Computational modeling of spatial attention

Michael C. Mozer
Mark Sitton

Department of Computer Science &
Institute of Cognitive Science
University of Colorado
Boulder, CO 80309–0430

Appears in: H. Pashler (Ed.), *Attention.* London: UCL Press. 1998. pp. 341-393

Contents

Introduction **1**

What is a computational model? **2**
Connectionist models 2

A basic model of object recognition **4**
Training the model 6
Performance of the model 8

A model of attentional selection **9**
Dynamics of the attentional network 11

Simulations of spatial selection **13**
The benefit of attentional precuing 13
Time course of attention shifts 16
Effect of spatial uncertainty 17
The effect of irrelevant stimuli 18

Attention as a spotlight? **19**

Modeling various selection criteria **23**

The relationship of object-based and location-based attention **24**

Simulations of visual search **27**
Simulation methodology 29
Simple feature search 29
Conjunction search 31
Discussion of visual search 32

The role of selective attention **33**

Contrasting theoretical perspectives on selective attention **34**

Issues in computational modeling **35**
Why build computational models? 35
What makes a model compelling? 36
When is a model right or wrong? 37
What about other models that also explain the data? 37
Depth versus breadth in modeling 37

Acknowledgements **38**

References **38**

Introduction

If we had really huge brains, say the size of watermelons, attention would play a much smaller role in our behavior. Its significance stems primarily from limitations in our processing hardware. We simply do not have sufficient brain capacity to analyze all information that passes through our sense organs, to reason exhaustively about all possible courses of action, and to maintain multiple interpretations of the world. Attentional selection is needed to determine what information will be processed by the available hardware.

Consider the task of recognizing objects in a visual scene. What sort of processing resources would be required to identify all objects in parallel, regardless of their positions, orientations, and size in the scene? If we are familiar with o different objects, and any object can appear in any of p horizontal or vertical positions and r orientations and s scales, the number of different object instantiations is op^2rs. This number would be far larger still if the objects are not rigid. Regardless of the nature of the recognition process, the number of possible object instantiations roughly determines the amount of processing resources required. You can plug in reasonable guesses as to how many object instantiations are possible; 100 million might be a reasonable ballpark figure. If we limit ourselves to one object at a time, however, and the object's position, orientation, and scale are computed first, then the number of object instantiations that have to considered at once is only o, or a number more like 10,000. Ballard (1986) and Tsotsos (1990, 1991) have presented computational complexity analyses of this sort to argue that the combinatorics of vision require some type of attentional selection to reduce the number of possibilities that need to be considered, and that attention can be particularly beneficial when exploiting knowledge of the particular task being faced by the visual system.

In accord with the computational arguments, human vision shows strong limitations on how many objects can be processed and identified in parallel (e.g., Duncan, 1987; Mozer, 1983, 1989; Pashler & Badgio, 1987; Shiffrin & Gardner, 1972; Schneider & Shiffrin, 1977; Treisman & Schmidt, 1982). In general terms, one can conceive of processing of a visual stimulus as occurring along a certain neural *pathway*. If the processing pathways for two stimuli are nonoverlapping, then processing can take place in parallel. But if the pathways cross—i.e., they share common resources or hardware—the stimuli will interact or interfere with one another. One role of attention is to reduce this interference by restricting the amount of information that is processed at once.

In this chapter, we examine the role of spatial attention from a computational perspective. Because the function of attention can be understood only in its relation to visual information processing, we must model not only the attentional system itself, but also the process of object recognition. We begin by presenting a basic model of object recognition, showing that interference prevents the system from reliably processing multiple, complex stimuli, and then we show how a simple mechanism of attentional selection can reduce this interference. Our initial goal will be to present a model that is computationally adequate, that is, a model that has the computational power to perform the sort of visual information processing tasks that people do. Psychologists are most concerned with another issue: whether the model can explain various experimental findings and whether it has any ability to predict the outcome of further experiments. In our view, the demands of computational adequacy and explanatory/predictive power are complementary, and a compelling account should satisfy both, and in so doing, allow one to understand the mechanisms that underlie information processing.

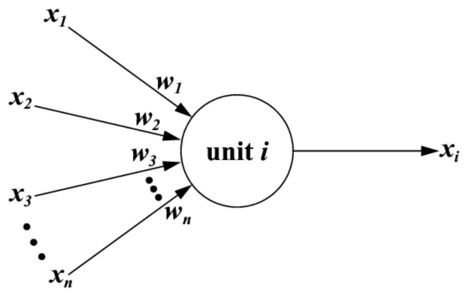

Figure 1: A connectionist processing unit. The arrows on the left depict inputs to the unit, and the arrow on the right depicts its output.

What is a computational model?

Models are often divided into two categories: *descriptive* and *process* models. Descriptive models primarily describe the data obtained from experiments via mathematical equations. In contrast, process models explain the cognitive mechanisms underlying performance in a task. Process models vary in their abstractness, from qualitative verbal descriptions to quantitative computer simulations that embody the cognitive process. Computational models are process models that lean toward the quantitative end of the spectrum.

Computational models come in many varieties. In the area of spatial attention, some models are abstract mathematical characterizations of behavior (e.g., Bundeson, 1990; Sperling & Weichselgartner, 1995), others are algorithmic, describing behavior in a sequence of steps much like a computer program (e.g., Ullman, 1984; Weismeyer & Laird, 1990), and still others, called *connectionist* or *neural network* models, attempt to capture the operation and functional architecture of the brain. Connectionist models are large networks of simple, autonomous, neuron-like processing elements (McClelland, Rumelhart, & Hinton, 1986).

We focus on connectionist models for several reasons. First, connectionist models have proven extremely useful for explaining psychological phenomena in visual perception and attention. Second, while connectionist models do not necessarily describe information processing at a neural level, connectionist models make contact with neurobiological data more readily than do other types of computational models. Third, connectionist models tend to offer a deeper level of explanation than do more abstract frameworks; for example, a mathematical model might treat an attention shift as a primitive operation, whereas in a connectionist model, the attention shift is an emergent consequence of the model's dynamics. Fourth, computer vision research has shown that tasks such as object recognition requires massively parallel, distributed processing (e.g., Marr, 1982), of the sort found in connectionist models. Fifth, and perhaps most important, it is our intuition that connectionist models are the *right* level of description for characterizing the essential properties of visual perception and attention. Ultimately, one must trust their intuition in selecting a set of modeling tools.

Connectionist models

The basic element of a connectionist model is a neuron-like *processing unit*. Figure 1 shows a typical unit. The unit conveys a scalar value, its *activity level*, to other units. The activity level can be thought of as something like the average firing rate of a biological neuron. The activity of unit i is

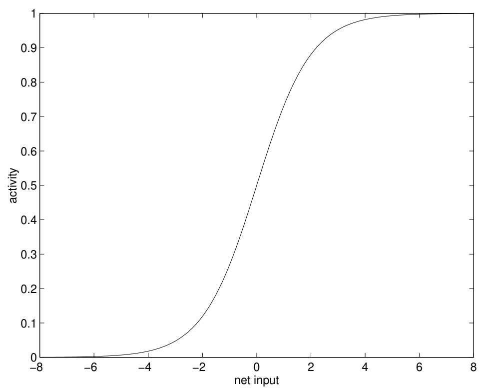

Figure 2: A typical activation function relating the weighted input to a unit and its output activity.

denoted x_i. The arrow on the right of the Figure depicts the flow of activity from the unit. The arrows on the left depict the flow of activity from other units into the unit. The unit's activity is a function of its inputs. In the Figure, there are n input lines. To compute its activity, the unit first calculates a weighted sum of its input, called the *net input*,

$$net_i = \sum_{j=1}^{n} w_{ij} x_j,$$

where w_{ij} is the weighting factor from unit j to unit i. The output of the unit is then a function of the net input:

$$x_i = f(net_i).$$

This *activation function* is typically monotonic and restricts activity between some minimum and maximum value. A common activation function is

$$f(net) = \frac{1}{1 + e^{-net}}.$$

As shown in Figure 2, this activation function maps a net input in the range of $-\infty \rightarrow +\infty$ to activities in the range $0 \rightarrow 1$.

If a particular weight, say w_{ji}, is zero, unit j will not influence the activity of unit i; if the weight is positive, activity in unit j will tend to produce activity in unit i; and if the weight is negative, activity in unit j will tend to suppress activity in unit i. Positive and negative weights are therefore called *excitatory* and *inhibitory* connections, respectively. Learning in a neural network involves modifying the connection weights which changes the response properties of units. We give an example of connectionist learning in a model we introduce below.

Because it is often important to model the time course of activation, we can add a further constraint to the activation dynamics that the rate at which information can flow from a unit is limited. This is achieved by defining the output of the unit as follows:

$$x_i(t+1) = \tau f(net_i(t)) + (1 - \tau) x_i(t)$$

where t is an index over time, assumed to be quantized into discrete steps, and τ, in the range $[0, 1]$, specifies the rate of change. A τ of 0.0 specifies that the rate is infinitely slow, while a τ of 1.0 specifies that the output instantaneously reflects the input state.

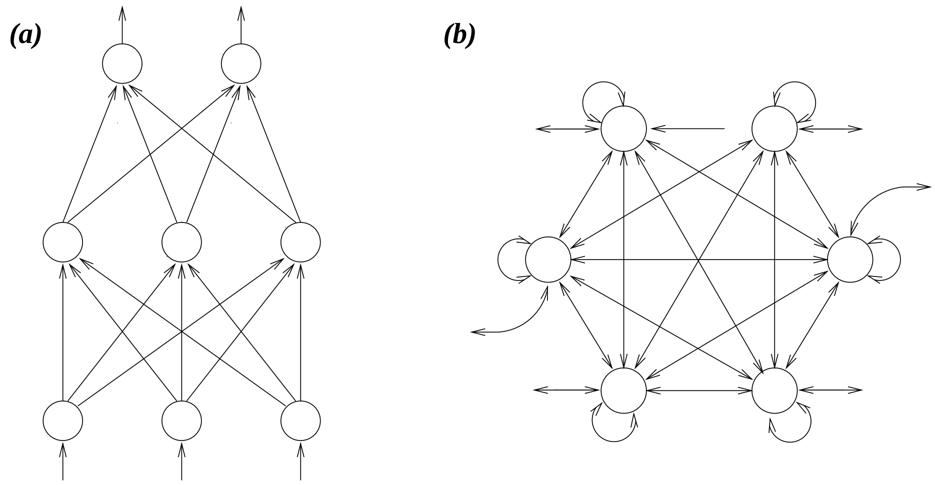

Figure 3: (a) A feedforward architecture in which activity flows from the bottom layer of units to the top layer; (b) A recurrent architecture in which activity flows in cycles.

Connectionist units can be interconnected to form two basic architectures: *feedforward* and *recurrent*. In a feedforward architecture (Figure 3a), activity flows in one direction, from input units to output units. The architecture shown in the figure is also layered by virtue of the fact that units in one layer communicate only with units in the next layer. In a recurrent architecture (Figure 3b), units are connected in a chain such that activity flows out of a unit, through other units, and can eventually influence activity in the unit itself.

A basic model of object recognition

We begin by introducing a general, relatively noncontroversial connectionist model of visual information processing and object recognition. It may strike experimental cognitive psychologists as unusual to propose a model without reference to specific data. However, the strategy we pursue is to first put forth a mechanism that is sufficient to perform the sort of information processing tasks that we believe are essential to cognition. In the case of visual perception, this includes recognizing objects and making judgements about visual stimuli. While the model embodies a basic theoretical perspective on visual information processing, we will not attempt to model specific experimental data until the basic framework has been laid out. The point of the model is not to explain object recognition per se, but to motivate the need for attention and to study how attention interacts with object recognition. Later, we validate the model as psychologically plausible by showing that it can account for experimental data.

Before describing the model itself, we begin by explaining the input and output of the model. To present a visual stimulus to the model, a pattern of activity is imposed on the model's *retina*. The retina is a collection of *feature maps*. Each feature map is a topographic array of units that detect the presence of a particular visual feature in a particular location of the visual field. The version of the model we'll describe has an array of 15×15 units in each feature map, and 5 feature maps: oriented line segments at 0°, 45°, 90°, 135°, and line-segment terminators. We refer to these inputs as *primitive features*.[1]

[1] The name "retina" should not be interpreted literally. The primitive feature representation is more like that found

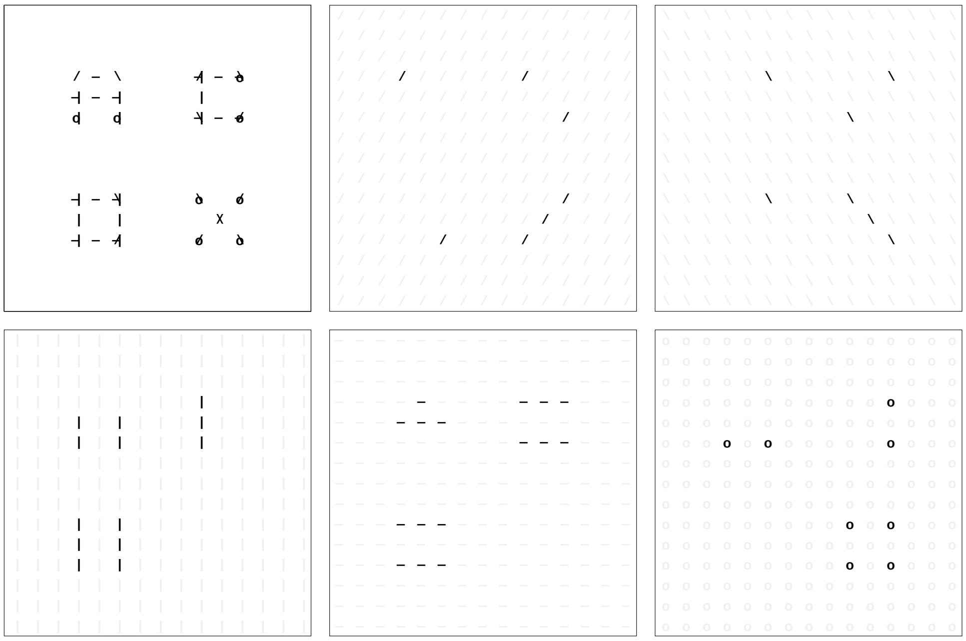

Figure 4: The top left panel shows the set of primitive features that form four letters, **A**, **C**, **D**, and **X**. The small circles depict terminators. The remaining five panels show the activity in each feature map, with a dark symbol indicating that the corresponding feature unit is active, and a light symbol indicating that the feature unit is inactive.

We use a simple font for uppercase letters in which each letter occupies a 3×3 region of retina (Mozer, 1991). Figure 4 shows the pattern of activity that corresponds to four letters—**A**, **C**, **D**, and **X**—on the retina. The activity of a feature unit is represented by the shading of the symbol, dark for activity 1.0 or light for activity 0.0.

In the version of the model we have implemented, the model's task is to recognize letters of the alphabet. There is one output unit for each letter. A unit should be active if its corresponding letter is present in any location on the retina.

