Simulating Visual Attention

Peter A. Sandon
Dartmouth College

Abstract

Selective visual attention serializes the processing of stimulus data to make efficient use of limited processing resources in the human visual system. This paper describes a connectionist network that exhibits a variety of attentional phenomena reported by Treisman, Wollord, Duncan, and others. As demonstrated in several simulations, a hierarchical, multiscale network that uses feature arrays with strong lateral inhibitory connections provides responses in agreement with a number of prominent behaviors associated with visual attention. The overall network design is consistent with a range of data reported in the psychological literature, and with neurophysiological characteristics of primate vision.

INTRODUCTION

Computational models of visual object recognition involving parallel processing of data are popular in machine vision work. These models are well matched to the large amounts of input data that must be processed at the low-level. Pyramid models (Uhr 1972; Hanson and Riseman 1978; Dyer 1982) and connectionist (neural) network models (Fukushima 1980; Sabbah 1985) are examples of such data parallelism. Although it is clear that parallelism is necessary for processing the large amount of data representing visual input in a relatively short amount of time (Uhr 1980; Feldman 1985), complete parallelism is not possible, because it requires too many processors and connections (Tsotsos 1987). Instead, a balance must be found between processor-intensive parallel techniques and time-intensive serial techniques. One way to implement this trade-off is to process all data in parallel at the early stages of vision, and then to select only part of the available data for further processing at later stages. This serialization of the computation at higher processing levels is referred to as attention.

Surprisingly, little work has been done investigating attentional models for machine vision. With some exceptions, most machine vision work has implicitly assumed either a strictly serial or a strictly parallel paradigm. Meanwhile, cognitive psychologists have been studying attentional behavior in human subjects (cf. Treisman and Gelade 1980; Wollord and Shum 1980; Duncan and Humphreys 1989), and work has been done in studying the neural processing related to attention (Crick 1984; Andersen et al. 1985; Moran and Desimone 1985; Anderson and Van Essen 1987). Our purpose in the work described here is to make use of the data from these other fields, particularly the more extensive behavioral data, to develop a computational model of visual attention that will serve researchers in both human and machine vision.

We have begun the development of a connectionist network model of visual attention. In some preliminary simulations, we have obtained results that qualitatively match some of the attentional phenomena reported in the psychological literature. Feature integration (Treisman and Gelade 1980) has been central to theories of the psychology of attention for many years. We have chosen this theory and the supporting behavioral data as the basis for our initial studies. In addition, we have included related work by Duncan and by Wollord, which provide additional data and alternative theories of attention. The network model is designed to integrate these seemingly conflicting data, as well as other attentional effects, while satisfying goals of computational efficiency. Such efficiency is relevant both to the implementation of machine vision and to a teleological understanding of the structures of biological vision.

Our goal in developing this model is to make contributions in three areas. First, this work will provide a common computational substrate for the study of machine vision and human vision. Second, our model will explain and make predictions regarding many disparate and seemingly conflicting behavioral attentional phenomena. Third, the connectionist structure will provide a link between behavioral and neurophysiological data related to attention.

The paper is organized as follows. In the next section, we review the work most closely related to the initial design of the attentional model. In the following section, we describe the model, relating its structure to this earlier work. Next, results of a number of simulations of two subnetworks of the model are presented. Finally, we discuss the model in terms of additional attentional be-
haviors that have not yet been simulated, and indicate several ways in which the current model is incomplete.

RELATED WORK
This work makes use of results from a number of different disciplines, including work in artificial intelligence on computational vision and spatial representation, work in psychology on attention and low-level perception, work in connectionist models on object recognition, and work in neurophysiology on visual areas and pathways in the primate brain. We review here some of the major points of contact between our work and these others.

COMPUTATIONAL VISION
Low-level vision is concerned with the extraction from an image of visual primitives. These primitives are then used in the recognition of objects and their relations in the scene. There are two competing views of how low-level vision works. In one view, the function of low-level vision is to produce a depth map of the scene being imaged (Marr 1982; Poggio et al. 1985). This depth map can then be used to extract 3-D features for the recognition of objects. An alternative view, which is more in agreement with our own outlook, is that carefully selected 2-D features provide sufficient information for the bottom-up component of image segmentation and object recognition (Lowe 1987). Features such as parallel or collinear lines, symmetric shapes, and smoothly changing regions in the 2-D image are invariant over wide ranges of 3-D viewpoints and are extremely unlikely to arise by accident (Witkin and Tenenbaum 1983; Pentland 1986). The use of such features would fit well into a parallel, hierarchical structure such as Uhr’s (1972, 1978) pyramid model, Fukushima’s (1980) neocognitron, or Sabbah’s (1985) connectionist origami world network.

Parallel architectures for machine vision have been built in hardware (cf. Batcher 1980) and simulated in software (cf. Tanimoto and Klinger 1980). Due to the large amount of uniform and relatively simple data comprising an intensity image, these designs invariably make use of large numbers of relatively simple processing elements. Although these highly parallel architectures fit well the requirements of low-level vision, complete parallelization of the recognition process would require resources that are not available to either the human vision system or to artificial vision systems. For example, Tsotsos (1987) argues that even the most conservative estimates on the size of the recognition problem faced by humans would require over $10^{10}$ processors for completely parallel processing. This and similar arguments have led to an emphasis on efficient representation of visual information, and combined parallel-serial models.

Among the computational characteristics that are important to efficient representation are hierarchy, locality, and viewpoint invariance. For domains in which objects can usefully be defined as combinations of simpler sub-objects or features, hierarchical representations reduce the combinatoric size of the representation by combining related features into groups that can then be shared among objects. Block provides an excellent illustration of this representational efficiency (Block et al. 1964). A local computation is one requiring access to only a limited segment of the entire representation. Locality requires fewer pathways among parallel processes and is particularly effective for vision tasks. Viewpoint invariant representations reduce the space requirements of object models by efficiently encoding object characteristics independent of viewing conditions.

