm} (y^{GN}_{lmij}) \tag{8}$$

At this point the general nature of the four constraints implemented by the grouping-selection network should be noted. The constraints are not specific

about which feature belongs to which layer. The exact assignment of the groups to the different layers in the network is achieved by the top-down influence from the contents network and the knowledge network. These networks feed featural information from known objects back into the grouping-selection network, setting the featural preference of each layer for the groups. The top-down influence, paired with the generality of the grouping constraint, is at the heart of the flexible grouping approach in G-SAIM.

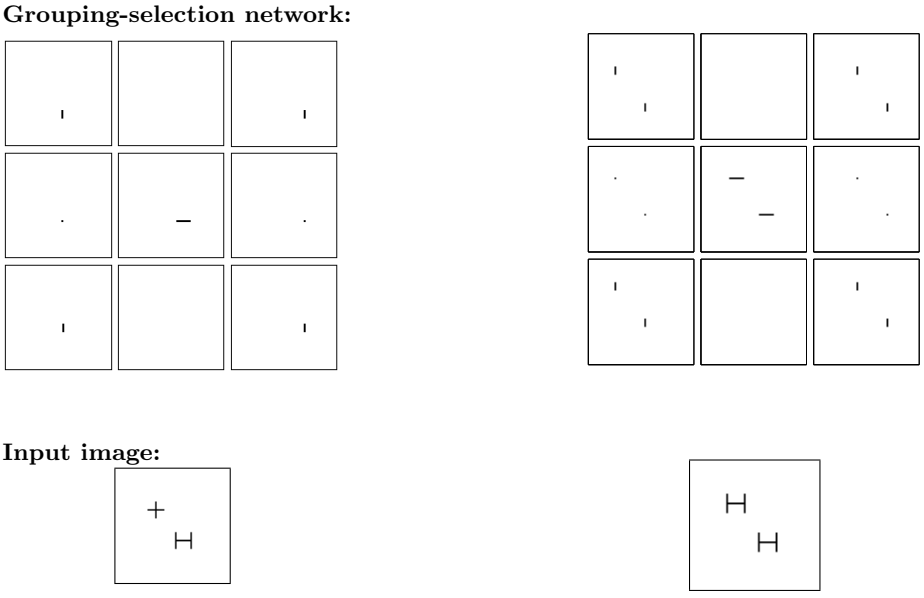

Fig. 4. Simulation results. The result on the left illustrates a case with two different objects (H and cross). The cross is suppressed and the H is selected. The simulation result on the right shows a result of grouping the same object as the two Hs are selected together.

2.5 Knowledge network

The energy function of the knowledge network is defined as

$$E^{KN}(\mathbf{y^{KN}}, \mathbf{y^{CN}}) = a^{KN}(\sum_k y^{KN} - 1)^2 - \sum_k \begin{cases} \sum_{lmn} y_k^{KN} \cdot y_{lmn}^{KN} & \text{If } \mathbf{w^k}\text{'dont select'} \\ \sum_{lmn} w_{lm}^* \cdot (y_{lmn}^{CN} - w_{lmn}^k)^2 \cdot y_k^{KN} & \text{otherwise} \end{cases} \quad (9)$$

The index k refers to the template unit. The term $(\sum_k y^{KN} - 1)^2$ restricts the knowledge network to activate only one template unit [13]. Each grid field

of a template has a flag attached to it indicating if this particular grid field contains information relevant for the object. For instance, in case of the cross the areas between the arms do not belong to the objects; consequently these positions are assigned a "don't select" flag. In future simulations the "don't select"-flag will served to deal with cluttered scenes. In case of an activated "don't select"-flag the term $\sum_{lmn} y_k^{KN} \cdot y_{lmn}^{KN}$ aims at suppressing any activation in the contents network and grouping-selection network. On the contrary, the term $\sum_{lmn}(y_{lmn}^{CN} - w_{lmn}^{k})^2 \cdot y_k^{KN}$ ensures that the best-matching template unit is activated. a^{KN} and b^{KN} weight these constraints against each other.

3 Results and discussion

Two sets of simulations were run to test the grouping mechanism in the grouping-selection network. Two objects were used, a cross and an H. The knowledge network had these two objects as templates (see Fig. 1 for an illustration). In the first set of simulations input images with one object (cross or H) were used (see Fig. 3). The results show that G-SAIM successfully segments parts of the objects (H and cross) into sub-groups, as indicated by the fact that the arms of the H/cross appear in separated layers of the grouping-selection network. This is a direct outcome of the feature extraction of lines which leads to similar feature values along the arms of the H and dissimilar feature values at its cross points. These simulations illustrate the fact that the layers in the grouping-selection network are not connected firmly to a certain feature. For instance, in the simulation with cross as input the centre layer represents the cross point whereas for the H the centre layer encodes the horizontal bar of the H. Compared to other neural network approaches to grouping (e.g. [14]), where groups in one layer for each feature, the grouping process in G-SAIM is both efficient and flexible. Moreover, it allows the model to maintain vital information about the spatial relation between groups within objects in a natural way. For instance, for the H spatial relations between its arms are important information to distinguish it from other objects, e.g. the cross. In G-SAIM this information is simply encoded by assigning the different segments to different layers of the grouping-selection network. It is unclear how such this form of coding could be achieved in a grouping mechanism where is a layer per feature. For instance in the case of the H the arms might be represented in one layer, but then an additional mechanism would be necessary to encode the spatial relations between the parts.
In the second set of simulations images with two objects were used. In these simulations the knowledge network had a slight bias towards activating the template unit for the H. Figure 4 show the simulation results of the grouping-selection network for two input images: on the left a cross together with an H and on the right two Hs. With two different objects, the cross and the H, G-SAIM selects the H and suppresses the cross. Hence, as in old SAIM, G-SAIM still performs object selection. The second simulation examines G-SAIM's behaviour with two Hs in the input image (Fig. 4 right). In this case G-SAIM selected both objects together. The decision concerning whether two objects are selected together or