Figure 5 shows a sketch of the model. It is a hierarchical feedforward architecture in which each layer of units feeds to the next layer. The bottom layer in the Figure is the input, the top layer is the output. The basic idea of the architecture is to transform low-level, location-specific visual features into high-level, location-invariant object identities. By "low level" or "high level," we mean that the features respond to either simple or complex patterns, respectively; by "location specific" or "location invariant," we mean that the feature detector responds to stimuli only in a particular location on the retina or over the entire retina, respectively. The transformation from input to output is accomplished in several stages. At each stage, the number of feature maps increases, the features respond to increasingly more complex patterns, and the region of the retina over which they respond increases. The logic of the architecture is that by increasing the number of feature maps at each layer, information about spatial relations among features in the layer below can be

in early visual cortical areas than on the human retina. Further, we do not even wish to claim that the coordinate frame of the primitive features is retinotopic. We have simply stated that the features are arranged topographically, but we have not specified whether the feature maps are defined with respect to a coordinate frame that is retinotopic, head-centered, body-centered, or environmental. We avoid this difficult issue because it is not critical to the discussion that follows.

Table 1: Architecture of the recognition neural network

layer	dimensions	number of feature types	receptive field size	receptive field characteristics
4c	1x1	26	2x2	
4s	2x2	26	2x2	overlapping
3c	3x3	20	2x2	nonoverlapping
3s	6x6	20	2x2	overlapping
2c	7x7	15	2x2	nonoverlapping
2s	14x14	15	2x2	overlapping
1	15x15	5		

encoded implicitly and hence the explicit representation of spatial relations (i.e., the dimensions of the feature maps) can be reduced.

The details of the architecture, not too important for the rest of our presentation, are as follows. Units in a layer receive projections only from a local spatial region of the layer below. Neighboring units in a layer receive projections from neighboring regions of the layer below. Table 1 summarizes the architecture. The input layer, layer 1, has an array of 15×15 cells of 5 feature types. The output layer, layer 4c, has an array of 1×1 cells (i.e., there is no explicit representation of location) and 26 feature types (the letters of the alphabet). Between the input and output are three transformation stages, each composed of a "simple" layer and a "complex" layer. The simple layer forms higher-order feature detectors by integrating information over space and feature types in the layer below, while the complex layer integrates only over space, resulting in a representation of the same features with lower spatial resolution. Thus, one will note that the number of feature types in the simple layer is greater than in the layer below, while the number of feature types in the complex layer is the same as in the simple layer. The terms "simple" and "complex" are a reference to cell types in visual cortex.

The ideas embodied in this architecture are traditional. Barlow (1972) and Milner (1974) have described hierarchies of feature detectors for vision. Fukushima and Miyake (1982), Sandon and Uhr (1988; Uhr, 1987), Le Cun et al. (1989), Mozer (1991), and others have built hierarchical connectionist architectures for vision tasks. The idea of dividing each stage of the transformation into simple and complex layers comes from Fukushima and Miyake and Le Cun et al.

Training the model

We have described the basic pattern of connectivity in the model—which units are connected to which other units. The response of the model also depends on the strength of connections between units, the network weights, which are found by a neural network *training procedure*. We sketch the training procedure but it is not essential to understanding the rest of the chapter.

We first generate a set of *training examples*, each of which consists of an *input pattern* and a *target output*. For instance, given the input pattern shown in Figure 4, the target would be an activity level of one for output units corresponding to **A**, **C**, **D**, and **X**, and an activity level of zero for all other output units. The training examples included displays of containing between one and four letters, 104 examples of each display size. Letters in each example were selected at random and always appeared in one of the four standard positions shown in Figure 4.

The goal of the neural network training procedure is to find a set of weights that allow the

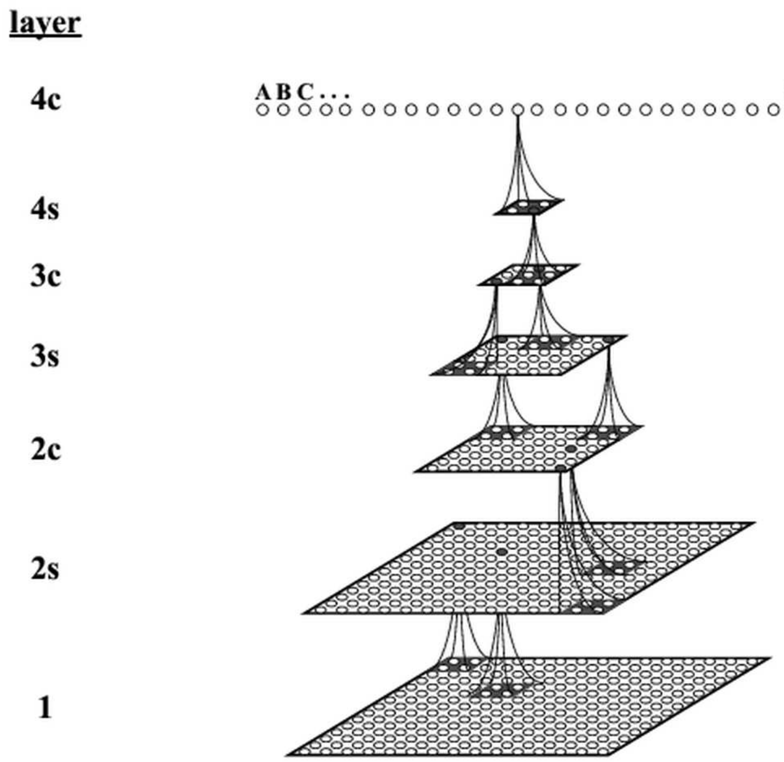

Figure 5: The object recognition model. Each layer is a topographic array of processing units. The bottom layer is the input, the top layer is the output. Activity flows from bottom to top. Each circle represents a *collection* of processing units that detect different features. The connectivity in the network is illustrated by shading a rectangular region in layer l and the location in layer $l+1$ into which this region feeds. Only a small fraction of the connections are depicted.

Table 2: Generalization performance of the recognition neural network

number of letters	**miss rate**	**false alarm rate**
2	10%	0%
3	21%	1%
4	37%	1%

network to perform correctly on the training examples. That is, when any input pattern in the training set is presented, the network should produce an output pattern closely matching the corresponding target output. This is achieved by a commonly used algorithm called *back propagation* (Rumelhart, Hinton, & Williams, 1986). This algorithm starts with random initial values for the weights and makes small incremental changes to the weights such that with each successive weight change, the network produces outputs that better match the target outputs. In order for units of the same feature type in different locations to repsond to the same pattern, their incoming weights must be identical. This is achieved by imposing weight constraints among the units, a common approach for visual object recognition networks (details can be found in Rumelhart, Hinton, & Williams, 1986).

Performance of the model

Following training, we can present any single letter in any of the four standard positions and the model will give a strong response to the appropriate letter (output) unit and a weak response to all other letter units. We can quantify the model's performance in terms of *misses* and *false alarms*. A miss occurs when the model fails to activate the output unit corresponding to a letter present in the image above a threshold of .5; a false alarm occurs when the model activates the output unit corresponding to a letter not present in the image above a threshold of .5. By these criteria, the training set of single letters produce a miss rate of 0% and a false alarm rate of 0%. When we test the model on letters presented in novel positions, i.e., not one of the four standard positions, the model shows a fair degree of generalization, achieving a miss rate of 30% and false alarm rate of 5%. This is not surprising, as the local receptive field architecture and the constraints among the neural network weights favor, but do not strongly enforce, translation invariance.[2]

Table 2 shows performance on test examples of double, triple, and quadruple letter displays. The test examples were formed by selecting random combinations of distinct letters and selecting a location randomly from among the four standard letter positions. Displays that were used in training were excluded.

Performance drops as the number of stimuli increases. One can understand why this must be the case when one views the model in terms of *processing channels*. Information flowing from one letter position in the input to the output passes through a set of intermediate units. The units involved in the processing of one letter position overlap with those involved in the processing of other letter positions, especially at higher layers of the model. The processing channels are thus dependent, and if information is flowing from two channels simultaneously, interference can occur, resulting in the loss of information. Thus, while the model was designed to process visual stimuli in parallel, finite resources result in *capacity limitations*. This motivates the need for some type of attentional processing that can limit the amount of information that the model attempts to handle

[2] *Translation invariance* means that the response of the system is the same regardless of the position of a visual stimulus on the retina.

at once. We now can present an attentional mechanism that performs this function.

A model of attentional selection

We have described a simple network that can recognize single letters well, and can recognize pairs of letters in parallel for the most part, but as the number of letters increases, the quality of the recognition degrades. On computational grounds, then, some means is required to select a subset of the locations in the visual field where letters appear. By selecting locations sequentially, the attentional system can control the flow of information and prevent the recognition system from being overloaded.

The attentional model we present is most similar to an early-selection model described by Mozer (1991). However, there are many related models in the literature, including Ahmad (1991), Koch and Ullman (1985), LaBerge and Brown (1989), and Sandon (1990). We have attempted to synthesize and incorporate the most promising features of each. The core of the model is a set of units arranged topographically, in one-to-one correspondence with retinal locations. This *attentional map* is depicted in Figure 6, along with the primitive feature maps. (Other details in the Figure can be ignored for the time being.) In other models, the attentional map is also referred to as the *priority map* or the *saliency map*. Activity of a unit in the attentional map indicates that units in the corresponding location on the retina are being attended. Attending to a region of the visual field requires activating a compact, contiguous set of units on the attentional map. For the time being, we won't discuss how activity patterns arise in the attentional map. Assume that the activity pattern has been established that indicates attention to a particular region. We will refer to this activity pattern as the *attentional state*.

How might attention control the flow of information in the visual system? The most straightforward notion is to allow the attentional units to *gate the activity flow* from the primitive input features through the object recognition network. If an attentional unit is active, all primitive features at the corresponding location transmit their activities to the recognition network. If an attentional unit is inactive, the activity of primitive features at the corresponding location is not available to the recognition network. This is consistent with experimental findings that once a location is selected, all features at that location are processed (Kahneman & Henik, 1981). This gating operation is depicted in Figure 6 for two locations by a convergence of the output from an attention unit onto the bundle of outputs from the primitive features.

A mathematical specification of this gating operation is simple; it basically involves multiplying the activity of the attentional unit by the activity of the primitive feature units. Let a_{xy} denote the activity of a unit at location (x, y) of the attentional map, and suppose this activity level ranges from a minimum of 0.0 to a maximum of 1.0. Let r_{qxy} denote the activity of a primitive feature type q at location (x, y) on the retina. Then the activity from this primitive feature unit that is conveyed to the recognition network, $\hat{r}_{qxy}$, is

$$\hat{r}_{qxy} = a_{xy}(r_{qxy} - \bar{r}) + \bar{r}$$

where $\bar{r}$ is the resting activity level of the primitive feature units. If the attentional unit has activity 0.0, only the resting activity is conveyed. As the attentional unit activity rises to 1.0, the activity conveyed approaches the actual primitive feature activity. This type of multiplicative junction between processing units is common in connectionist models (see, e.g., Hinton, Rumelhart, & McClelland, 1986).

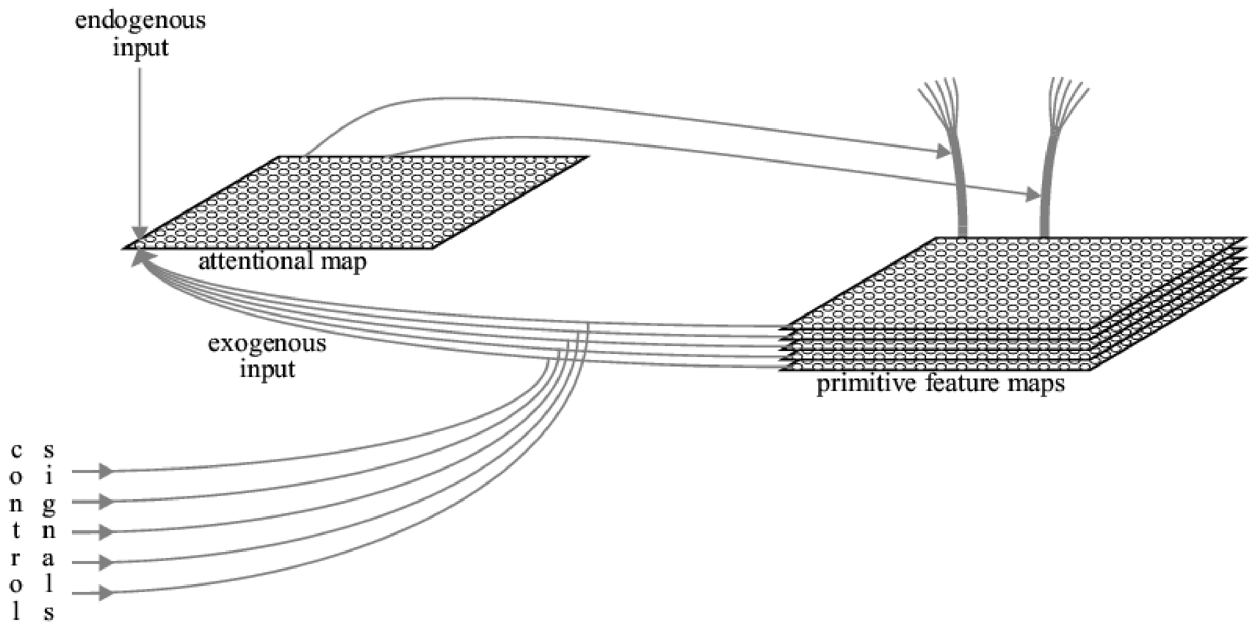

Figure 6: Each "slab" of circles on the right depicts a primitive feature map, in which the circles represent processing units that respond to a particular feature in a particular location. The slab on the left depicts the attentional map, with a set of units in one-to-one correspondence with the primitive feature maps. Further details of the Figure are explained later in the chapter.

Information in the unattended field seems not to be completely ignored in at least some situations (e.g., Gatti & Egeth, 1978; Eriksen & Hoffman, 1973; Shaffer & LaBerge, 1979). This suggests that activity of the primitive feature units in unattended locations should not be completely suppressed, because if it was, unattended information could have no effect on behavior. We will thus suppose that unattended information is attenuated, using a gating function of the form

$$\hat{r}_{qxy} = g(a_{xy})(r_{qxy} - \bar{r}) + \bar{r}$$

where $g(.)$ is a monotonic function, such as

$$g(a) = \lambda + (1 - \lambda)a^{\phi}.$$

For the moment, ignore ϕ. The constant λ determines the degree to which unattended information is passed through the recognition network. With $\lambda = .05$, which we use throughout this chapter, 5% of activity is conveyed even in unattended regions of the visual field. The consequence of a nonzero λ on behavior is not immediately obvious. The fact that 5% of activity is conveyed does not necessarily imply that 5% of unattended objects will be recognized or that their activity will be 5% that of attended objects. This depends in a fundamental way on the operation of the recognition network, and might well interact with familiarity of the unattended stimuli (i.e., the extent to which the recognition network is tuned to processing a particular stimulus) and the nature of task demands (e.g., the information on which responses are based).