A number of workers have addressed the problem of translation invariant object recognition in parallel models (Fukushima 1980; Hinton 1981; Ballard 1984). The networks used by Ballard and Hinton are such that, in order to limit the amount of processing hardware to a realistic range, only a single object can be recognized at a time. The explicit representation of image location of objects in these models, and the constraint that only a single image location can be processed at a time, provides a basis for an attentional mechanism mediated by image location.

PSYCHOLOGY OF ATTENTION
As the basis for the initial development of our model, we have chosen a small number of studies from the psychological literature. These include Treisman’s experiments supporting her feature integration theory, Wolford’s study of feature perturbations, and Duncan’s data supporting his stimulus similarity theory.

The work of Treisman is of primary importance to this work. Treisman and her colleagues (Treisman and Gelade 1980; Treisman and Schmidt 1982) have collected data on human visual performance for a variety of tasks in which attention is implicated. In particular, her data suggest that in a visual search task, when the target to be detected differs from nontarget distractors along a single primitive feature dimension, the target can be detected in an amount of time that does not depend on the number of distractors. Thus, it appears that the search for the target proceeds in parallel, with all objects, target and nontarget alike, being processed simultaneously. On the other hand, when the target is distinguished from the distractors by a combination of values in two feature dimensions, the amount of time required to detect the target increases linearly with the number of distractors in the input. Thus, it appears that the search for the target proceeds serially, with all objects being processed in sequence.

Treisman’s feature integration theory explains these data as follows. Primitive features are represented independently of one another in the visual system. Whenever the target contains a feature that is unique to the stimulus,
it can be immediately detected within the representation of that feature, and passed by the attention mechanism. Since all objects, targets and distractors, can be filtered with respect to this feature at once, this feature search is a parallel mechanism. When the target is distinguished from distractors only by a combination of two or more features, this parallel filtering is no longer possible. Instead, each object must be considered in turn to test for the combination of features that characterizes the target. This serial process is called conjunctive search, since it involves the detection of a conjunction of features to identify the target.

Duncan provides an alternative theory to explain the reaction time dependencies of visual search (Duncan and Humphreys 1989). Setting aside the parallel versus serial distinction, Duncan claims that the search times depend only on two measures: the similarity between the target and the nontarget distractors (T-NT similarity), and the similarity among the nontargets (NT-NT similarity). Higher T-NT similarity increases the search time for the target, while lower NT-NT similarity has an even more pronounced effect in increasing the search time. This is referred to as the stimulus similarity theory.

Another aspect of Treisman’s feature integration theory concerns the confounding of objects in short exposure presentations. When a multiple object image is presented tachistoscopically, the object reported is frequently one that was not in the image, but whose features were present as components of the objects that were in the image. This phenomenon is referred to as an illusory conjunction. For example, in a field of Ps and Qs, the subject might report having seen an R, since it can be formed by a conjunction of features from P and Q. Treisman’s theory ascribes such an error to a failure to focus attention, which allows the free floating features of unattended objects to be combined incorrectly.

Wolford reports data on perceptual errors in letter recognition tasks that are related to those just described (Wolford 1975; Wolford and Shum 1980). His feature perturbation model provides an alternative explanation of the incorrect combining of features that occurs when stimulus presentation times are short. In this theory, a feature corresponding to one object can migrate to an adjacent object, leading to the perception of an object having a combination of features not present in the stimulus. Wolford characterizes the manner in which these migrations depend on retinal locus, order in which objects are reported, order in which objects are scanned, and report delays. These dependencies are not addressed in Treisman’s work.

In the work described below, we develop a connectionist network to model the various phenomena reported by these authors. Our goal is to provide a more comprehensive framework from which to analyze these phenomena, and to clarify the points of agreement and points of contention of the individual theories that have been proposed.

**CONNECTIONIST MODELS**

Prior to this work, a handful of network models have been proposed to explain individual pieces of the attentional data. Hinton and Lang (1985) used a network similar to Hinton’s (1981) mapping network to simulate illusory conjunctions. Using a winner-take-all (WTA) array of processing elements to represent the attentional focus, they found that illusory conjunctions could be made to occur if a random input pattern was presented prior to settling of the WTA process in the attention array. Mozer (1988) used a similar network structure to simulate a probabilistic attentional mechanism having variable focus size in his MORSEL system. Sandon and Uhr (1988) used a network similar to Hinton’s mapping network, but implemented the representation of location hierarchically, as a means of more efficiently representing and learning translation-invariant object recognition. These ideas are all incorporated in the model described below.

**NEUROPHYSIOLOGY OF VISION**

Finally, we take note of a number of neurophysiological findings that are relevant to the work reported here. First, the two visual pathways in the primate brain (Mishkin et al. 1983) provide separate representations of object identity and spatial relations. These two major pathways are in turn composed of separate subpaths (Hubel and Livingstone 1985). Object recognition involves temporal cortex (Bruce et al. 1981) while spatial relations are represented in parietal cortex (Andersen et al. 1985; Zipser and Andersen 1988). This same structure is evident in the network models of Hinton, Ballard, and Mozer. The computational reason for separately representing features used for identification of an object and characteristics related to the spatial configuration of the object is that it provides more powerful invariant features to be used for both identification and localization.

Second, these visual pathways are organized hierarchically (Van Essen and Maunsell 1983; Van Essen 1985). The identification pathway includes primate areas V1, V2, V4, and IT, while the spatial pathway, which also involves motion processing, includes areas V1, V2, VP, MT, and parietal cortex. Higher areas in each pathway have successively larger receptive fields, and respond to successively more complex stimuli, as would be expected in a hierarchical organization.

