Controlling the Focus of Visual Selective Attention

Ernst Niebur
Laurent Itti
Christof Koch

ABSTRACT
Selecting only a subset of the available sensory information before further detailed processing is crucial for efficient perception. In the visual modality, this selection is frequently implemented by suppressing information outside a spatially circumscribed region of the visual field, the so-called "focus of attention." The model for the control of the focus of attention in primates presented here is based on a "Saliency Map" which is a topographic representation of the instantaneous saliency of the visual scene.

1 Introduction

Access to information about an organism's environment is essential for its survival. As a consequence, highly sensitive and efficient sensory organs were developed during evolution. Nearly as essential as acquiring information is sorting through it and deciding which part of it requires more detailed analysis and which is irrelevant. For animals with a highly developed sensorium (which includes all higher phyla), it is impractical to process all sensory input at all times. The problem is the mismatch between the need for sensors which provide information about the outside world required *at some time*, and the information processing capacity of the brain which is not capable to treat the information provided by all sensors *simultaneously*. The solution adopted by biology is to provide animals with a multitude of powerful sensors capable of providing detailed information in different modalities, to *select* a small portion of the available information, and to discard all the other information. This process is usually called "selective attention."

The triage of the incoming sensory information into parts to be discarded and parts to be retained can go along different lines, i.e. along different feature dimensions. Selective filters can be employed which selectively enhance one or more distinct visual "channels" across the visual field, such as to facilitate perception of all stimuli of a specific color, spatial frequency, or direction of motion. One of the most important of these filters uses spa-

tial location as the selection criterion, by suppressing all stimuli outside a spatially circumscribed region of the visual field relative to stimuli inside this region. Functionally, this makes sense because the location of stimuli is only weakly correlated with their other properties. In other words, most objects can appear with nearly equal probabilities anywhere in the visual field. This allows the system to reduce the complexity of the object recognition problem significantly since it can then operate in a space constructed as the *sum* of the feature space and the locality space, rather than the outer *product* of these states. Indeed, there is psychophysical evidence that space does play a special role among features [Nis85, TL93, SS93]; but see [Bun91] for a different view. The most direct evidence for a spatially defined focus of attention was recently found by imaging methods [BD99]. We refer the reader to ref. [NK98] for a more general discussion of selective attention from a computational point of view.

A space-based mechanism is also supported by the observed anatomical and physiological structure of the primate visual system. This system uses anatomically distinct pathways for encoding spatial information of objects in the environment and the specific features of these objects. The locations of visual stimuli are represented in the "dorsal" (or "where") pathway, while detailed feature processing and object recognition are localized in the "ventral" (or "what") pathway. In earlier work, we have presented neuronal models for attentional control in the ventral pathway [NKR93, NK94]. Here, we focus on the selection process in the dorsal pathway.

The purpose of this chapter is to present a neuronally based model of visual selection, paying attention to the underlying neurophysiology and anatomy. Two more technically oriented reports about aspects of this work have been published previously [NK96, INK98]. The following section 2 describes the model we have developed and implemented. Section 3 presents the results obtained on different classes of stimuli, ranging from simple synthetic stimuli akin to those used in prototypical psychophysical experiments to images of natural scenes. Finally, section 4 discusses the results, establishes the relation to previous work, and concludes with an outlook to future developments. A C++ installation of the model and numerous examples of predictions of the model can be retrieved from *http://www.klab.caltech.edu/~itti/attention*.

2 A Computational Model of The Dorsal Pathway

2.1 Model Assumptions

Our model is limited to the bottom-up control of attention, i.e. to the control of selective attention by the properties of the visual stimulus. The present model is based on the following hypotheses:

- Selection is based on iconic (appearance-based) scene representations rather than on categorical decisions.

- Visual input is represented in subcortical and early cortical structures in feature maps. A crucial step in the construction of these representations consists of center-surround computations in every feature, to avoid excessive redundancy in the processing.