The constant ϕ has been included to help suppress the effect of attentional units with weak activity. If $\phi > 1$ and the attentional unit activities are in the range $[0, 1]$, small activity values will be squashed more than large activity values. Throughout the chapter, we use $\phi = 4$.

Having described how an attentional state affects processing in the recognition network, we now specify how the model forms attentional states.

Dynamics of the attentional network

In a model of location-based selection, the attentional state should indicate a contiguous spatial region on the retina; attentional units within the region should be active and all others inactive. It turns out to be somewhat tricky to design an *elastic-spotlight* model that permits regions of varying size and shape. However, we do not need to deal with this problem right now, because letters are of constant size and are always presented to the model in one of four positions. Consequently, we will describe a simplified implementation that captures the essence of the model but, by its simplicity, is easier to interpret and analyze. In this *rigid-spotlight* model, four attentional states suffice, corresponding to the four quadrants of the retina. To attend to a letter position, say the upper left corner of the retina, all attentional units in that quadrant should have activity 1.0 and all units in the other three quadrants have activity 0.0. Because of the redundancy in this attentional state, we can collapse all attentional units in a quadrant to a single unit.

The rigid-spotlight model requires just four units. What determines how active these units will be? There are two sources of input to the attentional network: exogenous and endogenous. Exogenous input comes from sensory data: in any quadrant where primitive features are present, attention should be directed to that quadrant. This will cause attention to shift to locations where stimuli appear. Endogenous input results from previous learning, priming, or cueing which gives rise to expectations about the location of interesting sensory data. Both exogenous and endogenous input directly activate the appropriate attentional units. In the case of exogenous input, one can think of each primitive feature as having a small-weighted connection to the attentional unit in the corresponding location. In the case of endogenous input, one can think of input from an unspecified source to each attentional unit. This is depicted, for the elastic-spotlight model, in Figure 6.

Because only one attentional unit should be active at a time—corresponding to the selection of a particular location—the units should compete with one another. If each unit has an inhibitory connection to each other unit, the unit that is most active will inhibit all others. This is known in the connectionist literature as a *winner-take-all network* (Feldman & Ballard, 1982; Grossberg, 1976). Additionally, each unit should have an excitatory connection to itself, so that if it is active, it will tend toward the maximum activity level. Figure 7 shows a schematic depiction of the attentional model. Algebraicly, the net input to the attentional unit in location (x, y) at time t is:

$$net_{xy}(t) = ext_{xy}(t) + \alpha a_{xy}(t) - \beta \sum_{q,r \neq x,y} a_{qr}(t)$$

and the activity update rule is

$$a_{xy}(t+1) = \tau h(net_{xy}(t)) + (1-\tau) a_{xy}(t)$$

where $ext_{xy}(t)$ is the external input, endogenous and exogenous, to the attention unit, α is the strength of the excitatory self-connection, β is the strength of the inhibitory connection between pairs of units, and h is a threshold linear function that limits activity to the range $[0, 1]$:

$$h(net) = \begin{cases} 0 & \text{if } net < 0 \\ net & \text{if } 0 \leq net \leq 1 \\ 1 & \text{if } net > 1 \end{cases}$$

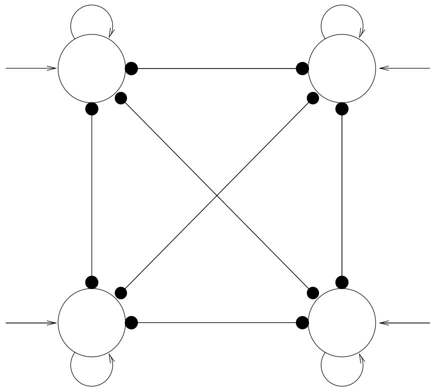

Figure 7: The rigid-spotlight attentional model. Each attentional unit represents a quadrant of the retina. Each unit is self-excitatory and inhibits each other unit. Each unit receives input from exogenous and endogenous sources. Excitation is represented by connections with arrows, inhibition by connections terminated with small circles.

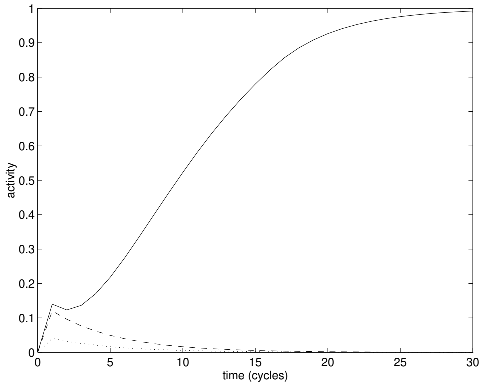

Figure 8: Activity of four attentional units as a function of time. All units have initial activity 0.0. Unit 1 (solid line) has external input .7, unit 2 (dashed line) has external input .6, and units 3 and 4 (dotted lines) have external input .2.

In simulations, we use $\alpha = .4$, $\beta = 3.5$, $\tau = .2$ for the attentional net. Figure 8 shows a graph of the activity over time of four attentional units, given fixed inputs, ext_{xy}. Unit 1 wins the competition and becomes fully active, while the other units are suppressed. Exogenous input to an attentional unit is based on the number of primitive features present in the quadrant represented by the unit. Each feature has a probability ρ of being included in the external input at each time. When detected, the feature contributes a constant σ to the external input. In simulations, we used $\rho = .8$ and $\sigma = .1$.

Curiously, this model does not treat attention as a limited resource of which there is only a finite amount to go around; if we wanted to, we could reduce the value of β so that units would no longer compete so strongly as to shut one another off. The limited resource in this model is found in the object recognition system. Without attentional selection, objects in the visual field will interfere with the processing of each another.

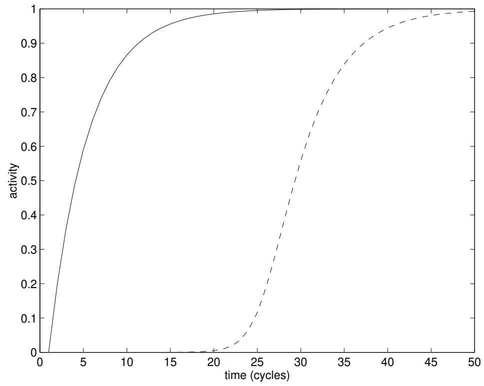

Figure 9: Build up of activity in an attentional unit (solid line) and the letter **S** unit (dashed line) in response to the stimulus **S**.

Simulations of spatial selection

The benefit of attentional precuing

With the attentional network in place, we can run simulation experiments using the model. Initially, the model is reset to a neutral attentional state in which all attentional units are inactive. A stimulus is presented to the model by introducing a pattern of activity over the primitive features. The primitive features provide input to the attentional network, leading to activation of attentional units. At first, the inactive attentional network prevents most primitive feature activity from entering the recognition network, but as the competition takes hold in the attentional network, one location becomes preferred and primitive feature units in this location are allowed to pass their activity through the recognition network. Figure 9 shows the response of the model when the letter **S** is presented in the upper left quadrant. The Figure depicts both the activity of the attentional unit in the stimulus location and the activity of the letter unit. The time required to activate the letter unit is due to the gradual build up of activity in the attentional network as well as the slow propagation of activity through the recognition network (as determined by the constant τ).

A simulation trial does not have to begin with a neutral attentional state. If the model has been cued to a location in advance of the trial, endogenous input to the attentional network will sustain activation at that location in the attentional map prior to stimulus onset. When the stimulus is presented, activation from the primitive features will immediately flow through the recognition network. Consequently, one might expect more rapid response to the stimulus.

Posner (1980) has studied a speeded detection task with location precuing. Subjects were asked to detect the onset of a suprathreshold target stimulus at one of several possible locations. Prior to target onset, the subject might be provided with a spatial cue indicating the location in which the target is likely to appear. Subjects were faster to detect the target when a cue was given than when no cue was given. Our description of the model's behavior is consistent with this result. Further, Posner manipulated the cue to be either a *valid* or *invalid* predictor of target location. Responses in the valid cue condition were faster than in the *neutral* cue condition, while responses in the invalid cue condition were slower (second column of Table 3).[3]

[3]Most of the human data we use for comparison and for setting model parameters has been extracted from figures of the referenced experimental papers, and/or has been averaged over several experiments. Because the details of our simulation experiments do not match the details of the human experiments (e.g., stimuli, presentation conditions),

Table 3: Reaction time to detect target onset

cue condition	**human RT**	**model cycles**	**model RT**
valid	230 ms	15.7 cycles	234 ms
neutral	260 ms	17.5 cycles	256 ms
invalid	300 ms	20.9 cycles	301 ms

Simulating experimental results even as basic as these nonetheless requires further assumptions about the operation of the model.

1. How do the different cue conditions correspond to states of the model? We assume that in the neutral condition, all attentional units are inactive. A cue—valid or invalid—guides attention endogenously to the cued location prior to presentation of the target.

2. How does the model formulate a response? The detection response must be based on some representation in the model; it could be the primitive input features, on the letter units, or on any level between. We assume that read out is based solely on the outputs of the recognition network.[4] Because the detection response depends only on whether a stimulus is present, not its identity, we assume that the evidence upon which a response is based is the total activity of the letter units.

3. When does the model initiate a response? One might assume a response is initiated when the total evidence passes some threshold. If there were no noise in the model, the threshold could be set to zero. However, our recognition model is noisy in that letter units have slight activity even when no stimulus is present. Additionally, most models assume some built-in noise that reflects sources of variability not modeled explicitly. Thus, the threshold should be set as low as possible such that responses can be initiated rapidly and without producing false detections due to noise.

   The response generation procedure we adopted is a variant of the procedure used by McClelland and Rumelhart (1981). We describe the general procedure. For each possible response r the model might be asked to make at time t, the *evidence* for the response, denoted $e_r(t)$, is computed. The probability of producing a response r at time t is then:

$$p(r,t) = \frac{\exp(\xi e_r(t))}{\sum_s \exp(\xi e_s(t))},$$

   where ξ is a constant that translates evidence into response strengths. The numerator is the strength of response r, and the denominator normalizes the probabilities to sum to 1. This rule will always choose a response at each time t it is applied. However, we would like to prevent the model from making a response unless sufficient evidence has accumulated. To do this, we add an additional response category, which we call "no response", that has constant evidence e_{NR}. This constant behaves as a probabilistic threshold; if e_{NR} is large relative to the evidence for the other responses, then the model will likely hold off making a response. e_{NR} is a free parameter of the model, and it essentially controls the speed-accuracy trade off:

there is little to be gained by trying to determine and model the exact outcome of a specific human experiment.

[4] It may seem strange to read out at a high level when the task does not call for stimulus identification. We find it most parsimonious to make the strong assumption of a single level of read out, and thereby avoid the issue of determining where read out occurs on a task-by-task basis.

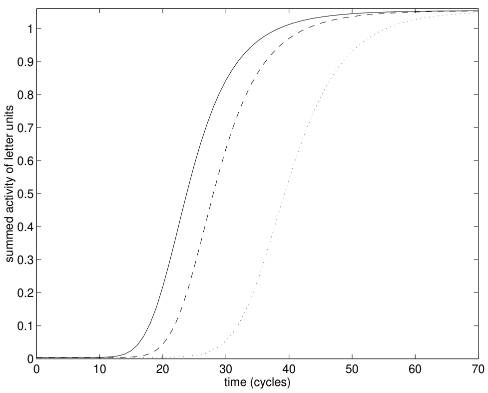

Figure 10: Summed letter unit activity as a function of model *cycles* for the valid (solid line), neutral (dashed line), and invalid (dotted line) cue conditions, averaged over a large number of stimulus presentations. A cycle is a single update of the activities of all units in the model.

the larger e_{NR} is, the more evidence must accumulate before a response is initiated. In all our simulations, $\xi = 10$; in detection tasks, $e_{NR} = .3$ and in discrimination tasks, $e_{NR} = 1.2$, reflecting the fact that more evidence should be required for a discrimination response than a detection response.

Figure 10 shows the summed letter unit activity as a function of time for the three cue conditions. Time is measured in *cycles*; each cycle is a single update of the activities of all units in the model. One can clearly see that activity rises most rapidly in the valid cue condition, followed by the neutral cue condition, followed by the invalid cue condition. The third column of Table 3 shows the mean number of cycles for the model to initiate a detection response, over a large number of stimulus presentations. The simulation response times are qualitatively in accord with the data. We scaled the model response time in cycles, RT_{model} to real-world response times, $RT_{realworld}$, according to a formula that assumed a fixed number of milliseconds per cycle, γ, and a fixed amount of time, κ, for input preprocessing and motor response:

$$RT_{realworld} = \gamma RT_{model} + \kappa,$$

We chose values for these constants—12.8 for γ and 32.8 for κ—to obtain a reasonable fit to this data. The same constants will be used in all subsequent simulation experiments.[5]

Cohen et al. (1994) and Jackson, Marrocco, & Posner (1994) have modeled the effect of cues on speeded detection using essentially the same approach—a set of attentional units that compete to select a location and preactivating units at the cued location. While the activation dynamics and competition mechanisms vary among the three models, and while Cohen et al. and Jackson et al. do not simulate the perceptual system in any detail, all three models show the same effect. This suggests that the effect is robust under a variety of implementations of the same key notion—that attention is the result of a competition among locations. Jackson et al. have attempted to provide a more neurobiologically plausible mechanism, localizing various components of their model to different brain regions. Both Cohen et al. and Jackson et al. also account for data of neurological patients with attentional disorders by "lesioning" their models in a manner consistent with the form of damage the patients are known to have suffered.

[5]Surprisingly, some modelers (e.g., Cohen et al., 1994) have allowed themselves the freedom of fitting the results of different experiments with different scaling parameters.

We attempted to extend our model in a somewhat different direction. Shiu and Pashler (1994), summarizing the literature on the effect of advance knowledge of stimulus location in processing single-item displays, concluded that although a spatial precue results in significant speedups in detection tasks, the effect is more modest in speeded suprathreshold discrimination tasks. We simulated a discrimination task in which the model was given a valid, neutral, or invalid cue, which was followed by one of two visually confusable targets, such as **X** and **Y**, and a forced-choice speeded response was required. For the discrimination task, a response cannot be initiated until the model is confident that one stimulus was presented and not the other. This is particularly critical because a stimulus like **X** will often produce activity for visually confusable letters like **Y**. Thus, we set the evidence for the **X** response, $e_{\mathbf{X}}$, to be the difference in activity between the **X** and **Y** units, and symmetrically for the **Y** response.

After experimenting with parameter values, activation functions, and response functions for over a week, we had to admit defeat: The model always produced a cue-validity effect in the discrimination task which was as large as the effect in the detection task. Figure 10 suggests one argument for why this might be. The curves for the three cue conditions appear identical except for a shift in time. Although the Figure shows summed activity of all letter units, this is true for the individual unit activity curves too. Any response initiation procedure based on these parallel curves will necessarily produce response times for the cue conditions that differ by the time shift. Thus, the detection cue-validity effect must be the same as the discrimination effect. Although it is theoretically possible that certain parameter settings might result in nonparallel curves, we were unable to discover such settings.