Third, the shifter circuits that Van Essen finds plausible for stabilization of retinal images (Anderson and Van Essen 1987) have the same logarithmic connectivity that we have used to support translation invariance in the hierarchical network mentioned above (Sandon and Uhr 1988). The network model described below uses a somewhat different connectivity in a structure that would also support image stabilization. This different connectivity would presumably be equally plausible to that described...
We have designed a network model of visual attention in which stimuli within a limited, contiguous region compete for attention, while those from spatially distant locations do not interact. In IT on the other hand, the attention behavior is global. Here, all stimuli in the entire visual field interact such that "unattended stimuli yield no response. We discuss this further when presenting the simulation demonstrating hierarchical attention.

Fourth, Desimone cites evidence for two different attentional mechanisms in two different neural loci (Moran and Desimone 1985; Wise and Desimone 1988). In V4, the behavior is that of a "local" attention mechanism, in which stimuli within a limited, contiguous region compete for attention, while those from spatially distant locations do not interact. In IT on the other hand, the attention behavior is global. Here, all stimuli in the entire visual field interact such that "unattended" stimuli yield no response. We discuss this further when presenting the simulation demonstrating hierarchical attention.

Fifth, many models of attention assume a spatially contiguous region for the attended image features. Crick (1982) presents arguments for the locus of attentional activation in the reticular complex of the thalamus. This "attentional searchlight" is sometimes assumed to involve full processing of information within the selected region, and no processing without. Evidence exists, based on event-related potential measurements, for a processing gradient, which falls off with distance from the center of the attended region (Mangun and Hillyard 1988).

A CONNECTIONIST MODEL OF ATTENTION

We have designed a network model of visual attention based on constraints drawn from three fields: computational principles developed in the machine vision literature, knowledge of the neurophysiology of vision, and behavioral data. In this section, we describe the overall model, referring to the literature from which constraints were derived. In the following section, we describe a number of simulations involving pieces of the model, which demonstrate some of the behaviors that have been reported in the psychology literature.

The underlying structure of the network is based on the original network used by Hinton (1981) to model translation- (as well as rotation) invariant object recognition. Ballard's (1984) parameter networks use an equivalent structure to address the more general problem of 3-D viewpoint-invariant recognition. The idea is to represent both shape features and spatial location of objects in separate arrays, and to combine these two sources of information multiplicatively (using so-called conjunctive connections) in recognizing objects. This design provides a computational structure that allows for the representation of multiple competing hypotheses about the location and identity of objects, and produces an interpretation of the image that is most consistent with the dual set of constraints (location features and shape features).

The structure proposed by Hinton requires one conjunctive connection for every possible combination of location and shape feature. For an $N \times N$ array of location nodes, and an $M \times M$ array of translation-invariant feature locations (object-centered features) for each of $F$ shape features, the number of connections required is $N^2 \times M^2 \times F$. As the size of the location array increases, the number of connections required becomes excessive.

As stated in the previous section, use of a hierarchical structure reduces the space requirements (number of processors and connections, in this case) of a representation. The use of pyramid data structures (Tanimoto and Pavlidis 1975; Uhr 1978) and pyramid processor architectures for machine vision is due mainly to their intrinsic hierarchical structure.

In previous work (Sandon and Uhr 1988), we augmented a pyramid structured shape network with a location subnetwork to efficiently perform translation-invariant object recognition. We used a two layered hierarchical representation for the location subnetwork, which reduced the required connectivity, allowing us to train the network to recognize objects in given positions and then generalize to novel positions. The hierarchical representation of location in that network used about $2 \times N$ location nodes, rather than $N^2$, and therefore required $2 \times N \times M^2 \times F$ connections. In a vision system of even moderate size, this factor of $N^2$ fewer nodes and connections can represent a space savings of one to two orders of magnitude. In the model described here, we use a modification of this hierarchy, in which about $N^2$ nodes are used to represent location, but due to the local nature of the lower layer of the hierarchy, the number of connections is still proportional to $N$ rather than $N^2$.

One apparent drawback of this object recognition model is that only a single object can be processed at a time. To process a multiple object image, an additional computational mechanism is needed to inhibit the processing of all but one object at a time. Hinton and Lang (1985) implemented a winner-take-all (WTA) mechanism (Feldman and Ballard 1982) in the location array, to force all but the strongest response in that array to zero. Interruption of the input during the relaxation phase of this WTA network resulted in illusory conjunctions of the type reported by Treisman. Sequential attention to different objects in the image is presumed mediated by an inhibition of the activation corresponding to the object currently being processed, allowing a different region of this network to become active.

Mozer (1988) used this same idea of representing attentional locus based on spatial location in his object recognition network. Mozer added the capability to attend to different sized regions of the image, through the high level control of a parameter defining the extent of inhibition used in the WTA procedure.

The model proposed here uses a hierarchical representation of spatial location as the basis for an attentional mechanism. This structure automatically provides translation-invariant recognition of objects, since its underlying structure is that of the translation-invariant network. To this basic structure, we add a capability for multiple scale analysis. Instead of Mozer's approach, which allows...
processing at only a single scale at a time, and does not transform the image features to a normalized scale, this model uses multiple data paths and scale transformation. Like the representation of location, which allows multiple image locations to compete (through the WTA mechanism) for access to higher level processing, the network uses these multiple data paths to process images of varying resolution, allowing competition among scales for high level processing.

This multiple scale processing structure is suggested both by machine vision work in multiresolution representation and processing (Dyer et al. 1980; Ballard 1981; Crowley 1983; Witkin 1983) and by evidence for multiple channels in the human visual system for processing image data at various scales (Marr 1982). We use as input to each data path the appropriate layer from a gaussian pyramid (Burt and Adelson 1983), which is computed by repeatedly smoothing (with a gaussian weighting function) and reducing the intensity image.

Given the multiscale input provided by the gaussian pyramid, what features should be extracted from the image? Both neurophysiological evidence (Hubel and Wiesel 1962; Jones and Palmer 1987) and psychological evidence (Attneave 1954) suggest the use of oriented edges as the lowest level features. Although features such as motion, color, and depth are also presumed to be primitive visual features, they are not applicable to the current model, which considers static, monochromatic, planar images.