- Information from these feature maps is combined in an additional feature map which represents the local saliency.

- *Per definitionem,* the instantaneous maximum of this saliency map is the most conspicuous location at a given time. The focus of attention has to be pointed to this location.

- The saliency map is endowed with internal dynamics which allow the perceptive system to scan the visual input such that its different parts are visited by the focus of attention in the order of decreasing saliency.

2.2 General architecture

Figure 1 shows an overview of the model *Where* pathway and selective attention mechanism. Input is provided in the form of digitized images from a variety of sensors, such as NTSC cameras or image scanners. Low-level vision features (red, green, blue and yellow color channels, orientation, and brightness) are extracted from the original color image, at several spatial scales, using linear filtering. The different spatial scales are created using Gaussian pyramids [AAB+84], which consist of progressively low-pass filtering and subsampling the input image. Pyramids have a depth of 9 scales, providing horizontal and vertical image reduction factors ranging from 1:1 (level 0; the original input image) to 1:256 (level 8) in consecutive powers of two. Each feature is computed in a center-surround structure akin to visual receptive fields. Center-surround operations are implemented as differences between a fine and a coarse scale for a given feature: the center of the receptive field corresponds to the value of a pixel at level $n \in \{2, 3, 4\}$ in the pyramid, and the surround to the corresponding pixel at level $n + \delta$, with $\delta \in \{3, 4\}$. We hence compute six feature maps for each type of feature. The feature maps are then normalized so as to allow the direct comparison between *a priori* not comparable modalities, as well as to enhance the feature maps in which a small number of highly significant stimuli are found. Within each modality, the features are linearly combined across scales, yielding three *conspicuity maps* for color, intensity and orientation. These are normalized, linearly combined and fed to a network of integrate-and-fire neurons representing the *saliency map.* The saliency map input, as well

as the three conspicuity maps, is a single image lying at level 4 (reduction factor 1:16 horizontally and vertically).

Short descriptions of the different feature maps are presented in the next section (2.3). We then (section 2.4) address the question of the integration of the input in the saliency map, a topographically organized map which codes for the instantaneous conspicuity of the different parts of the visual field, and present the dynamical mechanism controlling the focus of attention.

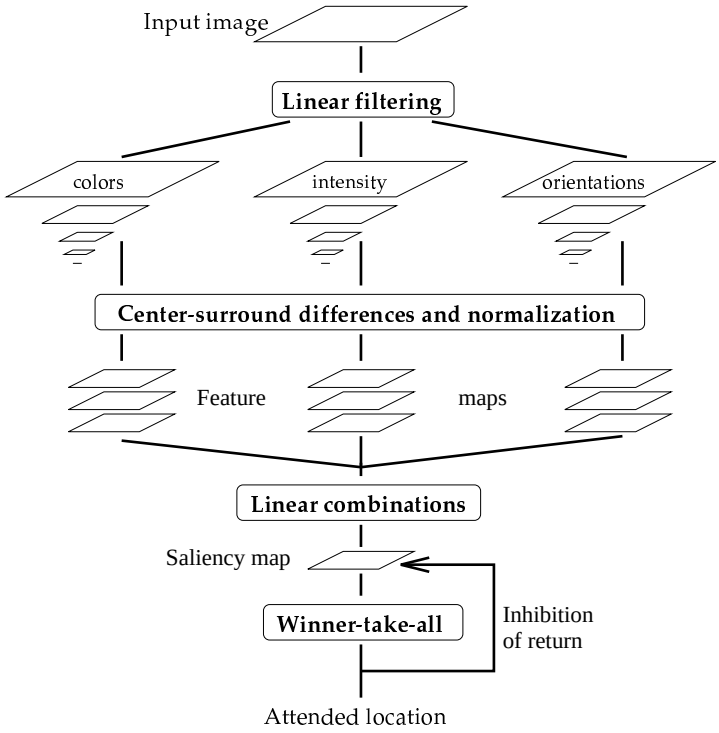

FIGURE 1. Overview of the model *Where* pathway. Features are computed as center-surround differences between several fine and coarse scales. The feature maps are normalized, combined and integrated in the saliency map which provides input to an array of integrate-and-fire neurons with global inhibition. This array has the functionality of a winner-take-all network and provides the output to the ventral pathway as well as feedback to the saliency map through an inhibition of return mechanism.