Two lessons might be learned from this exercise. First, the model shows a parameter-independent, qualitative behavior, indicating that it represents a strong, testable theoretical perspective. Large computational models often arouse suspicion because they appear sufficiently malleable that they can be made to account for any piece of data. More often than not, this belief is misguided, as we discuss later. Second, if one has strong confidence in the model, one might question Shiu and Pashler's conclusion from the literature, which is based on studies of Posner (1980) and Posner, Snyder, and Davidson (1980). Although both studies appear to show smaller cue-validity effects for discrimination than detection, this conclusion was not backed up by statistical analyses. Further, the detection and discrimination tasks were performed with different stimulus materials and experimental procedures, making it problematic to directly compare results. (Our simulation results assume that detection and discrimination tasks are carried out under identical stimulus and experimental conditions, except for the response required of subjects.) Resolving whether the model or the characterization of the data is right is beyond the scope of this chapter, but the model—right or wrong—has clearly pointed to an avenue of further investigation.

Time course of attention shifts

In the cue-validity simulation, we assumed that the cue was presented sufficiently far in advance of the target that attention could settle on the cued location prior to the target onset. What happens if the stimulus-onset asynchrony (SOA) between cue and target is varied so that the target is presented before attention becomes fully active at the cued location? Experimental studies have shown that response times decrease monotonically for increasing SOA, up to about 200 ms in both detection and discrimination tasks with peripheral cueing (e.g., Eriksen & Hoffman, 1974; Posner, 1980).

Figure 11 shows a simulation result for the model on a detection task in which a cue is presented for a varying number of cycles, and is then replaced by a target item to be detected. We assume

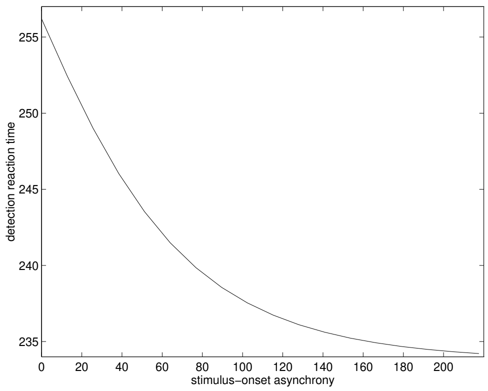

Figure 11: Response time of the model to detect a stimulus as a function of cue-target SOA. The simulation shows the same pattern as human performance: Response times decrease monotonically for SOAs up to about 200 ms.

that the cue initiates activity in the attentional network but not in the recognition network. The same detection procedure is applied as in the cue-validity simulations. Clearly, the model shows the same pattern as human performance.

Effect of spatial uncertainty

Speeded response to a visual stimulus is delayed by the presence of irrelevant stimuli, even when sensory interference, discriminability difficulties, and response conflict are ruled out as contributing factors. In a study by Kahneman, Treisman, and Burkell (1983), observers were asked to read as rapidly as possible a word that appeared unpredictably above or below the fixation point. On half the trials, another object was presented on the opposite side of fixation, either a word or a word-sized patch of randomly placed black dots. The mere presence of the second object resulted in a reading time delay of 30–40 msec.

We simulated this experiment by presenting a letter in one of the four letter locations and a "black dot patch" in one of the other locations. We assume the black dots activate some unspecified primitive visual features that drive attention to the location of the black dots, as do the other primitive features, but they do not activate the letter features used in the recognition network. We also assume that the letter features provide strong exogenous input to attention, causing attention to eventually select the letter location.[6]

Using the discrimination task, response time of the model was 568 ms in the condition with a letter alone and 605 ms in the condition with a letter and the black dot patch. The explanation for this behavior is straightforward: When the dot patch competes with the letter for attention, the activity of the letter location in the attention network grows more slowly, causing a delay in propagating information through the recognition network.

[6]As we elaborate later, the model requires the ability to modulate the degree to which each primitive feature type can drive attention; in this task, choosing the letter location is desirable, and hence letter features should drive attention more strongly than features of the black dots.

Table 4: Reaction time to target

distractor type	**human data**	**model**
compatible	460 ms	459 ms
neutral	500 ms	493 ms
incompatible	540 ms	546 ms

The effect of irrelevant stimuli

When subjects are asked to make a speeded response to a target letter in a known location, their responses can be influenced by the identies of other letters nearby in the display (e.g., Eriksen & Eriksen, 1974; Eriksen & Hoffman, 1973; Eriksen & Schultz, 1979; Miller, 1991). Consider the task of pressing one response key for the target **A** or **U**, and another key for the target **H** or **M**. Letters can be presented flanking the target which are either *compatible* with the target (i.e., selected from the same response category), *incompatible*, or *neutral* (i.e., not belonging to either response category). Responses are fastest on compatible trials and slowest on incompatible trials (Table 4). Flanker effects are significantly reduced when the flankers are presented one degree of visual angle or more from the target, but this may well be due to reduced acuity at greater distances (Egeth, 1977).

The flanker effect appears to be a failure of focal attention, in that subjects are unable to prevent the processing of letters adjacent to a target even if the target location is known in advance. This effect can be eliminated under some conditions, however (LaBerge et al., 1991; Yantis & Johnston, 1990).

The model has a simple explanation for the flanker effect. When a location is unattended, activity from that location is not completely suppressed; a small amount of activity stemming from that location—represented by the constant λ—is transmitted to and analyzed by the object recognition network. This may result in letter activity that will strengthen the evidence for one response category in the case of compatible flankers or weaken the evidence in the case of incompatible flankers.

We performed a simulation study in which a target letter was presented in a fully attended location, and two flankers appeared in adjacent unattended locations. A large number of trials were run, varying the response sets and the stimulus locations. The response initiation procedure was that of the discrimination task we modeled earlier. The results in the three flanker conditions are shown in Table 4. When unattended information is fully suppressed by setting λ to zero, the effect vanishes.

The model has now been shown to account for the results of four quite different phenomena related to selective attention. Although the model produces excellent quantitative fits to the human data, the reader should recognize that there is a bit more going on behind the scenes than we have told you about. For example, in the dot patch experiment, we had the freedom to manipulate the strength of the exogenous input representing the dot patch, enabling us to produce an effect of the right magnitude. Nonetheless, it would be impossible to manipulate the model to alter the qualitative pattern of results, e.g., to cause the effect of attention to diminish as the cue-target SOA increased. At the end of the chapter, we return to the issue of qualitative versus quantitative modeling of data.

Attention as a spotlight?

Spatial attention has been likened to a spotlight (e.g., Eriksen & Hoffman, 1973; Posner, 1980). This metaphor implies that attention is allocated to a contiguous, possibly convex, region of the visual field. If the spotlight metaphor is appropriate, then the spotlight should have an adjustable diameter (Eriksen & Yeh, 1985; LaBerge, 1983). The rigid-spotlight attentional model simulated in the previous section does indeed select a contiguous region, but the region is of fixed size and shape—an entire quadrant of the visual field. We now discuss an implementation of the elastic-spotlight attentional model, which is able to select regions varying in size and shape. In this model, the attentional map has the same dimensions as the retinal map, and a region of the visual field is attended by activating all attentional units in that region.

The elastic-spotlight model is identical to the rigid-spotlight model, except that the dimensions of the attentional map are increased and the computation of the "net input" to the attentional unit at map location (x, y), net_{xy}, is changed to

$$net_{xy}(t) = ext_{xy}(t) + \alpha \sum_{\substack{(i,j) \in \\ \mathrm{NBHD}_{xy}}} a_{ij}(t) - \beta(\gamma \bar{a}(t) - a_{xy}(t)) \ ,$$

where $ext_{xy}(t)$ is the external input to the attentional unit, as before, NBHD_{xy} is the set of nine locations immediately adjacent to and including (x, y)—the *neighbors*, $\bar{a}$ is the mean activity of units with nonzero activity, α and β are the same constants as before, and γ is an additional constant.

The first term encourages each unit's activity to be consistent with the external input, as before. The second term encourages each unit's activity to be as close as possible to that of its neighbors; if a unit is off and the neighbors are on, the unit will tend to turn on, and vice versa. The third term encourages units having activity below the mean to shut off, and units above the mean to turn on. The constant γ serves as a discounting factor: with γ less than 1, units need not be quite as active as the mean in order to be supported. Instead of using the average activity over *all* units, it is necessary to compute the average over the *active* units. Otherwise, the effect of the third term is to limit the total activity in the network, i.e., the number of units that can turn on at once. This is not suitable because we wish to allow large or small spotlights depending on the external input.

To explain the activation function intuitively, consider the time course of activation. Initially, the activity of all units is reset to zero. Activation then feeds into each unit in proportion to its external input (first term in the activation function). Units with active neighbors will grow the fastest because of neighborhood support (second term). As activity progresses, high-support neighborhoods will have activity above the mean; they will therefore be pushed even higher, while low-support neighborhoods will experience the opposite tendency (third term).

This model has been used to explain data from neurological patients suffering from attentional disorders (Mozer & Behrmann, 1990; Mozer, Halligan, & Marshall, 1996). We have adopted the parameter values from the earlier work: α was set to .11, β to .5, and γ to .11 times the total external input, with lower and upper limits of .75 and 1.0. A feature contributes external input not only to its corresponding attentional unit—as in the rigid-spotlight model, the contribution is σ with probability ρ—but also to its neighboring locations with $\sigma_{neigh} = .02\sigma$. The original intent of this blurring was to give the input a more continuous spread of activity.

Figure 12 shows an example of the attentional model selecting a single region when the external input specifies three blob-like patterns that represent distinct objects. The region chosen by the

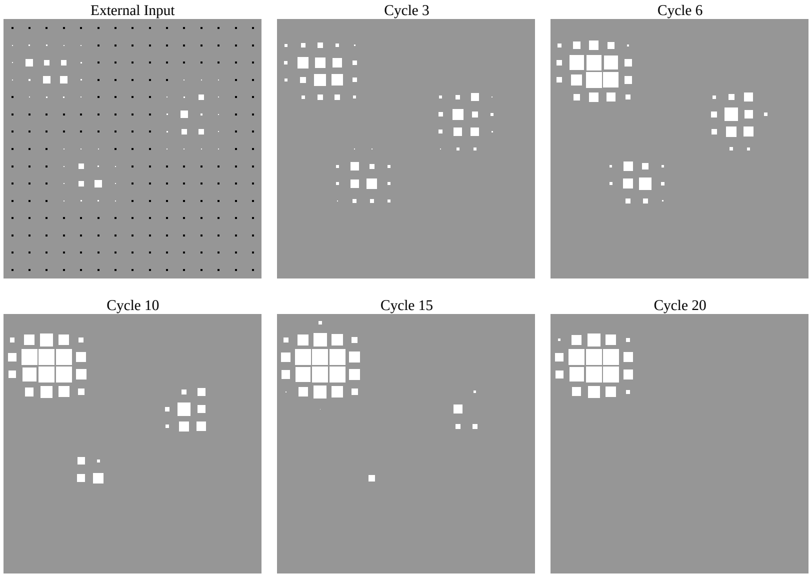

Figure 12: The upper left panel depicts the external input to the attentional model. The panel consists of a 15×15 array of squares. The area of a white square corresponds to the amount of external input to the corresponding unit of the attentional model. The small black dots are drawn in locations where the external input is zero, to show the extent of the array. The external input pattern is meant to indicate three objects, the largest one—the one with the strongest external input—is in the upper left portion of the field. The next five panels show the activity as the network settles. By cycle 20, the network has reached equilibrium and has selected the region of the largest object.

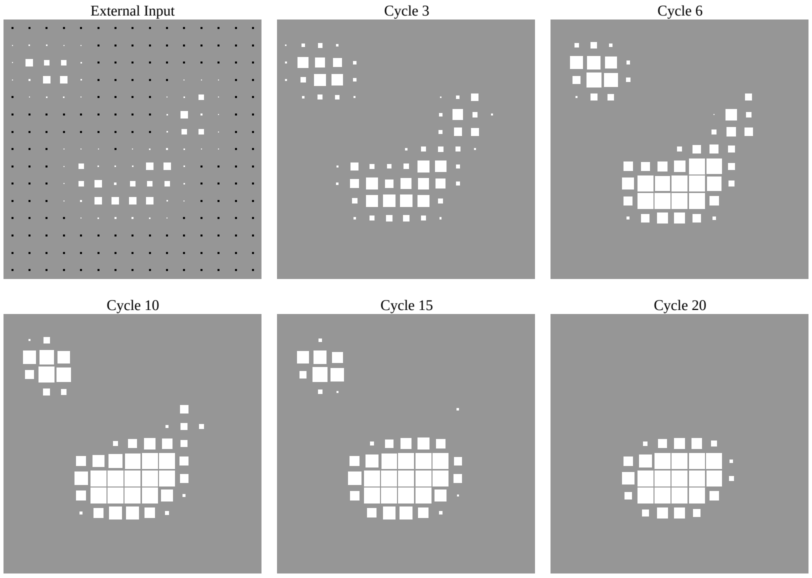

Figure 13: The response of the attentional model to an external input pattern in which there are three objects, the largest of which is at the bottom and center of the field.

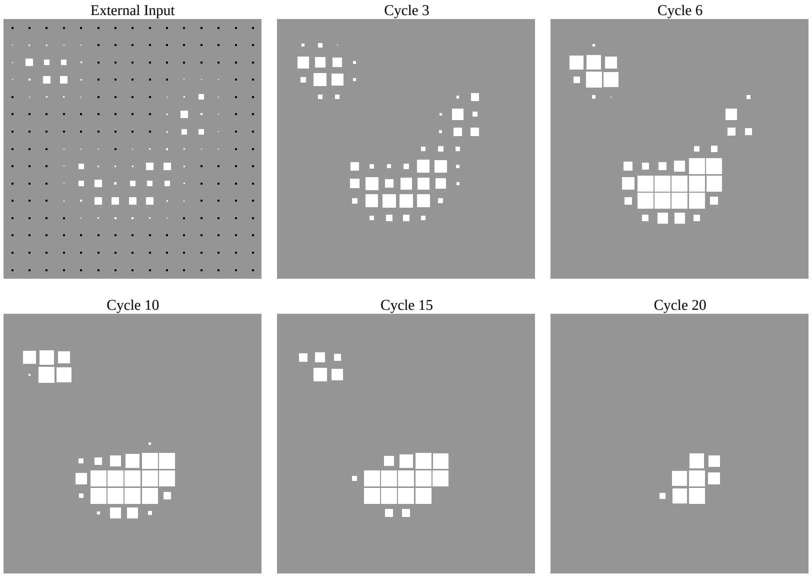

Figure 14: The response of the attentional model when the β parameter is raised from 0.5 to 0.7. The external input is the same as in Figure 13, but the region selected is clearly smaller.

model corresponds to the object with the strongest external input. Figure 13 shows a similar example when one of the blobs is made larger, and a correspondingly larger region is selected. Comparing the two figures, it is clear that the model can select regions of varying size. Model parameters can also be adjusted to vary the size of its spotlight without changing the input. Figure 14 shows the response of the model when the β parameter is raised from 0.5 to 0.7. The external input is the same as in Figure 13, but the region selected is clearly smaller.