For the purpose of recognizing relatively simple geometric shapes, as are generally used as stimuli for psychological and neurophysiological studies, a shape feature hierarchy such as that used by Uhr (1978) or Sabbah (1985) is appropriate. Since the evidence for various shape feature detectors in the human visual system is still a matter of debate (Treisman and Gelade 1980; Sagi 1988; McLeod et al. 1988; Duncan and Humphreys 1989), this model makes no a priori commitment to a particular set of features. On the other hand, as we discuss in the conclusion, our work in modeling various behavioral characteristics will lead us to make very specific choices as to the appropriate shape features to be used.

Next, we address the question of how attention is guided. That is, what features of the image are used to activate the attentional layer? Note that the features used for recognition need not be the same features used to guide attention. There are two aspects to be considered in addressing this question. First, what are the attentional features themselves, and second, how do these features interact to produce the attentional activation?

Regarding the features themselves, there are two lines of evidence for choosing attentional shape features. First, behavioral studies produce attentional effects using short lines of varying orientations and simple combinations of these (cf. Sagi and Julesz 1985). This suggests that oriented edges or lines provide attentional input. Evidence for more complex features, such as corners, crossings, and line terminations is less conclusive (cf. Treisman and Gelade 1980 versus Duncan and Humphreys 1989 on corners, or Bergen and Julesz 1983 versus Iwama and Maida 1989 on crossings). Again, the model does not commit to a particular feature, though we address the question of corner features in one of the simulations described below. Second, evidence for the ability to spontaneously form perceptual groupings of image tokens suggests that the shape features most effective in guiding attention in complex images are groupings exhibiting symmetry, collinear or parallel lines, or adjacent line terminations (Lowe 1987; Witkin and Tenenbaum 1983).

The proposed model uses this first type of shape feature for guiding the low level attention mechanism, and the second type of feature to guide the high level attention mechanism. We do not intend to rule out additional levels in the attentional hierarchy, but for simplicity and lack of evidence to the contrary, we consider only two levels in the current model.

The second consideration for the guidance of attention has to do with how the input features interact to yield an attentional activation. In earlier models (Hinton and Lang 1985; Mozer 1988), the input to the attention layer was the sum of activation of all primitive feature detectors at each location. The WTA mechanism within the attention layer then inhibits all but the strongest activation within the array. This combining ignores the evidence that like features interact more strongly with each other than with unlike features (Treisman and Gelade 1980). Since this behavior is central to the attentional phenomena of interest here, we implement an interaction among like features prior to their introduction to the attention array. In particular, a central-excitatory, peripheral-inhibitory interaction among features of a given type is applied to each of the feature arrays used as input to the attention array. This contrast enhancement of features produces input to the attention array only when a given feature occurs in the image in relative isolation from other features of the same type.

The effect of the attentional activity is to gate the features from a particular region of the image up to higher layers of the network, where object recognition occurs. As previously noted, the features comprising the input to this recognition process are location invariant. Similarly, these features are made scale invariant by transforming each possible scale to a normalized size. The individual data paths representing the different processing scales are combined prior to recognition processing. The criterion used to choose which data path is gated to the recognition process is based on a computational principle that is used in machine intelligence applications (cf. Levine 1980). This principle is evident in behavioral and neurophysiological data as well (Hughes et al. 1990), though these data apply to a lower level mechanism than the one we describe. The principle is called
coarse-to-fine processing. It is a search strategy that makes use of coarse level (more global) information to find an approximate solution, and then refines this solution using more detailed information. In psychology, this principle is referred to as global precedence, referring to the use of global information prior to the use of more fine-grained information. The criterion used in the model, then, is that the coarsest data path that exhibits significant activity in its highest attention layer is gated to the recognition processor.

Figure 1 summarizes the attentional model described in this section. The leftmost data path is for fine scale processing and includes two levels of attention. Data paths are bidirectional (McClelland and Rumelhart 1981), providing a pathway for attentional priming, and other task-directed responses. The middle data path is similar, but starts with a lower resolution intensity image, and requires only one attention layer to select features for processing by the recognition processor. The rightmost data path involves the coarsest resolution intensity image, whose features are passed directly to the recognition processor. These three data paths provide processing at three scales. The choice of which scale to process is made by the scale arbitrator, which implements the global precedence criterion.

SIMULATION RESULTS
We have run simulations of two different pieces of the network just described. The first subnetwork is used to demonstrate a hierarchical attention mechanism, and then to simulate Wolford's feature perturbation effect. The second subnetwork is used to demonstrate perceptual popout and to simulate some of Duncan's stimulus similarity effects.

HIERARCHICAL ATTENTION
In our work on translation-invariant object recognition (Sandon and Uhr 1988), we used a hierarchical representation of object location as a means of efficiently implementing the mapping units suggested by Hinton.
As others have observed (Hinton and Lang 1985; Mozer 1988), the representation of object location can be used to mediate an attentional spotlight. This suggests that the attentional focus, like object location in our previous work, can be efficiently represented in a hierarchy (Sandon 1989). We have simulated a small network that implements a two-layer attention mechanism at a single scale, which chooses a particular region of an input image for recognition processing.