not is influenced through the constraints implemented in the contents network. If units in the grouping-selection network correspond to the same features in the input image, the activation of these units contribute to the minimization of the energy function. In contrast, if the units correspond to different feature values, the activation in the grouping-selection network is suppressed, as was the case for the cross-H simulation. These two contrasting behaviours, selection of identical objects and competition of different objects, match extensive evidence from human search (e.g. [10]) and from phenomena such as visual extinction in patients with parietal lesions (e.g. [15]).
Finally, it should be noted that the representation of objects in the FOA possesses interesting properties. In the FOA the average feature values for each group are represented. Since the features within the groups are similarly determined by the grouping-selection network, averaging provides a good approximation of the features values within the group. Consequently, when the size of the regions changes and the features in the regions stay the same, the representation in the FOA does not change, leading to a size-invariant representation of objects. Moreover, the size-invariant representation in the FOA is very compact, e.g. it represents the cross through a small number of units.

4 Conclusion

G-SAIM represents the first step towards modeling grouping and selective attention in an integrative approach. The core mechanism for the grouping process is activation spreading, modulated by bottom-up similarity between pixels and top-down influence from object knowledge. G-SAIM simulates the grouping of regions within objects as well as grouping across multiple objects. To the best our knowledge no other models integrate these two forms of behaviour. For instance, the saliency-based approach to selective attention [6] does not have a grouping mechanism. The same is true for the biased-competition approaches (e.g. [7]). In terms of architecture G-SAIM is closest to the dynamic shifter circuits [16], but again this model does not contain a feature extraction and grouping mechanisms. Moreover, other models on grouping utilize similar activation spreading mechanisms (e.g. [14]. However, it is not clear how these models could cope with multiple object situations.
In addition it should be noted that G-SAIM also keeps crucial features of the old SAIM, especially the interaction of competitive and cooperative processes and the multilayer structure of the grouping-selection network. Both elements were responsible for the modeling successes of old SAIM. Hence, we expect G-SAIM to be able reproduce the simulation results of old SAIM, whilst having added capability in simulating grouping effects.
In future work we will aim at replacing the present feature extraction process by a more biologically-plausible approach (e.g. using a Gabor filter) and at simulating psychological data on grouping and attention, especially experimental evidence on the interactions between grouping and attention [9].

Acknowledgment This work was supported by grants from the EPSRC (UK) to the authors.

References

1. 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.
2. K. M. Heilman and E. Valenstein. Mechanisms underlying hemispatial neglect. *Annals of Neurology*, 5:166–170, 1979.
3. D. Heinke and G. W. Humphreys. Computational Models of Visual Selective Attention: A Review. In G. Houghton, editor, *Connectionist Models in Psychology*. Psychology Press, in press.
4. M. C. Mozer and M. Sitton. Computational modeling of spatial attention. In H. Pashler, editor, *Attention*, pages 341–393. London:Psychology Press, 1998.
5. G. W. Humphreys and H. J. Müller. SEarch via Recursive Rejection (SERR): A Connectionist Model of Visual Search. *Cognitive Psychology*, 25:43–110, 1993.
6. J. M. Wolfe. Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994.
7. G. Deco and J. Zihl. Top-down selective visual attention: A neurodynamical approach. *Visual Cognition*, 8(1):119–140, 2001.
8. S. E. Palmer. *Vision Science – Photons to Phenomenology*. The MIT Press, 1999.
9. J. Driver, G. Davis, C. Russell, M. Turatto, and E. Freeman. Segmentation, attention and phenomenal visual objects. *Cognition*, 80:61–95, 2001.
10. J. Duncan and G. W. Humphreys. Visual Search and Stimulus Similarity. *Psychological Review*, 96(3):433–458, 1989.
11. J. J. Hopfield and D.W. Tank. "Neural" Computation of Decisions in Optimazation Problems. *Biological Cybernetics*, 52:141–152, 1985.
12. Shun-ichi Amari. Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields. *Biological Cybernetics*, 27:77–87, 1977.
13. E. Mjolsness and C. Garrett. Algebraic Transformations of Objective Functions. *Neural Networks*, 3:651–669, 1990.
14. S. Grossberg, E. Mingolla, and W. D. Ross. A Neural Theory of Attentive Visual Search: Interactions of Boundary, Surface, Spatial and Object Representation. *Psychological Review*, 101:470–489, 1994.
15. I. Gilchrist, G. W. Humphreys, and M. J. Riddoch. Grouping and Extinction: Evidence for Low-Level Modulation of Selection. *Cognitive Neuropsychology*, 13:1223–1256, 1996.
16. B. Olshausen, C. H. Anderson, and D. C. Van Essen. A Multiscale Dynamic Routing Circuit for Forming Size-and Position- Invariant Object Representations. *Journal of Computational Neuroscience*, 2:45–62, 1995.