Two properties of the network are worth noting. First, the units on the edge of the spotlight tend to have less activity than the units in the center of the spotlight. One is tempted to relate this to the claim that sensitivity falls off gradually at the perimeter of the attended region (Eriksen & St. James, 1986; Downing & Pinker, 1985; LaBerge & Brown, 1989). Second, all stimulus locations become active in the initial phase of processing. It isn't until competitive mechanisms take reign that a winning location emerges. Thus, the model is unfocused initially, but over the course of time it narrows in on a single object. Because the recognition network begins processing immediately—and before the attentional network has settled to equilibrium—it initially tries to handle all information in the field simultaneously. If one were to observe the activity of units in the recognition network, it would appear as if the units responded to unattended stimuli at first, but this activity was eventually suppressed. In single cell studies of monkey visual cortex, this behavior has been observed: 60 msec after stimulus onset a response is triggered in the extrastriate cortex, but not until 90 msec does attention kick in and suppress unattended stimuli (Desimone & Duncan, 1995).

The model was not designed with these data in mind, but it does appear a natural consequence of such a filtering mechanism. One can envision two basic designs: (1) a *cautious* system that does not allow the processing of any information until selection is complete; and (2) an *audacious* system that allows the processing of all information until selection is complete. The audacious system will respond more rapidly, but is more prone to error because items in the visual field may interfere with one another when attention is unfocused. The model, and apparently the primate

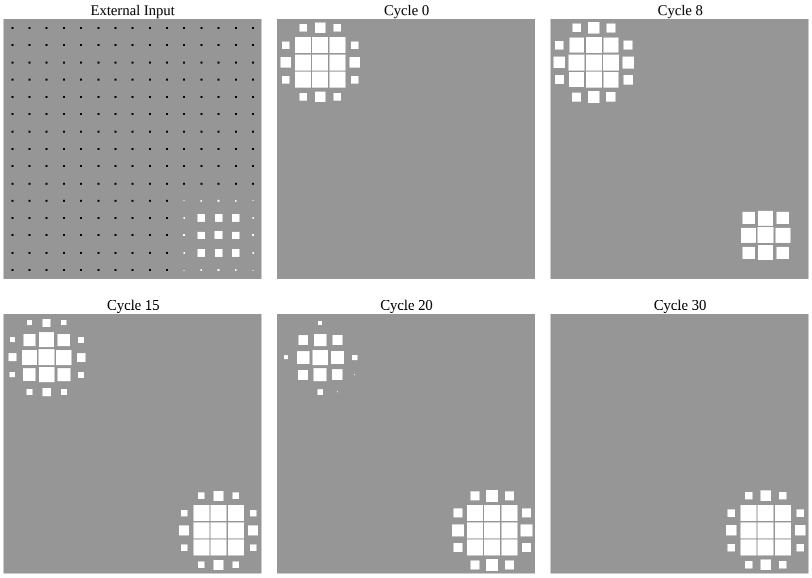

Figure 15: The response of the attentional model when it is already attending somewhere and the external input changes, triggering an attention shift. The external input (upper-left panel) appears at cycle 0, when the model is in the state depicted in the upper-middle panel. The remaining panels show the shift of attention from the upper-left corner of the field to the lower-right corner.

brain, is audacious. This is a sensible strategy if the cost of slow responses is greater than the cost of occasional errors.

In the simulations above, the initial state of the model was neutral. Essentially, the model was attending nowhere, then a multi-item display appeared which initiated a competition among the stimuli for attention. What if the model is already attending somewhere when the display appears, requiring a shift of attention from one location to another? Figure 15 illustrates this situation. Attention fades out from the old location and in to the new. The spotlight metaphor does not seem appropriate for describing the attention shift. If the focus of attention were like a spotlight, one would expect attention to move across the field in an analog fashion, illuminating intervening points along the way. One would also expect that the time required to shift attention would be monotonically related to the distance between foci. The model does not show this behavior either: The time required for attention to shift from a focus at (2,2) to stabilize on a focus at (12,12) is 32 cycles (Figure 15). The time required for a shift half as far—from (2,2) to (7,7)—is also 32 cycles.

Early evidence in the literature did appear to support an analog view of attentional shifts (Shulman, Remington, & McLean, 1979; Tsal, 1983). However, several critiques have appeared of this interpretation of the original data (Eriksen, 1990; Yantis, 1988) and recent experiments suggest that the time to shift attention is independent of the distance traversed and of the presence of interposed visual obstacles (Sperling & Weichselgartner, 1995). The current consensus is that the spotlight of attention turns off at one location and then on at another (Eriksen, 1990; Kinchla, 1992). Thus, our attentional model, which was not designed to behave this way, appears to capture the key property that attention shifts are discrete and distance independent.

Modeling various selection criteria

We have described several simulations in which the model selects items for report based on location. Other tasks require selection by different properties of the stimulus, for example, reporting the identity of the red letter or the location of the brightest dot. An important aspect of a computational model is that, beyond explaining data, it can also carry out the same sort of operations as can people. Thus, in this section, we endow our model with a mechanism that allows it to perform selection by simple physical attributes such as color or brightness. The mechanism assumes that selection by attributes other than location is nonetheless mediated by location selection, consistent with findings of Snyder (1972), Nissen (1985), and Tsal and Lavie (1985). The mechanism is based on models by Mozer (1991) and Wolfe, Cave, and Franzel (1989).

Earlier, we characterized the input to the model in terms of primitive feature maps. Each map is a spatiotopic array of detectors tuned to a particular feature. Until now, we have only required features of letters—oriented line segments and line terminators—but suppose that the primitive feature maps include other dimensions of the stimulus, such as color and brightness.

To perform selection by arbitrary features, the model needs the ability to specify which of the feature maps provide exogenous input to the attentional network. This is achieved through a set of *control signals*, one per feature map, as shown in Figure 6. The control signals modulate the probability that features in that feature map are detected by the corresponding unit in the attentional map. We referred to this probability earlier as ρ, but we now add the index q to indicate the control signal for feature type q, ρ_q. By default, the ρ_q will have value .8, as we assumed for ρ. The control signals in Figure 6 are shown only for a single location, but the gating is performed at every location across the spatiotopic map.

If the task requires selecting the red item for report, then the system should be configured such that only activity from the "red" feature map drives the attentional network, causing selection of red items. If the display contains only a single red item, it will be selected, activity from all feature types in its spatial location will then be allowed to pass through the recognition network, and the output of the recognition network will be the identity of the red item.

What are the primitive feature dimensions that can drive attention? In addition to edge orientation and termination, color, and brightness, there is evidence to support dimensions such as size, direction and speed of motion, binocular disparity, and three-dimensional surface properties (Driver, Mcleod, & Dienes, 1992; Enns, 1990; Hillstrom & Yantis, 1994). Discontinuities or singletons in all of these dimensions appear capable of attracting attention as well (Pashler, 1988; Sandon, 1990; also, see chapter by Yantis, this volume). And multiple spatial scales of resolution must be encoded.[7] By this reckoning, there are at least fifteen primitive feature dimensions, and to coarse code a value on each dimension (e.g., to specify the value "red" on the color spectrum) would require a bare minimum of, say, five feature types, resulting in at least 75 primitive feature types.

Having argued for voluntarily control over which features can drive attention, we must add that this control is certainly limited. Some visual features may attract attention willy nilly (e.g., Jonides & Yantis, 1988; Pashler, 1988; Treisman & Gormican, 1988), indicating that it is difficult or impossible to gate out these features. And, based on evidence we discuss below, there are probably bounds on the visual system's ability to gate in or out various feature types.

[7]Our framework makes no strong assumptions about the nature of the primitive visual features. Many features we have listed would not ordinarily be considered "primitive." The framework allows for considerable preattentive parallel visual processing prior to "feature" registration.

In terms of the model, we propose a simple limitation on control over which features can drive attention. Let us allow the control signal, ρ_q, for each feature type q to be continuous in the range $[0, 1]$. The control signal then determines the degree to which a feature type will attract attention. Suppose that each ρ_q has a default setting which has been determined by past experience, based on what features in the environment tend to be most important and need to be responded to quickly. Modulating the value of a ρ_q requires some type of limited resource, let us call it *regulatory juice*. The amount of juice is sufficient to, say, fully open or close one gate, or to make small adjustments in several gates. It may even be that some gates are easier to modulate with a fixed quantity of juice. The point is that the ρ_q cannot be adjusted arbitrarily.

The introduction of control signals into the model allows us to explain selection on the basis of primitive features other than location. The notion of regulatory juice allows us to explain how selection criteria are adjusted in response to task demands. Experimental data are consistent with this notion, e.g., short-term experience performing a task can affect the degree to which certain feature types drive attention, and this effect can be either excitatory or inhibitory, i.e., increasing or decreasing the ρ_q (Hillstrom, 1995; Maljkovic & Nakayama, 1994). Treating the regulatory juice as a limited resource allows us to account for limitations on attentional selectivity. The general issue of voluntary control over exogenous influences on attention is beyond the scope of this chapter, although we find it difficult to build a computational model without at least specifying the "hooks" for such control from unspecified higher cognitive processes.

The relationship of object-based and location-based attention

Studies have shown that attention can select stimuli on the basis of object shape or structure (e.g., Behrmann, Zemel, & Mozer, 1996; Duncan, 1984; Egly, Driver, & Rafal, 1994; Kramer & Jacobson, 1991; Vecera & Farah, 1994). For example, Kramer and Jacobson examined the influence of flankers on a target stimulus, similar to the experiments described earlier. When the flankers and the target were considered part of the same object, there was a response-compatibility effect; when the flankers and target were part of different objects, there was no effect, even though the spatial separation between the target and flankers was the same in the two conditions.

The data argue for *object-based* selection: Visual features are attended to not on the basis of their spatial location but according to which object they belong, even if the features are not spatially compact and overlap with features of other objects. Two very different processes could underlie object-based attention. One possibility is that attention is allocated to an object-based representation, perhaps a high level, abstract representation of object identity. The other possibility is that attention is allocated to a set of spatial locations, possibly noncontiguous, at which features of an object are present. Evidence from Vecera (1994) supports the latter interpretation.

What is the relationship between object-based and location-based attention? Both forms of attention can be observed in the same experiment (Egly, Driver, & Rafal, 1994), suggesting that the two are not mutually exclusive. Consequently, one must ask which type of attentional selection operates first, or whether there is an interactive process in which both types of selection occur in parallel. Experimental work like that of Kramer and Jacobsen (1991) argues that object-based segmentation must precede or interact with location-based selection. Assuming that object-based segmentation is related to perceptual processes that group distinct display elements into coherent regions, additional support for this hypothesis can be found (Driver & Baylis, 1989; Duncan, 1995), and there are several recent theoretical proposals that embody the hypothesis (Grossberg, Mingolla, & Ross, 1994; Humphreys & Müller, 1993; Rensink & Enns, 1995; Trick & Pylyshyn, 1994).

Given that object-based selection involves allocating attention to spatial arrays of features, and that object-based selection operates prior to or simultaneously with location-based selection, the mechanism of attentional gating we have already proposed is adequate to explain object-based selection. We must, however, posit an additional process that segments features of a display according to which object they belong and can guide attention to the locations of a single object's features.

Several computational models have been proposed to segment displays into their component objects. Humphreys and Müller (1993) and Grossberg, Mingolla, and Ross (1994) have built connectionist models that group display elements on the basis of similarity and spatial proximity. Mozer, Zemel, Behrmann, and Williams (1992) have designed a connectionist model that *learns* which features are likely to be grouped together or apart based on a set of presegmented examples. It thus extends the notion from Gestalt psychology of fixed grouping principles to a more dynamic process based on statistics of the environment. (Figure 16 shows the model segmenting a simple image.) Both types of models use *heuristics* to guide the grouping process, rather than whole-object knowledge. Although the heuristics will not be infallable, the hope is that they will suffice for most segmentation tasks, and even when they fail, recognition processes will be robust to some degree of segmentation error (Enns & Rensink, 1992). This avoids the chicken-and-egg problems of how to segment a display without knowing the component objects, and how to recognize the objects without knowing the segmentation.[8]

Assuming some process has segmented the visual field into feature groups, how do the groups influence attention? Here is one proposal in terms of our model. The attentional model selects a single region—a contiguous set of locations—because each unit in the attentional map inhibits all units outside its neighborhood. However, for object-based attention, the possibility of selecting noncontiguous locations must be allowed. Thus, units representing locations of features of the same group should excite rather than inhibit one another. The result of grouping processes, then, should be to increase temporarily the connection strengths between attentional units that represent grouped locations. The notion of dynamic, short-term weight adjustments in response to grouped features was proposed by von der Malsburg (1981; von der Malsburg & Schneider, 1986).

The eventual attentional state will then be a complex interaction between the dynamic links formed among grouped features and exogenous and endogenous inputs to the attentional network. This brief sketch is hardly a compelling answer to the difficult and important question about how object-based and location-based attention work together. Existing computational models do not directly address how the two forms of attention are integrated, with the exception of preliminary work by Goebel (1993) and Grossberg, Mingolla, and Ross (1994). This is clearly fertile ground for future exploration and simulation.

[8] All segmentation models use some information about objects. The information can be as basic as the fact that two features appearing in a certain spatial relation are more often part of the same object than parts of different objects. The information can be as complex as restrictions on how a feature can appear with respect to all the other features that are part of the object, which we have referred to as *whole-object* knowledge. One can characterize the information along a continuum of what *order* statistics comprise the knowledge. Second-order statistics describe relationships between pairs of features; very high-order statistics are required to describe whole objects. The information used by the Mozer et al. (1992) model is of intermediate order, based on spatially local configurations of features. Vecera and Farah (1993) found that upright overlapping block letters are segmented more readily than the same stimuli inverted. This experiment rules out the use of only low-order statistics, such as continuity between pairs of lines, because upright and inverted letters are identical in terms of low-order statistics. While the use of whole-object knowledge for segmentation could explain the experimental results, the results are also consistent with the use of intermediate-order statistics that are different for upright and inverted letters. For example, English letters are more often open on the right than on the left. Inverted letters violate this configural property.