A schematic drawing of this network is shown in Figure 2. The 64 × 64 pixel intensity image used in this simulation is shown in Figure 3a. The image is processed as follows. First, eight oriented edge images are extracted. Figure 3b shows a composite of two oriented edge images, where light regions indicate diagonal edges changing from dark to light when moving northwest (up and to the left) while dark regions indicate diagonal edges of the opposite polarity (dark to light when moving down and to the right). The set of eight edges provides the input to the first level attention layer. Each edge image is decomposed into a set of 16 × 16 subarrays, such that adjacent arrays have 50% overlap. This operation yields a 7 × 7 matrix of subarrays. Each subarray corresponds to a "local" attention region. Associated with each attention region is a 9 × 9 attention array. The receptive field of each of the attention nodes is an 8 × 8 region of each of the edge images. Each node computes the total edge strength in its receptive field. The set of attention nodes in each of the 9 × 9 arrays comprises a WTA network, which inhibits all but the most strongly activated node. The winning node in each local attention array indicates the 8 × 8 area of the corresponding local attention region that exhibits the most edge activity.

Figure 4a shows the total edge input, simply the sum over all eight edge orientations, to each of the attention regions. Figure 4b shows each of the attention arrays, with the maximally firing node highlighted. The winner of the WTA competition for each attention region gates the corresponding 8 × 8 area in each edge image up to the next level for further processing. Figure 4c shows the total edge input corresponding to the winning subregions.

Given these subregions, which are deemed "interesting" due to their total edgeness, the second attentional step chooses just one of these subregions to be processed for recognition. In a hierarchical representation, we work with increasingly complex features at each successive stage of processing. At the same time, at any given level, we work with features that are relatively easy to compute from the features at the next lower level. For this simulation, we have chosen pairs of parallel but opposite polarity edges to mediate the focusing of the second attention process. We refer to these features as opposing edge pairs. In addition to such features as collinear edges and termination proximity, parallel edges have been suggested by Lowe (1987) as useful features for low-level segmentation.

Figure 5a shows the edges from Figure 3a that were...
attended in each of the local attention regions of the first attention level. Four sets of opposing edge pair features are computed, one for each of the horizontal, vertical, and two diagonal orientations. The results of this computation for the edges of Figure 5a are shown in Figure 5b. Finally, the total opposing edge pair strength is used as the input to the second level attention array, which contains one node for each of the \(7 \times 7\) first layer attention regions. This array is shown in Figure 5c, though the maximum value, in the bottom row, third from the right, is barely discernable from the figure. To see what part of the input image this corresponds to, refer to Figure 4c, bottom row, third subregion from the right. The features we have used to guide attention at the two levels apparently tend to choose the biggest blob that fits within the attention region. We did not implement the recognition network, shown in dotted lines in Figure 2.

Our purpose in simulating this network was to demonstrate the general structure that a two level attention mechanism might have. Notice that this network is useful for performing translation-invariant object recognition, but for objects that are of a certain size. The problem of scale invariance is treated separately, in the following subnetwork. The goal of the overall work is to develop a network model that can provide a bridge between human vision and machine vision, between psychological data and neuroscientific data, and between disparate behavioral theories of human attention. There are three points of contact between this simulation and other work.

First, Desimone has reported cells in V4 of monkey that respond as if they were implementing a local attentional mechanism (Moran and Desimone 1985; Wise and Desimone 1988). These cells fire when an attended stimulus is present in the receptive field (RF), do not fire if an unattended stimulus is present in the RF when attention is directed to that RF, and fire when any stimulus is present in the RF but attention is outside the RF. These are the characteristics of our first level attention mechanism, which involves Competition for attention by stimuli that share an attentional region, but no competition for stimuli in different regions. Note that although in our model there are many “attended” regions at the lower level, it is the single region chosen at the second layer that corresponds to the behavioral attention region referred to in Desimone’s studies. Desimone also reports that no attentional behavior is found in cells in V1, while the attentional behavior of cells in IT reflects a global attention mechanism, like our second level of attention.

Second, part of the debate in the psychology literature regarding early versus late selection mechanisms involves the fact that unattended stimuli are partially processed. Some models of attention are probabilistic (cf. Mozer 1988) to allow unattended features to be pro-
cessed occasionally, in addition to processing of attended features. In our hierarchical model, many regions of the image get processed partially because they are passed beyond the first attention level. We will need to model the recognition and response functions before we can make predictions about what unattended input gets processed to what extent. This further modeling will also be required before we try to reconcile the evidence for a spotlight and that for an attentional gradient (Sagi and Julesz 1986; Mangun and Hillyard 1988).

The third point of contact between this simulation and other work is addressed in the next simulation.

**FEATURE PERTURBATIONS**

We noticed in the previous simulation that the segmentation occurring at the first level of attention was fairly good, in that the subregion chosen for further processing generally corresponded to an object in the image. However, referring to Figure 4a once again, in the first column, the fourth subregion down, we see a segment that does not correspond to an object. The attention mechanism for this region has chosen a segment consisting of parts of two objects. If we view the attention mechanism as a segmentation process, we would label this as an incorrect segmentation. Given the weak information used to guide this segmentation process, it is not surprising that such incorrect segments would sometimes result.

In the next simulation, which uses the same network as described above, a behavior similar to the feature perturbation phenomenon described by Wolford is exhibited (Wolford 1975; Wolford and Shum 1980). In character recognition tasks involving speeded responses, the incorrect characters reported by subjects often contain combinations of features of two different characters in the stimulus. Treisman refers to this as an illusory conjunction, and explains it as the correct segmentation of two characters, followed by attentional gating of both sets of features to the recognition process. Wolford refers to a similar behavior as a feature perturbation, and explains it as a migration of a feature from one character to another.

Using an input stimulus similar to that used by Wolford (see Fig. 6a), we studied the segmentation obtained in the first attention layer, which is again guided by oriented edge features. Figures 6 and 7 are the analogs of Figures 4 and 5 using this new input image. Figure 7c shows the segmentation that was obtained. The segment in the fourth row from the top, third column from the left is similar to the incorrect segment found in the previous simulation, as it too contains parts of two different objects.