2.3 Input Features

Input is represented in a hierarchical scheme based on basis functions which are similar to those realized in the primate visual system. For the represen-

tation of shape in our model in the form of oriented edges, it has been shown that similar functions can be learned from natural scenes [OF96, BS99]. For the chromatic information, it was shown by principal components analysis that the most efficient transformation of incoming data, either from a spectrally flat Gaussian noise process [BG83] or from natural scenes [Moo85] is a three-stage mechanism. The three principal components are, in the order of decreasing signal energy, (1) intensity, (2) red minus green, and (3) blue minus yellow. These features will be discussed in the following two sections.

Intensity

Intensity information is obtained from the chromatic information of the original color input image. With R, G, and B being the red, green and blue channels, respectively, the intensity I is obtained as $I = (R + G + B)/3$. A more accurate computation could be implemented if only images from a unique known source were to be used (e.g. always the same camera). A Gaussian pyramid is created from the intensity image. The entries in the intensity feature maps are given by the modulus of the contrast, i.e., $|I_{center} - I_{surround}|$. This corresponds roughly to the sum of two single-opponent cells of opposite phase, i.e. bright-center — dark-surround and *vice-versa*.

Chromatic Input

Psychophysical results [LN93] show that color information is available for preattentive selection. We therefore implemented inputs to the saliency map which depend on the chromatic contrast of the input image. The red, green and blue components of the original color image are first normalized by the intensity image computed in the previous step, in order to clearly decorrelate intensity and hue information. Three gaussian pyramids are then created for these three isoluminant color components. At any given scale, yellow is computed as $(R + G)/2$. A quantity corresponding to the double-opponency cells in primary visual cortex is finally computed by center-surround differences across scales. For instance, for the red-green filter, we first compute at each pixel the value of (red-green) at the scale of the center. From this, we then subtract (green-red) of the surround. Finally, we take the absolute value of the result.

Shape: orientated edges

Local orientation is obtained at all scales through the creation of oriented Gabor pyramids from the intensity image. In such pyramids, orientation-selective Gabor filters of increasing spread are successively applied to down-scaled versions of the original intensity image. The oriented Gabor pyramids were implemented in a computationally efficient way by using overcomplete steerable pyramids [GBG+94]. Four different orientations are used (0, 45,

90 and 135 degrees). Center-surround operations are performed, within each oriented pyramid, yielding orientation feature maps. These maps encode, as a group, how different the average local orientation is between the center and surround scales. It is possible in our implementation to use an arbitrary number of orientations, but we noticed that using more oriented filters than the four abovementioned did not alter the performances of the model drastically.

2.4 The Saliency Map

Implementation of the Saliency Map

In section 1, we have discussed that "decoupling" of feature dimensions is computationally advantageous, and that one implementation of such decoupling is sequential scanning of different parts of the visual field. This leads to the metaphor of the "focus of attention" and requires a spatially defined selection scheme which controls where the focus of attention is deployed at any given time. In 1985, Koch and Ullman [KU85] suggested that an efficient way for the coordination of this control mechanism is in the form of a spatially organized feature map, which codes for the "saliency" of every location in the visual field.

The task of the saliency map is the computation of the salience at every location in the visual field and the subsequent selection of the most salient areas or objects. At any time, only one such area is selected. The output of the saliency map consists of a spike train from neurons corresponding to this selected area in the topographic map which projects to the ventral ("What") pathway.