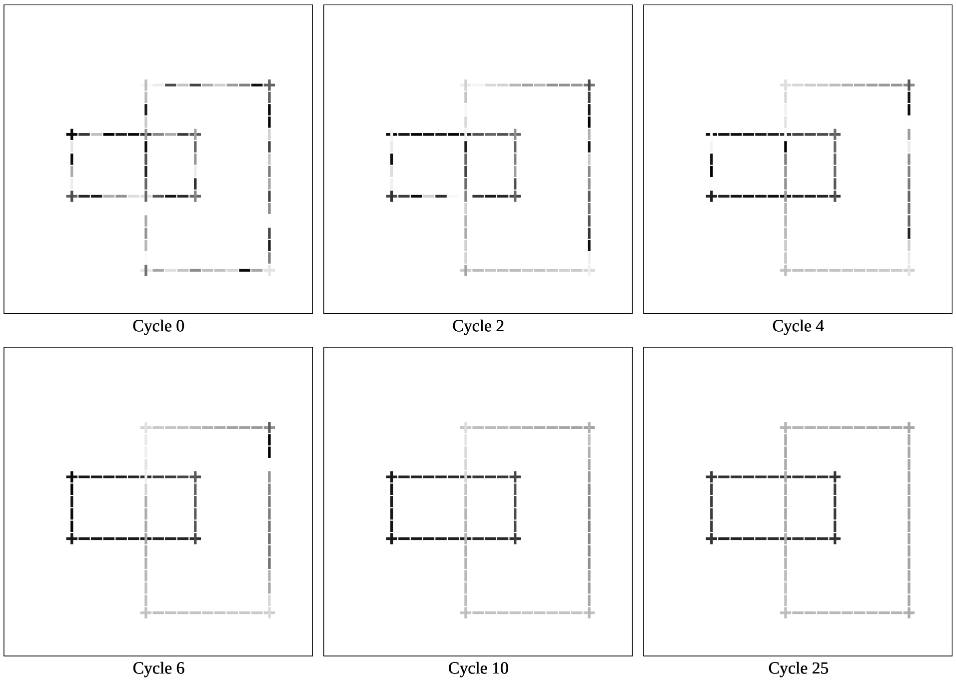

Figure 16: The adaptive grouping model of Mozer et al. (1992). The six panels show the state of the model at various points in processing a display consisting of two overlapping rectangles. The upper left panel is the initial state of the model; the lower right panel is the final output of the model. Each oriented line segment is a primitive input feature. The coloring of the features indicates the object label assigned to the features. The initially random labeling is transformed into a pattern in which the features of each rectangle have a unique label.

Simulations of visual search

In the previous three sections, we have discussed diverse aspects of attention: adjustable attentional spotlights, shifts of attention, selection on the basis of object attributes, and the relationship of location-based and object-based attention. All of these aspects must be addressed if we hope to model the vast and complex literature on visual search.

In a visual search task, subjects are commonly asked to detect the presence or absence of a target in a display containing distractor elements. Response time is measured as a function of the number of elements in the display. The shape of this curve indicates something about how subjects perform the search task. Flat curves, in which response time does not increase with the number of elements or increases very gradually (less than about 10 milliseconds per element), is suggestive of a parallel search across the visual field. Curves with steep slopes, in which each additional element increases the response time, are suggestive of a serial search. For example, searching for a vertical bar among horizontal bar distractors produces a flat curve; searching for a plus among vertical and horizontal bar distractors produces a positively sloped curve. Characterizing search using a serial-parallel dichotomy has turned out to be an oversimplification (see Chapter by Wolfe, this volume). Response time curves are often nonlinear, and slopes vary across tasks continuously, from flat to steep. It is thus more appropriate to view search on an easy-to-hard continuum.

A variety of promising computational models have been devised to replicate various aspects of the data (Ahmad, 1991; Ahmad & Omohundro, 1991; Gerrissen, 1991; Grossberg, Mingolla, & Ross, 1994; Humphreys & Müller, 1993; Mozer, 1991; Niebur & Koch, 1996; Sandon, 1990). Most of these models are based on feature-integration theory (Treisman & Gelade, 1980; Treisman & Gormican, 1988; Treisman & Sato, 1990) or the guided-search model (Wolfe, Cave, & Franzel, 1989). We will discuss the processes and mechanisms underlying visual search in terms of the model we have developed for this chapter, but our account overlaps significantly with these theories and earlier computational models. An outline of this account is as follows.

- We assume that the target and distractor sets are known in advance. For each primitive feature type, an analysis must be performed to determine how well the feature discriminates targets from distractors. That is, if all display elements containing (or not containing) the feature are discarded, have we done a good job in eliminating distractors and keeping targets? Consider, for example, searching for a red vertical among blue verticals and blue horizontals. If all red elements are ruled in (or equivalently, all blue elements are ruled out), the target has been reliably separated from the distractors. However, if all verticals are ruled in (or horizontals are ruled out), we are left with a set of elements that includes both targets and distractors.[9]

- The control signals, ρ_q, of highly discriminative feature types should be modulated such that potential target elements will be more likely to capture attention and potential distractor elements will be less likely. In our example of searching for a red vertical among blue verticals and horizontals, ρ_{red} should be increased, causing red elements to drive attention more than blue elements. The modulation of control signals might be subject to a limited amount of regulatory juice. The model also has the flexibility to adjust other parameters that influence its performance, including the diameter of the attentional spotlight, controlled by β, and the

[9] Judging the discriminative power of a feature requires additional assumptions about the nature of the stimulus displays, such as the relative likelihood of various distractors and the relative frequency of target-present trials.

response criterion, controlled by e_{NR}.[10]

- When a search display is presented, features in the display will drive the attentional network exogenously, gated by the control signals.
- A competition ensues within the attentional network to select one region. The region may contain one or multiple display elements. The size of the region will depend on the density and arrangement of elements, segmentation and grouping processes, and the adjustable parameter of the attentional network that controls the spotlight diameter.
- As selection takes place, display elements are processed and identified by the recognition network. Even display elements that are commonly thought of as simple features, such as a vertical bar, are processed by the recognition network. The vertical bar is an *object* which might be composed of vertical bar and terminator primitive features.
- The output layer of the recognition network contains a set of units that represent identities of the different display elements that might appear. Target detection would occur using the response initiation procedure of earlier simulations.
- If the target has not been detected by the time that the outputs of the recognition network have stabilized, the selected region is deemed not to contain a target, and attention should be prevented from returning there. This can be accomplished by forcing off the currently active attentional units, possibly by assigning them a strong negative bias that gradually decays back to zero, and resetting the recognition system.[11] [12]
- The attentional state is reset, and this process is repeated until all stimulus locations in the display have been explored, at which point the model reports "target absent." It is possible that the model could quit after only one or a small number of attentional fixations, or, at the other extreme, that it could return to locations to verify the absence of a target.

This is a complicated, ill-specified story, but visual search is a complicated, ill-specified task—ill specified in the sense that subjects must make a variety of strategic and control decisions that are not part of the task instructions. To simplify our simulation, we will model search in relatively small displays, of up to nine elements. This allows us to avoid limitations on peripheral visual acuity, eye movements, and—as we will show—the need for sequential attentional fixations. We also neglect target-absent data, because modeling performance on these displays requires additional mechanisms which, e.g., determine when to switch attention, when to quit searching, and how to suppress locations such that they are not repeatedly searched.

[10] It is a difficult optimization problem to configure the model's parameters so as to minimize errors or response time, especially under the constraint of a finite amount of regulatory juice. Fine tuning the system parameters is no doubt a matter of learning and experience.

[11] A bias is a tonic input to a unit. A negative bias causes the unit to shut off unless there is overwhelming positive input to the unit via excitatory connections from other units.

[12] Although we have not specified the coordinate frame in which the attentional units operate, it seems most natural to interpret it as retinotopic. There is evidence, however, that inhibition of return, the likely mechanism for preventing the human visual system from returning to an already searched location, operates in a coordinate frame that does not depend on eye position (Posner & Cohen, 1984).

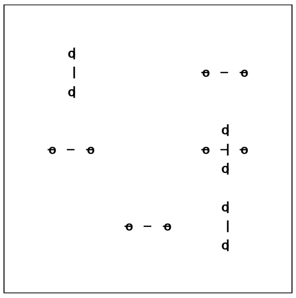

Figure 17: A pattern of activity that corresponds to five bars and a plus in random locations on the model's retina.

Simulation methodology

For these simulations, we trained a version of the recognition model that recognizes three "objects": a vertical bar, a horizontal bar, and a plus sign. The objects can appear in any of nine locations in the field. Figure 17 shows a sample display. Note the distinction between vertical-bar objects and vertical-bar features; the former is composed of the latter. The network was trained on 450 example displays of one to nine elements, similar to those used in the visual search simulations described below. The training set was unbiased in that it contained equal numbers of examples from each condition in the visual search simulations.

During testing, displays are presented to the model with a target and a variable number of distractors. The elements are arranged randomly on the model's retina. The elastic-spotlight attentional network, guided by control signals, selects a subset of the display elements, and the recognition network reports the identities of the selected elements. Although we imagine that detection responses are trigged by the response initiation process described earlier, we took a short cut which is a deterministic approximation to the stochastic process, and simply used a fixed activity threshold, generally around .5, as the all-or-none threshold for initiating a response. If a response has not been initiated within 100 cycles, the model reports "target absent" and is considered to have made an error.[13] The threshold we selected was as low as possible, to produce responses as fast as possible, such that the rate of false detection in target-absent displays was nearly zero.

Simple feature search

Searching for an element with a distinctive feature is easy. In a display containing a single vertical target among a variable number of horizontal distractors, response time will be independent of the number of distractors. Our model can explain this finding by assuming that the control signal for the distinctive feature is increased, causing attention to be driven directly to the location of the distinctive feature. Once that location is attended, the object at that location is recognized and a response is made.

We have simulated the search for a vertical among horizontals and a horizontal among verticals. On each simulation trial, the control signal for the primitive feature unique to the target is increased from .8 to 1.0, and the control signal for the primitive feature unique to the distractor is decreased from .8 to 0.

[13]The processing involved in deciding the target is absent is undoubtably more complex than this. Indeed, Chun and Wolfe (in press) suggests that absent responses are unlikely to be triggered by a fixed passage of time.

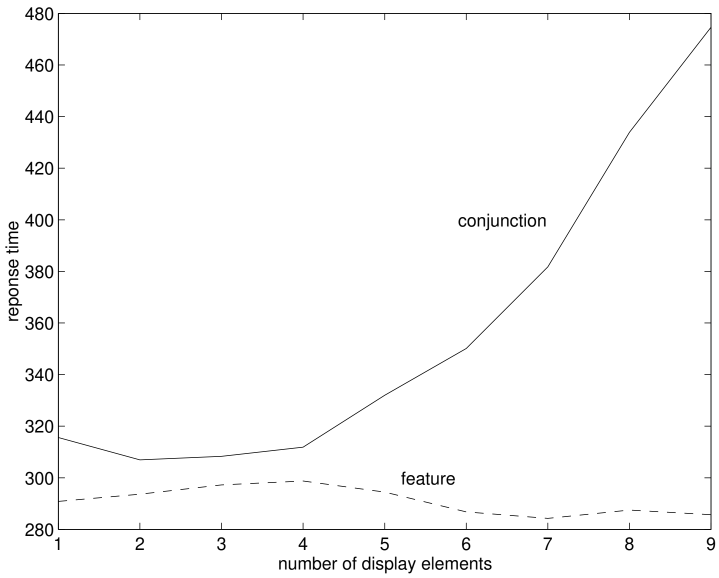

Figure 18: Time for the model to initiate a detection response as a function of display size for target-present trials in feature (dashed line) and conjunction (solid line) search.

The dashed curve in Figure 18 shows the model's performance on target-present trials as a function of the number of display elements. Response times are not dependent on display size. The model never fails to detect the target.

Theories of visual search generally assume that feature search does not require selective attention, and more strongly, that feature search does not benefit from selective attention. We tested whether this assumption is consistent with our model by forcing attention to be distributed across the visual field. This is achieved by setting $\lambda = 1$, which causes all perceptual data to be fully analyzed by the recognition network, regardless of the attentional state. One might conjecture that if simple feature displays can be processed in parallel and if there is a benefit of allocating attention prior to stimulus onset, as we observed in the cue-validity effect, response times might actually be *faster* with distributed attention. Indeed, there is a statistically reliable benefit for small displays, replicating the cue-validity effect, but there is also a statistically reliable cost for large displays. Cost and benefit are both on the order of 30 ms, and over the various display sizes, they tend to cancel. Thus, the attentional network is not really helping processing for simple feature displays, nor is it hurting, consistent with the traditional view of feature search.

The model offers a nontraditional perspective in two other respects, however. First, simple feature search is viewed as an object recognition task, albeit one which the recognition system has capacity to perform in parallel. Second, while the guided-search model and feature-integration theory consider the role of attentional guidance only in conjunction search, modulation of control signals is critical in our model in feature search. Because the attentional network always acts to select display elements, it is necessary to modulate control signals to select the target features or else the target will be suppressed and may not be detected. It would be a challenge to develop an experimental test that could distinguish this perspective from the traditional view.

Conjunction search

Subjects are slow to search for an element defined by a conjunction of features, such as a red vertical target among red horizontal and blue and red vertical distractors. An explanation in terms of our model for the difficulty of conjunction search is not obvious. Suppose that control signals were set such that exogenous input from only the red and vertical feature maps was able to reach the attentional network. Locations of red elements would receive a certain amount of input, locations of vertical elements would receive roughly same input, but the locations of red vertical elements would receive twice as much input. The attentional network should reliably select the location of the target, independent of the number of distractors. Regardless of recognition and verification processes, one would expect the response curve to be flat, in contrast to typical human data. Thus, it might seem that our model is too powerful, even though there appear to be at least some conjunction searches that are easy (e.g., color/depth and motion/depth, Nakayama & Silverman, 1986).

One account of the difficulty of conjunction search, suggested by the guided-search model (Wolfe et al., 1989) is to postulate that recognition and attention operate in an intrinsically noisy environment. Although the attentional system should be directed more strongly to the target location than to the distractor locations, the strength of the direction may not be sufficient to overcome noise, and will therefore not be reliable, and serial search will be required. A second way account of the difficulty of conjunction search is to postulate limits on the voluntary adjustment of the control signals—the regulatory juice. These two accounts are complementary; weak limits on regulatory juice and a high intrinsic noise level should yield performance similar to that with strong limits on regulatory juice and a low intrinsic noise level. In simulating our model, we discovered that it provides a somewhat different account altogether, which we detail below.

In the canonical conjunction search task, the target is composed of features on two different dimensions. We could simulate this experiment by adding red and blue feature types to the model and then training a net to recognize red and blue verticals and horizontals. We could then use the control signals to bias attention toward the red and vertical feature maps, if the target was a red vertical. Instead, we have chosen to simulate an experiment which is more challenging to the model. Our simulation experiment involves a target plus symbol embedded in a distractor array of verticals and horizontals. Even without modulating the default values of the control signals, the exogenous input to the target location should be twice that of the exogenous input to any of the distractors because the target is composed of twice as many features.[14] It would thus seem that selection should be strongly biased toward the location of the target—a problematic result for the model.

With trepidation, we ran the conjunction search simulation. To our surprise, the model's performance nicely matched the human data, as shown by the solid curve in Figure 18.[15] As the number of display elements increases, response times increase. For small displays, the curve is flat. This is in accord with the finding of Pashler (1987) that nearly flat search slopes can be observed for small displays, and it reflects the fact that the recognition network is able to detect conjunctions in parallel, albeit with limited capacity. (Mordkoff, Yantis, and Egeth, 1990, present further experimental evidence of limited-capacity parallel conjunction detection.)

Response times increase with display size for two reasons. First, the competition among elements

[14] The guidance to the target is the same as it would be in the colored bar experiment if the control signals were set up to allow only target features to guide attention, i.e., assuming no limit on the regulatory juice.