In the absence of a complete simulation model, we can draw only tentative conclusions. The interpretation of this result is the following. During the relaxation process associated with the WTA mechanism at each of the attentional levels, multiple segments are partially gated to the recognition process as part of the search for the best region of interest. During this time, the incorrectly segmented region will be competing for attention at the second layer with other segments found at the first layer. In a speeded response task, a response would be required before the second level attention process settles on a single segment. The incorrect segment provides evidence for a box with a tick on the right, and from among the competing hypotheses, this object would occasionally be reported.

In Wolford's experiments, the set of characters that subjects expected to see were boxes having either no ticks, or a single tick on one of the four sides. Given the stimulus shown in Figure 6a, subjects would occasionally report the second character as a box with a tick on the right. This is exactly the incorrect perception that this partial model predicts.

Wolford's feature perturbation model explains this phenomenon as being due to a migration of the tick from one object to another. Our simulation suggests that this phenomenon is due not to migration of features, but to incorrect segmentation at this low level. Work is currently in progress (in collaboration with Wolford) to rerun the original experiments using a stimulus such as that shown in Figure 8a. Wolford's model predicts that subjects will report seeing the object shown in Figure 8b in the second position from the left, while our model predicts that subjects will report seeing the object shown in Figure 8c in that position.

**POPOUT**

We simulated the second network, schematically shown in Figure 9, to begin our study of some of the phenomena reported by Treisman and Duncan. This network implements a single layer attentional mechanism, but at four different scales. We discuss the processing at the finest resolution scale first.

Like the previous network, the input image is an intensity array of 64 × 64 pixels. The network first extracts simple features from the image, in this case, four oriented line features. Next, an on-center, off-surround gaussian mask is applied to these four oriented line images to get four feature contrast images. The feature contrast images are summed together in the attention layer to get the initial attentional activation. Finally, a WTA competition is executed within the attentional array to select a single active region.

The image is processed at three coarser resolutions by first computing reduced size intensity images. A 32 × 32 pixel intensity image is computed by applying a 3 × 3 gaussian filter to the original image, then eliminating every other row and every other column in the image. The 16 × 16 and 8 × 8 pixel intensity images are produced by applying this operation recursively. Figure 10 shows the four intensity images. These coarse resolution...
intensity images are processed similarly to the full resolution image, producing attention arrays of correspondingly smaller sizes.

Since we have not implemented the entire model, we make two assumptions regarding the results of these network simulations. First, the results of computation at the different scales would be combined by a scale arbitrator implementing a policy of global precedence. This policy would simply give preference to activity at the coarser, more global scale prior to that at a finer scale. Second, in cases where the target object is not the first one attended to, we assume the existence of a higher level inhibitory control mechanism, which would inhibit the currently active attentional region, allowing the region with the next strongest support to become active (Mozer 1988; Klein 1988).

The first phenomenon we studied in this network is parallel, or feature, search. This popout phenomenon occurs when a target object is distinguished from the distractor objects by a single feature. Popout is characterized by a response time to the stimulus that is independent of the number of distractor objects in the input.

Figure 10, on the left, shows the 64 x 64 pixel input image used in this simulation. The distractor objects are lines oriented like the middle stroke of the letter "N," and the target object is a line oriented like the middle
stroke of the letter "Z." Human subjects detect the Z-line in this image in a time that is independent of the number of N-line distractors. According to Treisman's feature integration theory, popout occurs in this case because there is a feature associated with the target that is unique among all objects. The purpose of this simulation is to demonstrate a computational mechanism that implements this parallel search behavior.

The two diagonal line features computed from the intensity images in Figure 10 are shown in Figure 11a and b. The corresponding feature contrast images are presented in Figure 12a and b. The sum of all feature contrast images, that is, the input to the attention arrays, is presented in Figure 13. We did not execute the WTA procedure for this simulation, since the attentional activity can be easily interpreted from its initial state.

We interpret the results shown in Figure 13 as follows. At the two most coarse resolutions, the feature detectors responded weakly so that no activity was registered in the attention array. This is simply an indication that there is no low spatial frequency (coarse resolution) pattern to be detected in this image. At the finest resolution, the attention array has many regions of activity. This occurs because the individual objects are spaced far enough apart that they do not appear in the inhibition fields of adjacent objects. Finally, at the second finest resolution, the attention array has a single region of activation. As can be seen by comparing Figure 13 with Figures 11 and 12, this occurs because the N-lines at this scale are dense enough to inhibit one another, while the single Z-line activates its corresponding attention region. Invoking our global precedence policy, we would expect the single

---

Figure 9. The subnetwork used to demonstrate popout (simulation #3), and to simulate stimulus similarity phenomena (simulation #4).
activation at this layer to be attended to prior to the
regions activated at the higher resolution. Since the
single activation in the attentional layer, corresponding
to the target object, is unaffected by the number of distrac-
tor objects, this configuration yields “immediate” atten-
tion to the target independent of the number of
distractors. This is the popout phenomenon.

STIMULUS SIMILARITY

Duncan suggests that the time it takes to detect a target
among distractors is a function of both the similarity
between target and distractors and the similarity among
distractors (Duncan and Humphreys 1989). Bergen and
Julesz (1983) made a similar observation about the sim-
ilarity between a target and its distractors. This is in
contrast to Treisman’s feature integration theory, which
identifies the determinants of detection time as the pres-
ence or absence of primitive features in targets and dis-
tractors. We have used the network shown in Figure 9
to simulate several of Duncan’s experiments, using the
same set of stimuli used in those experiments.

In one experiment, a letter “L” is presented among
upright and 90° CW rotations of the letter “T.” A sample
display of this type is shown in Figure 14, along with the
reduced resolution intensity images used in the network.
Duncan finds that the time required to find the target
“L” in this case is nearly independent of the number of
nontargets used. We have added oriented L-junctions as
features in this simulation. The feature planes for four of
these features (two oriented lines and two oriented L-
junctions) are shown in Figure 15a and b.

The corresponding contrast enhanced features are
shown in Figure 16a and b. Finally, the initial values of
the attention arrays are shown in Figure 17.