Fusion Of Information

Once all relevant features have been computed in the various feature maps, they have to be combined to yield the salience, i.e. a scalar quantity. The main difficulty in this task is that features from different modalities have different dynamic ranges, and consequently are not directly comparable. Normalizing all feature maps to the same dynamic range is not a satisfactory solution because it artifactually amplifies those maps in which only a low, noisy response was originally present. Moreover, evidence has been found in monkey primary visual cortex [CH94] for the presence of non-linear amplification stages, whose gains vary with the responses of neighboring cells. This supports the idea that different feature maps should not be normalized to a fixed range, but rather according to their content. Also, psychophysical studies performed on humans have shown that strong competition between spatially close stimuli yields mutual lateral inhibition of areas in which a large number of conspicuous responses are present [Bra94].

We currently use a very simple normalization scheme, consisting of promoting those feature maps in which a small number of strong peaks of

activity are present, while suppressing feature maps eliciting comparable peak responses at numerous locations over the whole visual scene. The first step is to normalize all the feature maps to the same dynamic range, in order to eliminate across-modality amplitude differences due to dissimilar feature extraction mechanisms. In biological systems, this step may reflect rapidly adapting fine tuning of weighting factors assigned to the various feature maps. Then, for each map, the global maximum M of the activity is found, and an average $\overline{m}$ of all the other local maxima in the map is computed. Finally, the map is globally multiplied by $(M-\overline{m})^2$. The direct consequence of this normalization scheme is to strongly amplify those feature maps in which the most conspicuous location (global maximum) elicits a much stronger response than, on average, the other conspicuous locations (secondary maxima). Such contents-based amplification of the feature maps might be implemented in biological systems (in a local neighborhood rather than for the whole visual field) by lateral inhibition mechanisms, suppressing large areas of rather uniform peak activities while enhancing strong, isolated peaks.

After normalization, feature maps are combined into three separate channels for intensity, color, and orientation. These combinations are performed across scales and within each channel to yield three conspicuity maps, at the spatial scale of the saliency map. The reduction of each feature map to the scale of the saliency map is obtained by using auxiliary Gaussian pyramids. The intensity conspicuity map is simply obtained by adding up the six normalized (using the normalization method described previously) and downscaled intensity feature maps. The color conspicuity map is obtained by adding up the six red-green and six blue-yellow normalized and downscaled feature maps. Four intermediate orientation maps are first obtained by adding up the six normalized and downscaled center-surround maps for a given orientation (0, 45, 90 or 135 degrees). Each of these four maps is then normalized by the same method as used previously. The four orientation maps are finally added up, to form the orientation conspicuity map. The three conspicuity maps for intensity, color and orientation each undergo one final normalization, and are then added together, to form the final input to the saliency map. A total of 50 maps is hence computed by our system (42 feature maps, 4 intermediate orientation maps, 3 conspicuity maps, and the final input to the saliency map neural network). The motivation for the creation of three separate channels and their individual dedicated normalization is the hypothesis that similar features compete strongly for salience, while different modalities contribute independently to the saliency map.

The biophysical assumption underlying the linear combination of normalized feature maps is that of linear summation of incoming synaptic potentials in the dendritic tree. This feed-forward procedure for the fusion of information has the advantage of being very fast compared to relaxation methods suggested in a similar context [MPW93].

Internal Dynamics And Trajectory Generation

By definition, the activity in a given location of the saliency map represents the relative conspicuity of the corresponding location in the visual field. At any given time, the maximum of this map is therefore the most salient stimulus to which the focus of attention should be directed next, to allow more detailed inspection by the "what" pathway with its powerful object recognition capabilities not available to the "where" pathway. To find the most salient location, we have to determine the maximum of the saliency map.

This maximum is selected by application of a winner-take-all mechanism. Different mechanisms have been suggested for the implementation of neural winner-take-all networks [KU85, YG89]. In our model, we used a 2-dimensional layer of integrate-and-fire neurons with strong global inhibition in which the inhibitory population is reliably activated by any neuron in the layer[1]. Therefore, when the first of these cells fires, it will inhibit all cells (including itself), and the neuron with the strongest input will generate a sequence of action potentials. All other neurons are quiescent.