[15] By experimentation, we found that the model performed best when the control signals for all feature types were lowered from .8 to .5 and when β was lowered from .10 to .04, creating a narrower focus of attention.

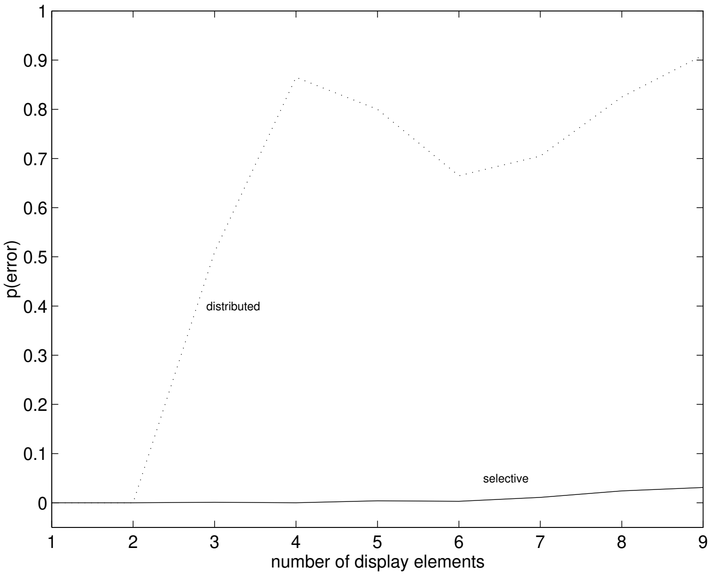

Figure 19: Error rate of the model on target-present trials for conjunction search as a function of display size. The solid line represents the condition in which the attentional network performs selective attention. The dotted line represents the condition in which attention is distributed across the field, i.e., all perceptual data enters the recognition network.

in the attention network increases. This can be shown by observing how long the attentional network requires to reach a stable state. Second, unattended elements in the display interfere with recognition. This can be shown by comparing performance of the network with $\lambda = 0$, i.e., unattended information fully suppressed, to the standard model, which has $\lambda = .05$. We find that response times are statistically slower with unattended information fully suppressed, 39 ms slower in the case of nine-element displays.

As we did with feature search, we can examine how important selective attention is for conjunction search. Here, we find a very different pattern. Figure 19 shows that the model's error rate skyrockets when attention is divided across display elements. No setting of the response threshold can achieve a low error rate over both target present and absent trials. The model cannot reliably detect the plus target without selective attention, consistent with the traditional theories of visual search.

However, our account of conjunction search is in part nontraditional, because it depends on subtle properties of the model—the influence of unattended elements on the detection of attended elements and the dynamics of the attentional model. Simulation studies were critical to discovering that the model behaved correctly and why it did. Although the current simulation did not require postulating noise in the attentional system or limitations on regulatory juice, these factors may contribute to conjunction search performance, and may be necessary in simulations of other experimental findings.

Discussion of visual search

In its present form, the model can explain other data relating to visual search, including the findings of faster search in low-density displays (Cohen & Ivry, 1991) and the difficulty of detecting the absence of features (Treisman & Souther, 1985). With minor extensions to the model, a wide variety of other data can be addressed, including response-time curves for target-absent displays, effects of target-distractor contrast (Treisman & Gormican, 1988), search asymmetries (Ivry & Cohn, 1992; Treisman & Gormican, 1988), effects of distractor homogeneity (Duncan & Humphreys, 1989), and rapid conjunction search (Wolfe et al., 1989). However, our immediate goal is not to present a comprehensive model of visual search, but rather to begin considering the underlying mechanisms. By addressing data from experimental paradigms as disparate as spatial cueing and visual search, we hope to have convinced the reader of the model's breadth and flexibility. This is the remarkable property of computational models—they can help one to integrate phenomena under a unified framework. The other lesson from these simulations of visual search is that, although the model shows some behavior that one would intuitively expect, other aspects of its behavior were found only via simulation. The model raises some intriguing possibilities, and addressing these possibilities requires further human experimental studies.

The role of selective attention

When one adopts a computational perspective, a natural question to ask is what computational role selective attention plays in visual information processing. Four distinct functional roles of attention fall naturally from the computational perspective presented in this chapter.

- *Controlling order of readout.* The attentional system allows the recognition system to selectively access information in the visual field by location. A task requiring sequential responses to items in various locations could not be carried out with the recognition system alone.

- *Reducing crosstalk.* As we illustrated earlier, when the recognition network analyzes multiple items in parallel, interactions within the network cause the processing of one item to interfere with another. Deploying attention to one or a small number of items at once will reduce or eliminate crosstalk.

- *Recovering location information.* The output of the recognition system we developed encodes identities but not locations. Computationally, it makes sense to separate identity from location, because often the same response should be made to a stimulus regardless of where it appears in the visual field. Neurophysiological evidence also suggests that, at least in the responses of individual cells, a great deal of location information is discarded in higher cortical areas involved in object recognition (Tanaka, 1993). And some psychological data suggests that stimulus identity is encoded apart from location (Mozer, 1989; Kanwisher, 1990). Thus, some means of recovering location information is critical. Because the current locus of attention reflects the spatial source of activations in the object recognition system, the attentional system can convey the discarded location information.

- *Coordinating processing performed by independent modules.* The heart of feature-integration theory is the notion that visual stimuli are analyzed by functionally independent modules specialized along certain attribute dimensions such as color, form, and motion. Because these modules operate autonomously, it is imperative to ensure that they coordinate their processing efforts. Otherwise, the system can encounter a *binding problem* in which attributes of multiple objects are simultaneously activated and it cannot be determined which attributes belong together, possibly resulting in illusory conjunctions (Treisman & Schmidt, 1982). By guiding all modules to analyze the same spatial region, attention can ensure that the attributes of a single object will be bound together.[16]

Contrasting theoretical perspectives on selective attention

The key properties of our model are common to most theories of selective attention. We summarize these properties, which collectively we call the *spatial-selection perspective*, as follows.

- Attention operates as a spatial gating mechanism. This is mandatory to perform selection by location.
- The mechanism includes a representation of visual field location—the attentional map—which is distinct from the representation of visual features used for object recognition.
- Attention acts to modulate the activity of visual features such that the signal strength of features at attended locations are enhanced relative to the strength of features at unattended locations.
- Object recognition is limited in capacity. While there may be some capacity to recognize objects in parallel, interference among objects arises which necessitates attentional selection early in the processing stream. Although selection is performed early, unattended information receives some degree of processing and causes some interference with attended information.
- Selection can be performed on the basis of object attributes, if these attributes can be characterized in terms of combinations of primitive features that discriminate the item of interest from other items in the visual field.
- Perceptual grouping operates prior to attentional selection and can influence the deployment of spatial attention.

An alternative theoretical perspective on selective attention has been suggested recently in which competition is ubiquitous and is not limited to competition among locations (Allport, 1993; Desimone & Duncan, 1995; Duncan, 1996; Phaf, van der Heijden, & Hudson, 1990). We call this the *ubiquitous-competition perspective*, and highlight its the main properties as follows.

- Attention is viewed as the competition among stimulus representations at many loci in the processing stream, from sensory input to response formation. Objects might compete within subsystems that represent color, shape, and location information, as well as a subsystem that represents possible actions.
- Within each subsystem, a winner-take-all process results in a gain in activity or representation for one object and a loss for others.
- The competitive mechanisms are integrated such that multiple subsystems tend to work concurrently on the same object.

[16]Note that this statement is not as strong as the claim of feature-integration theory that attention is *necessary* to perform all types of binding. Even if intra-dimensional bindings are performed automatically, or if experience might allow inter-dimensional bindings to be performed automatically, or if the modules are only weakly independent, there is still a role for attention to coordinate processing.

- Priming of representations within any subsystem acts to guide selection (see also Farah, 1994; Grossberg, Mingolla, & Ross, 1994; Mozer, 1991).
- Selection by location is no more fundamental than selection on the basis of other stimulus or response dimensions.

It is beyond the scope of this chapter to try to resolve the differences between this perspective and the one we have presented. However, we point out that the two are not altogether incompatible. One can accept the primacy of location-based selection, but also allow for competition among higher-order object representations. For example, in the model we have presented, inhibitory connections could be added between units that represent different letters, forcing a selection of a single letter. This competition among identity representations would be useful for response selection; the process could even be primed to a particular letter by preactivating the appropriate letter unit, in accord with the ubiquitous-competition perspective.

The difference between the two perspectives is primarily one of emphasis, the spatial-selection perspective addressing capacity limitations in object recognition and the ubiquitous-competition perspective focusing on the diverse sorts of cues that can be used for selection. However, the two perspectives suggest quite different mechanisms of selection on the basis of object identity. The ubiquitous-competition perspective allows for competition to operate fairly late in processing among high-level object representations, and then for cooperation among the subsystems to work its way back to select the same object everywhere in the processing stream. The spatial-selection perspective, as we have elaborated, suggests a variety of "quick and dirty" heuristics to guide spatial attention to objects of interest. It remains to be seen which perspective will be most useful in explaining the broad and complex corpus of psychological data on attentional selection.

Issues in computational modeling

We have presented an elaborate computational framework for analyzing and understanding spatial attention. Our goal has not been to convince you that the framework is necessarily correct, but rather that modeling is a valuable exercise that allows one to reason in concrete terms about the computational mechanisms. We suspect that some readers will still be skeptical as to the value of model building. For this reason, we conclude with a discussion of general issues in computational modeling.

Why build computational models?

It goes against the tradition of experimental psychology to construct large, complex computational models with dozens to hundreds of parameters. Nonetheless, as the field matures, computational models should play an increasingly important role, for the following reasons.

- As one tries to explain larger and larger bodies of data and data from diverse experimental paradigms, the complexity of the model must necessarily increase. Computational models with many components and parameters thus become better justified.
- Computational models provide a framework for integrating knowledge from behavioral studies with results from fields as neuroanatomy and neurophysiology.
- Computational models force one to be explicit about one's hypotheses and assumptions. To test a computational mechanism, it must be specified with precision and detail.

- Computational models provide the ultimate in controlled experimentation. Any simulation experiment can be replicated exactly. Stimulus materials can be generated that differ just on the dimension of interest, without any confounding factors. One can poke at and examine any part of the model. One can precisely lesion or adjust individual components of the model and observe the consequences.

- Computational models can make empirical predictions. The model can be presented with novel stimuli or a novel experimental paradigm, and its performance can be compared to that of human subjects. The ability of the model to predict nontrivial experimental findings that it was not explicitly designed to explain is an indication that the model correctly captures some aspect of cognition. In the best of circumstances, an experiment can be designed to distinguish predictions of one model or model class from another, thereby providing not just support for one model but evidence against another. Of course, the ability to predict experimental results is not unique to computational models.

- Computational models allow one to observe the consequences of interactions among mechanisms. In many models, the effect of changing one component trickles to others. It is difficult to anticipate these effects without computer simulation.

- Computational models help one to understand the tradeoffs involved in the design of the cognitive architecture. It is our conviction that limitations in human cognition are not arbitrary, but are the result of sensible, if not optimal, design decisions given various constraints on the cognitive architecture.

We do not mean to suggest that the mere fact that a model has been implemented in computer simulation gives it some intrinsic value, nor the fact that a model is described qualitatively instead of using equations implies that the model has little value. Any model is useful only to the extent it helps us understand some aspect of cognition.

What makes a model compelling?

A simple model that can explain a large, diverse corpus of data is very compelling. However, characterizing the complexity of a model is not a trivial task. For linear models, the number of parameters is a measure of model complexity and of how many data points it is guaranteed to account for. For nonlinear models such as connectionist models, no such direct relationship exists. Some parameters give the model a lot of flexibility, others practically none. For example, in our model, any individual connection strength in the recognition network can be changed with little effect on the model's qualitative or quantitative behavior; however, a parameter like λ, which determines the degree to which unattended information will be processed, dramatically affects the qualitative behavior of the model.

Perhaps the complexity of a model should be measured in terms of how many basic principles it embodies, rather than the total number of parameters. For example, our recognition model, while it has several thousand parameters, embodies just a few principles—local receptive fields, convergence of information from different regions of the retina, and so forth. The specific number of feature types in each layer and the specific pattern of connectivity is probably not central to the model's qualitative behavior.[17]

[17] To determine which aspects of the model are key and which are incidental, one must conduct simulation studies over a variety of different architectures. Unfortunately, this is computation intensive work, and is seldom done.

One question to ask when evalutating a model is whether more falls out of the model than has been built into it, that is, whether the model has emergent properties. A clear demonstration of emergent properties is when the model can make novel empirical predictions that are eventually validated. However, this is not the only criterion by which a model can be judged as useful. The Occam's Razor argument is that if a simple model can explain complex patterns of data, then there is likely to be some truth in the model, regardless of whether the data are old or new.[18] Ultimately, it is up to the reader to determine whether the model is indeed simple relative to the amount of data it explains.

When is a model right or wrong?

Odds are that the model is wrong, at least in some detail. This is not to say that the model has no value; it may be one's current best theory, and the only way one has of contemplating mechanisms of behavior. When the model makes a concrete prediction and this prediction is incorrect, one faces the challenge of modifying the model to incorporate the new effect. More often, the model will not be sufficiently well specified to predict the outcome of an experiment; in this case, the model will need to be elaborated to account for results. Thus, over time, the complexity of the model will grow as the corpus of data it can explain grows. If the model is a good one, the model's rate of growth will be far lower than the growth of the corpus. Each time the model is modified or elaborated, it becomes further constrained. Eventually, someone is likely to devise an experiment that the model is simply not suited to explain. At this point, the model has run its useful life, and a fresh conception of the underlying mechanisms is demanded.

What about other models that also explain the data?

One question that modelers are constantly asked is: Why should one believe in a particular model when there are probably dozens of models that are just as effective in explaining the corpus of data? The response of modelers is usually amusement; it is extremely difficult to build one model that can explain the data, let alone a hundred. Those who have never built a model often fail to appreciate this fact. The appropriate response is perhaps to challenge the questioner to propose an alternative model. Then, experiments can be devised for which the models make different predictions, or else the models are functionally equivalent.

Depth versus breadth in modeling

Ultimately, one would like a model both broad and deep, "broad" in that it can address a variety of experimental tasks and response paradigms, and "deep" in that it can explain subtleties and quantitative properties of the data. Traditionally, psychological models have aimed for depth over breadth, and the cost has been that a model of one phenomenon, say the word superiority effect (McClelland & Rumelhart, 1981) may have little in common with a model of some other phenomenon, say the Stroop task (Cohen, Dunbar, & McClelland, 1990), even though the two models are ostensibly of the same fundamental process, reading in this case.

[18] In model building, the distinction between old and new data is seldom clear. One often constructs the model with particular data in mind, and then discovers that the model, with no or minor changes, can explain other data as well. In this case, the additional data are in fact predicted by the model, even though the data may have been collected and published before the model was developed.