We have implemented a WTA relaxation procedure for
the attention layers, based on the following inhibition
rule (Koch and Ullman 1985):

\[
\frac{dy_i}{dt} = y_i(x_i - \sum y_j)
\]

In this rule, the \(x_i\) are the attention layer activations,
while the \(y_i\) are the activations of a set of auxiliary nodes.
The rule constrains the \(y_i\) always to sum to 1. The \(y_i\) are
initialized using

\[
y(0) = \frac{1}{N} + \eta
\]

where \(\eta\) is zero mean noise. The maximum attention
node is marked by running this relaxation procedure
until a single \(y\) node is 1, and all others are zero. The
 corresponding attention layer location is the winner of
the WTA competition. The noise term is added to the
initialization values to break symmetries that can oth-
erwise cause the procedure not to converge. We have used
a termination criterion that requires one of the \(y_i\) to
exceed a value of 0.5, which is sufficient to guarantee
that the corresponding region will win the competition.

For the initial activation pattern in the highest reso-
lution layer of Figure 17, the number of relaxation cycles

**Figure 10.** The intensity image used as input to the network for
simulation #3, along with the three coarser resolution intensity
arrays.

**Figure 11.** The diagonal line features detected at each resolution level for the image of Figure 10. (a) The 'N' type diagonal lines. (b) The 'Z'
type diagonal lines.
required for the target location to be selected is 17. When only four stimulus objects are presented, criterion again was reached in 17 cycles.

Duncan describes four extreme cases with respect to his similarity measures:

<table>
<thead>
<tr>
<th>Case</th>
<th>T–NT Similarity</th>
<th>NT–NT Similarity</th>
<th>Dependence on Display Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>B</td>
<td>High</td>
<td>High</td>
<td>Intermediate</td>
</tr>
<tr>
<td>C</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>D</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

In the simulation just described, the target object, "L," is dissimilar to the non-target "T"s, and the two orientations of nontargets are dissimilar to one another as well. Thus, this first stimulus pattern is one used by Duncan to demonstrate his case C.

Figure 18a shows a stimulus pattern corresponding to Duncan's case A. The initial attention activation for this pattern is shown in Figure 18b. The result of applying the WTA algorithm is again 17 cycles for both the four and six object images. Both of these first two cases correspond to parallel search for the target.

Figure 19a shows a stimulus pattern corresponding to Duncan's case B. In this case, the nontarget objects contain an "L" corner of the same orientation as the target object, yielding a high T–NT similarity. The resulting initial attention activation is shown in Figure 19b. The result of applying the WTA algorithm is 50 cycles for this stimulus, but 33 cycles for four objects. This condition leads to a significant dependence of the search time on the number of objects in the image, in contrast to the previous two cases.

Finally, Figure 20a shows a stimulus pattern corresponding to Duncan's case D. The resulting initial attention activation is shown in Figure 20b. For this case, the WTA relaxation time is 34 cycles, and 18 cycles for four objects. However, this case is qualitatively different from the others in that the attention region chosen is not the one corresponding to the target. Instead, the network converges to an image region corresponding to a non-target object. To find the target in this case, multiple
Figure 15. Some features detected at each resolution level for the image of Figure 14. (a) Horizontal and vertical line features. (b) Oriented L-junction features.

Figure 16. Contrast features corresponding to the features of Figure 15. (a) Horizontal and vertical line features. (b) Oriented L-junction features.
Figure 17. The initial activation of the attention arrays at each scale for the stimulus of Figure 14.

settlings of the WTA network would be required, each followed by an inhibition of the incorrect attention region, until the target is found (see discussion below). This case yields a search time dependence on the number of image objects that is stronger than any of the previous cases. Our simulation results are thus in agreement with Duncan’s experimental data.

DISCUSSION

The two previous sections describe the overall network model of attention and some simulation results associated with particular subnetworks. In this section, we summarize the theory of attention that is implicit in the model, as it relates to the previously described work of Treisman, Duncan, and Wolford. In addition, we discuss additional characteristics of the model, both in its current form and with extensions.

In our model, detection of a target in a time that is independent of the number of distractors occurs whenever the stimulus pattern produces an initial attention pattern having a single strong region of activation, with perhaps some additional regions having much weaker activation. In this case, the WTA computation in the attention array quickly converges on the strongly active region as the attended region. This is the case that Treisman refers to as parallel search. In the case in which the target differs from the distractors in a single primitive feature (one explicitly represented in a feature array in our model), a single strong attention region results, and so our model agrees with Treisman’s feature integration theory in this case. This situation is characterized by Duncan as one in which target–nontarget similarity is low (both cases A and C), and his stimulus similarity theory also predicts target detection in time independent of the number of distractors.

Whenever multiple regions of comparable activation occur in the attention array, the WTA computation converges more slowly. More importantly, when the activations due to targets and nontargets are similar, the WTA procedure is as likely to converge on a nontarget region as a target region. In this case, multiple WTA cycles are required to eventually detect the target. This is the case that Treisman refers to as serial search. When the target is defined as a conjunction of features shared with distractors, no single strong attention input results, and so we might expect a serial search in our model, as Treisman predicts. Duncan’s case D would be expected to exhibit this behavior in our model, since the case of heterogeneous nontargets yields less inhibitory interactions among them, due to their representation in different feature arrays. This lower inhibition leads to greater attentional activity associated with the nontargets.

There is a third case which is addressed by Duncan but not by Treisman. In our model, when multiple regions exhibit significant attention activity, but the initial activation corresponding to the target is strongest, the array will converge to the target region. However, due to the other activity in the array, the convergence time will be slower than in the parallel search case, and will depend on the number of other active regions. Duncan’s

Figure 18. (a) The second intensity image used as input to the network for simulation #4, along with the three coarser resolution intensity arrays. This image pattern corresponds to Duncan’s case A. (b) The initial activation of the attention arrays at each scale for the stimulus of (a).
Figure 19. (a) The third intensity image used as input to the network for simulation #4, along with the three coarser resolution intensity arrays. This image pattern corresponds to Duncan's case B. (b) The initial activation of the attention arrays at each scale for the stimulus of (a).