We have aimed for breadth in our presentation by discussing data across a variety of experimental tasks and paradigms. A consequence of this choice is a model with multiple components and parameters which can be configured differently for different tasks it is asked to perform. For a particular task, we presented arguments for why the model should be configured a certain way. This "configuration" includes specifying decision criteria, modulating control signals, and adjusting the diameter of the attentional spotlight, and in a more complete model it might also include the vigilance level (the degree to which units are modulated by the attentional network, the parameter λ), exogenous guidance of attention, and priming to bias selection.

When subjects are given verbal task instructions, they are able to configure their perceptual systems appropriately for the task. In addition to producing the right response to a stimulus—whether the response is a foot tap when a vowel is presented or a spoken report of the number of display items—response criteria are adjusted to trade off speed and accuracy, and performance is optimized, e.g., searching a display in parallel if subjects are capable of doing so. Understanding and modeling this configuration process is a tremendous challenge ahead for the next generation of computational models of human cognition.

Acknowledgements

This research was supported by NSF Presidential Young Investigator award IRI–9058450, grant 90–21 from the James S. McDonnell Foundation, and a grant from Lifestyle Technologies. Our thanks to Harold Pashler and James Juola for their fine editorial feedback, and to many helpful discussions with participants in a seminar on computational models of attention, particularly Don Mathis, Clark Fagot, Sigal Adoot, Richard Beach, Gina Cherry, Braden Craig, Julia Fisher, Audrey Guzik, Tracy Hansen, Carol Kealoha, Deb Miller, Bill Raymond, and Eran Tari.

References

Ahmad, S. (1991). *VISIT: An efficient computational model of human visual attention* (ICSI Technical Report 91-049). Berkeley, CA: International Computer Science Institute.

Ahmad, S., & Omohundro, S. (1991). Efficient visual search: A connectionist solution. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (pp. 293–298). Hillsdale, NJ: Erlbaum.

Allport, D. A. (1993). Attention and control: Have we been asking the wrong questions? A critical review of twenty-five years. In D. E. Meyer & S. Kornblum (Eds.), *Attention and Performance XIV* (pp. 183–218). Cambridge, MA: MIT Press.

Ballard, D. H. (1986). Cortical connections and parallel processing: Structure and function. *The Behavioral and Brain Sciences*, *9*, 67–120.

Barlow, H. H. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, *1*, ???–???.

Behrmann, M., Zemel, R. S., & Mozer, M. C. (1996). Object-based attention and occlusion: Evidence from normal subjects and a computational model. *Journal of Experimental Psychology: Human Perception and Performance.* Submitted for publication.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*, 523–547.

Chun, M. M., & Wolfe, J. M. (1996). Just say no: How are visual search trials terminated when there is no target present? *Cognitive Psychology*, *10*, 39–78.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review*, *97*, 332–361.

Cohen, J. D., Romero, R. D., Farah, M. J., & Servan-Schreiber, D. (1994). Mechanisms of spatial attention: The relation of macrostructure to microstructure in parietal neglect. *Journal of Cognitive Neuroscience*, *6*, 377–387.

Cohn, A., & Ivry, R. B. (1991). Density effects in conjunction search: Evidence for a coarse location mechanism of feature integration. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 891–901.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222.

Downing, C. G., & Pinker, S. (1985). The spatial structure of visual attention. In M. I. Posner & O. S. M. Marin (Eds.), *Attention and performance XI* (pp. 171–187). Hillsdale, NJ: Erlbaum.

Driver, J., & Baylis, G. C. (1989). Movement and visual attention: The spotlight metaphor breaks down. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 448–456.

Driver, J., McLeod, P., & Dienes, Z. (1992). Are direction and speed coded independently by the visual system? Evidence from visual search. *Spatial Vision*, *?*, 133–147.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, *113*, 501–517.

Duncan, J. (1987). Attention and reading: Wholes and parts in shape recognition – A tutorial review. In M. Coltheart (Ed.), *Attention and performance XII* (pp. 39–62). Hillsdale, NJ: Erlbaum.

Duncan, J. (1995). Target and nontarget grouping in visual search. *Perception and Psychophysics*, *57*, 117–120.

Duncan, J. (1996). Coordinated brain systems in selective perception and action. In T. Iaui & J. L. McClelland (Eds.), *Attention and performance XVI*. Cambridge, MA: MIT Press.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*, 433–458.

Egeth, H. (1977). Attention and preattention. In G. H. Bower (Ed.), *The psychology of learning and motivation, Volume 11* (pp. 277–320). New York: Academic Press.

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, *123*, 161–177.

Enns, J. T. (1990). Three-dimensional features that pop out in visual search. In D. Brogan (Ed.), *Visual Search* (pp. 37–46). London: Taylor and Francis.

Enns, J. T., & Rensink, R. A. (1992). A model for the rapid interpretation of line drawings in early vision. In D. Brogan (Ed.), *Visual search II.* London: Taylor and Francis. In press.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*, 143–149.

Eriksen, C. W. (1990). Attentional search of the visual field. In D. Brogan (Ed.), *Visual Search* (pp. ???–????). London: Taylor Francis.

Eriksen, C. W., & Hoffman, J. E. (1973). The extent of processing of noise elements during selective coding from visual displays. *Perception and Psychophysics*, *14*, 155–160.

Eriksen, C. W., & Hoffman, J. E. (1974). Selective attention: Noise suppression or signal enhancement? *Bulletin of the Psychonomic Society*, *4*, 587–589.

Eriksen, C. W., & Schultz, D. W. (1979). Information processing in visual search: A continuous flow conception and experimental results. *Perception & Psychophysics*, *25*, 249–263.

Eriksen, C. W., & St. James, J. D. (1986). Visual attention within and around the field of focal attention: a zoom lens model. *Perception and Psychophysics*, *40*, 225–240.

Eriksen, C. W., & Yeh, Y.-Y. (1985). Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance*, *11*, 583–597.

Farah, M. J. (1994). Neuropsychological inference with an interactive brain: A critique of the "locality" assumption. *Behavioral and Brain Sciences*, *17*, 43–104.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science*, *6*, 205–254.

Fukushima, K., & Miyake, S. (1982). Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition*, *15*, 455–469.

Gatti, S. V., & Egeth, H. E. (1978). Failure of spatial selectivity in vision. *Bulletin of the Psychonomic Society*, *11*, 181–184.

Gerrissen, J. F. (1991). On the network-based emulation of human visual search. *Neural Networks*, *4*, 543–564.

Goebel, R. (1993). Perceiving complex visual scenes: An oscillator neural network model that integrates selective attention, perceptual organisation, and invariant recognition. In S. J. Hanson, J. D. Cowan, & C. L. Giles (Eds.), *Advances in neural information processing systems 5* (pp. 903–910). San Mateo, CA: Morgan Kaufmann.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding. I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121–134.

Grossberg, S., Mingolla, E., & Ross, W. D. (1994). A neural theory of attentive visual search: Interactions of boundary, surface, spatial, and object representations. *Psychological Review*, *101*, 470–489.

Hillstrom, A. (1995). Singleton pop-out: Facilitation of uniqueness or inhibition of similarity?. Psychonomics abstract.

Hillstrom, A. P., & Yantis, S. (1994). Visual motion and attentional capture. *Perception and Psychophysics*, *55*, ???–???.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I: Foundations* (pp. 77–109). Cambridge, MA: MIT Press/Bradford Books.

Humphreys, G. W., & Müller, H. J. (1993). SEarch via Recursive Rejection (SERR): A connectionist model of visual search. *Cognitive Psychology*, *25*, 43–110.

Ivry, R. B., & Cohn, A. (1992). Asymmetry in visual search for targets defined by differences in movement speed. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 1045–1057.

Jackson, S. R., Marrocco, R., & Posner, M. I. (1994). Networks of anatomical areas controlling visuospatial attention. *Neural Networks*, *7*, 925–944.

Jonides, J., & Yantis, S. (1988). Uniqueness of abrupt visual onset in capturing attention. *Perception & Psychophysics*, *43*, 346–354.

Kahneman, D., & Henik, A. (1981). Perceptual organization and attention. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 181–211). Hillsdale, NJ: Erlbaum.

Kahneman, D., Treisman, A., & Burkell, J. (1983). The cost of visual filtering. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 510–522.

Kanwisher, N. G. (1990). Binding and type-token problems in human vision. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society* (pp. 606–613). Hillsdale, NJ: Erlbaum.

Kinchla, R. A. (1992). Attention. *Annual Review of Psychology*, *43*, 711–742.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227.

Kramer, A. F., & Jacobson, A. (1991). Perceptual organization and focused attnetion: The role of objects and proximity in visual processing. *Perception and Psychophysics*, *50*, 267–284.

LaBerge, D. (1983). Spatial extent of attention to letters and words. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 371–379.

LaBerge, D., & Brown, V. (1989). Theory of attentional operations in shape identification. *Psychological Review*, *96*, 101–124.

LaBerge, D., Brown, V., Carter, M., Bash, D., & Hartley, A. (1991). Reducing the effects of adjacent distractors by narrowing attention. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 90–95.

Le Cun, Y., Boser, B., Denker, J. S., Hendersen, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*, 541–551.

Maljkovic, V., & Nakayama, K. (1994). Priming of pop-out: I. Role of features. *Memory & Cognition*, *22*, 657–672.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological Review*, *88*, 375–407.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I: Foundations* (pp. 3–44). Cambridge, MA: MIT Press/Bradford Books.

Miller, J. (1991). The flanker compatibility effect as a function of visual angle, attentional focus, visual transients, and perceptual load: A search for boundary conditions. *Perception & Psychophysics*, *49*, 270–288.

Milner, P. M. (1974). A model for visual shape recognition. *Psychological Review*, *81*, 521–535.

Mordkoff, J. T., Yantis, S., & Egeth, H. E. (1990). Detecting conjunctions of color and form in parallel. *Perception and psychophysics*, *48*, 157–168.

Mozer, M. C. (1983). Letter migration in word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 531–546.

Mozer, M. C. (1989). Types and tokens in visual letter perception. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 287–303.

Mozer, M. C. (1991). *The perception of multiple objects: A connectionist approach.* Cambridge, MA: MIT Press/Bradford Books.

Mozer, M. C., & Behrmann, M. (1990). On the interaction of selective attention and lexical knowledge: A connectionist account of neglect dyslexia. *Cognitive Neuroscience*, *2*, 96–123.

Mozer, M. C., Halligan, P. W., & Marshall, J. C. (1996). The end of the line for a brain-damaged model of hemispatial neglect. *Cognitive Neuroscience.* In Press.

Mozer, M. C., Zemel, R. S., Behrmann, M., & Williams, C. K. I. (1992). Learning to segment images using dynamic feature binding. *Neural Computation*, *4*, 650–666.

Nakayama, K., & Silverman, G. H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature*, *320*, 264–265.

Niebur, E., & Koch, C. (1996). Control of selective fvisual attention: Modeling the "where" pathway. In D. S. Touretzky, M. C. Mozer, & M. Hasselmo (Eds.), *Neural Information Processing Systems IX.* Cambridge, MA: MIT Press.

Nissen, M. J. (1985). Accessing features and objects: Is location special? In M. I. Posner & O. S. M. Marin (Eds.), *Attention and performance XI* (pp. 205–219). Hillsdale, NJ: Erlbaum.

Pashler, H. (1987). Detecting conjunctions of color and form: Reassessing the serial search hypothesis. *Perception & Psychophysics*, *41*, 191–201.

Pashler, H. (1988). Cross-dimensional interaction and texture segregation. *Perception & Psychophysics*, *43*, 307–318.

Pashler, H., & Badgio, P. C. (1987). Attentional issues in the identification of alphanumeric characters. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 63–82). Hillsdale, NJ: Erlbaum.

Phaf, R. H., Van der Heijden, A. H. C., & Hudson, P. T. W. (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology*, *22*, 273.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*, 3–25.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. In H. Bouma & D. Bouwhuis (Eds.), *Attention and performance X* (pp. 531–556). Hillsdale, NJ: Erlbaum.

Posner, M. I., Snyder, C. R. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*, 160–174.

Rensink, R. A., & Enns, J. T. (1995). Preemption effects in visual search: Evidence for low-level grouping. *Psychological Review.*

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I: Foundations* (pp. 318–362). Cambridge, MA: MIT Press/Bradford Books.

Sandon, P. A. (1990). Simulating visual attention. *Cognitive Neuroscience*, *2*, 213–231.

Sandon, P. A., & Uhr, L. M. (1988). An adaptive model for viewpoint-invariant object recognition. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 209–215). Hillsdale, NJ: Erlbaum.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search and attention. *Psychological Review*, *84*, 1–66.

Shaffer, W. O., & LaBerge, D. (1979). Automatic semantic processing of unattended words. *Journal of Verbal Learning & Verbal Behavior*, *18*, 413–426.

Shiffrin, R. M., & Gardner, G. T. (1972). Visual processing capacity and attentional control. *Journal of Experimental Psychology*, *93*, 72–83.

Shiu, L.-P., & Pashler, H. (1994). Negligible effect of spatial precuing on identification of single digits. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1037–1054.

Shulman, G. L., Remington, R., & McLean, J. P. (1979). Moving attention through visual space. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 522–26.

Snyder, C. R. (1972). Selection, inspection, and naming in visual search. *Journal of Experimental Psychology*, *92*, 428–431.

Sperling, G., & Weichselgartner, E. (1995). Episodic theory of the dynamics of spatial attention. *Psychological Review*, *3*, 503–532.

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

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, *95*, 15–48.

Treisman, A., & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 459–478.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, *14*, 107–141.

Treisman, A., & Souther, J. (1985). Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, *114*, 285–310.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, *101*, 80–102.

Tsal, Y. (1983). Movements of attention across the visual field. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 523–530.

Tsal, Y., & Lavie, N. (1988). Attending to color and shape: The special role of location in selective visual processing. *Perception & Psychophysics*, *44*, 15–21.

Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *Brain and Behavioral Sciences*, *13*, 423.

Tsotsos, J. K. (1991). Computational resources do constrain behavior. *Brain and Behavioral Sciences*, *13*, 506.

Uhr, L. (1987). *Highly parallel, hierarchical, recognition cone perceptual structure* (Technical Report 688). Madison, WI: Computer Sciences Department, University of Wisconsin.

Ullman, S. (1984). Visual routines. *Cognition*, *18*, 97–159.

Vecera, S. P., & Farah, M. J. (1994). Does visual attention select objects or locations? *Journal of Experimental Psychology: General*, *123*, 146–160.

von der Malsburg, C. (1981). *The correlation theory of brain function* (Internal Report 81-2). Goettingen: Department of Neurobiology, Max Planck Institute for Biophysical Chemistry.

von der Malsburg, C., & Schneider, W. (1986). A neural cocktail-party processor. *Biological Cybernetics*, *54*, 29–40.

Weismeyer, M., & Laird, J. (1990). A computer model of 2D visual attention. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 582–589). Hillsdale, NJ: Erlbaum.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 419–433.

Yantis, S. (1988). On analog movements of visual attention. *Perception and Psychophysics*, *43*, 203–206.

Yantis, S., & Johnston, J. C. (1990). On the locus of visual attention: Evidence from focused attention tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 135–149.