Figure 20. (a) The fourth intensity image used as input to the network for simulation #4, along with the three coarser resolution intensity arrays. This image pattern corresponds to Duncan's case D. (b) The initial activation of the attention arrays at each scale for the stimulus of (a).

Case B corresponds to this situation, and his case D could also yield this situation rather than the previous one. In this case, we can get detection times that depend on the number of distractors, but the dependence is less than in the "serial" search case just described. The exact time dependence is a function of the relative activations due to the target and the distractors—the smaller the difference, the larger the dependence. Thus, our model clarifies what the term "similarity" means in Duncan's theory. Two stimuli are similar to the extent that they produce the same level of activity in the same primitive feature arrays. This appears not to be the definition assumed by Duncan, but is the definition required for our model to be compatible with his theory.

Feature perturbations, according to our model, result from incorrect segmentation of the image at the first attentional level. Under normal conditions, an incorrect segment may have good bottom-up support, and so will begin to win the WTA competition, but will have poor top-down support, and so will fade from consideration. When processing is not allowed to complete normally, as in the speeded response conditions under which feature perturbations and illusory conjunctions occur, the initial strength of the incorrect segment may be sufficient to gate it to the recognition process. The best match with a known object is then chosen as the corresponding percept. The illusory conjunctions described by Treisman occur due to a different mechanism in our model. In this case, two correct segments are competing for attention at the higher layer. As the WTA competition proceeds, both are partially active, and both are therefore weakly gated to the next level. The result is that features from each segment are composed during the settling of the attention layer. Under normal conditions, the network will settle on one segment or the other. Under speeded response conditions, recognition occurs based on the composition of the two segments.

In addition to satisfying these primary behavioral constraints, the network model exhibits other desirable characteristics. Translation-invariant recognition is achieved through the gating of segment features up the attentional hierarchy. Size-invariant recognition is achieved by pro-
cessing the image at multiple scales. Attentional priming is mediated by feedback connections in the network. For example, priming for target location by presenting a cue in the visual field prior to stimulus presentation can be modeled by an activation of that attentional region that persists until the stimulus presentation. Such prior activity of the attention layer would bias selection in favor of that location. Priming for object identity, on the other hand, would be modeled as an activation of the object representation in the recognition network, which would preactivate the attention mechanism through feedback connections to the features of the object.

Given this framework for modeling attentional behavior, there are a number of additional issues that can be addressed by modifying the network. One topic of current debate in the attention literature is the identity of the primitive features. Treisman names color, orientation, spatial frequency, brightness, and direction of motion as separable features (Treisman and Gelade 1980). However, Sagi (1988) presents evidence that orientation and spatial frequency are not separable. Similarly, evidence against the separability of motion and form has been reported (McLeod et al. 1988). Julesz identifies color, elongated blobs and blob terminators as primitive features (Julesz 1981; Bergen and Julesz 1983). In our model, primitive features are represented in feature arrays within which lateral inhibition enhances feature contrast. As we discussed when describing our preliminary simulation of Duncan's experiments, we must choose the correct primitive features to obtain the correct behavior. To the extent that we can find a set of primitive features for our model that is consistent with the behavioral data, we will be making a very explicit statement about what features are, and what features are not primitive.

Related to this question of primitive features is the question of how attention is guided. In our preliminary work, we used the sum of all oriented edges to guide the lower level attention mechanism, and the sum of all opposing edge pairs to guide the higher level mechanism. In another simulation, we introduced contrast enhancement of features prior to use by the attention layer. There have been various suggestions in the literature as to which perceptual grouping operators are most useful at the lower levels of visual processing (Lowe 1987; Marr 1982; Julesz 1981). We interpret these as being the features guiding attention, and thus producing image segments, at the low or intermediate levels.

In the current model, we use a simple object recognition network, which provides the response of the overall vision system. This provides only a crude model of response to visual inputs, which will be insufficient for studying certain aspects of the overall process. For example, our explanation of the partial processing of unattended stimuli is that there are many stimuli in the image that are attended to at lower levels, even though they may not reach the recognition processor to affect the primary response. However, later responses have been shown to depend on this partial processing. We must model the persistence of activity in the network, and more carefully model the recognition process itself, to emulate this behavior. Similarly, to more closely model the serial search process, we must implement the control mechanism that inhibits the attentional activation when a shift of attention is desired (Klein 1988; Tsal and Lavie 1988).

In addition to modeling these behaviors, we can use other psychological and neurophysiological data to make some more quantitative predictions. For example, Duncan's and Wolford's data on the size and spatial distance of interacting objects, coupled with Desimone's data on the receptive field sizes in V4 can be used to predict the structural characteristics of the inhibitory network that implements the contrast enhancement of features prior to the first attention layer. We hope also to characterize the time course of WTA relaxation in the attention regions under varying conditions when both bottom-up and top-down influences are active. By characterizing the way in which attentional activity is generated and resolved, we can make a more precise statement about the similarity metric required to make Duncan's theory work.

CONCLUSIONS

We have begun the development of a neural network model of visual attention. The overall goal of this work is to provide an understanding of the computational trade-offs involved in implementing a parallel-serial model of object recognition. Our approach is to use a neural network to model human attentional behavior, as a means of revealing some of the computational techniques used in this highly successful vision system. We have described network simulations of a number of attentional phenomena, and have provided a network design that supports a number of additional desirable characteristics.

Acknowledgments

I would like to thank Berrin Yanikoglu for helpful discussions and for assistance in developing the network simulations. I would also like to acknowledge the suggestions from two reviewers that led to improvements in the presentation of this material.

REFERENCES