For a static image, the system described so far would attend continuously the most conspicuous stimulus. This is neither observed in biological vision nor desirable from a functional point of view; instead, after inspection of any point, there is usually no reason to dwell on it any longer and the next-most salient point should be attended. An additional temporal effect comes into play after the focus of attention has moved away from a momentaneously attended location. There is strong psychophysical evidence that the visual system tries to avoid shifting back the focus of attention to a location which it has just visited. One of the most direct observations of this behavior is in terms of reaction times which are longer when the subject is requested to return attention to a location which had just[2] been attended [PC84, WH97, DKS98]. Kwak and Egeth [KE92] showed that this "inhibition of return" has a strong spatially defined component i.e. the return is inhibited to the *location* of the last attended item. Although location is not the only feature on which inhibition of return is based [LPA95], this behavior is in agreement with previously mentioned data [Nis85, TL93, SS93] which showed that spatial location has a somewhat special role among object features.

We achieve this behavior by introducing feedback from the winner-take-all (WTA) array to the saliency map. When a spike occurs in the WTA network, the integrators in the saliency map receive additional input with the spatial structure of an inverted Mexican hat, i.e. a difference of Gaus-

[1] A more realistic implementation would consist of populations of neurons. For simplicity, we model such populations by a single neuron with very strong synapses.

[2] A lengthening of the reaction time was observed if the time between the cue and the stimulus exceeded about 300 ms and was less than about 1.5 s

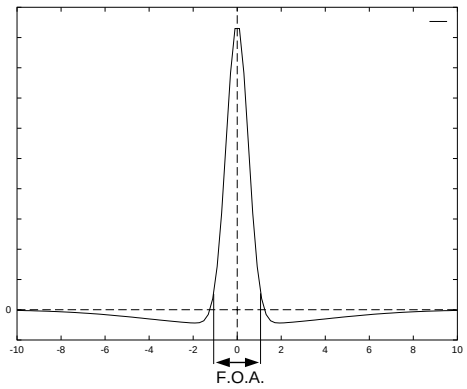

FIGURE 2. Radial section of the inhibitory feedback triggered around the last attended location. The central lobe provides strong inhibition which causes the focus of attention ("F.O.A.") to jump towards another location. The left and right excitatory lobes (whose sign is, of course, opposite to that of the central inhibitory lobe) slightly enhance the salience of the immediate surround of the currently attended location, favoring short jumps of the focus of attention over long jumps.

sians (Figure 2). The inhibitory center (with a standard deviation of half the radius of the focus of attention) is at the location of the winner. Cells in the saliency map located there receive maximal inhibition and this location, which used to be the global maximum of saliency, now becomes a (local or global) minimum of the saliency map. As a consequence, attention switches to the next-most conspicuous location (Figure 3). The function of the positive lobes (half width of four times the radius of the focus of attention) of the inverted Mexican hat is to favor locality in the displacements of the focus of attention: if two locations are of nearly equal conspicuity and one of them is close to the previously attended location and the other is far away, attention will jump to the closer location rather than to the distant one.

The inhibitory feedback of the saliency map is consistent with the aforementioned inhibition-of-return phenomenon observed psychophysically. Not all observed properties of the inhibition-of-return phenomenon can be explained, however, in the present version of our model. For instance, we assume that the location at which inhibition-of-return occurs is computed relative to the global stimulus environment. It has been observed, in fact, that this location can be relative to other stimuli [TDW91]. If more than one frame of reference is available (e.g. the global environment and a nearby object), the coordinates are combined in a nontrivial, not fully understood way [GE94]. From a functional point of view, it seems desirable that inhibition of return should act in a world-based frame of reference, rather than in a retina-based coordinate system as our model does. This requires that a model of attentional selection which includes the inhibition-of-return

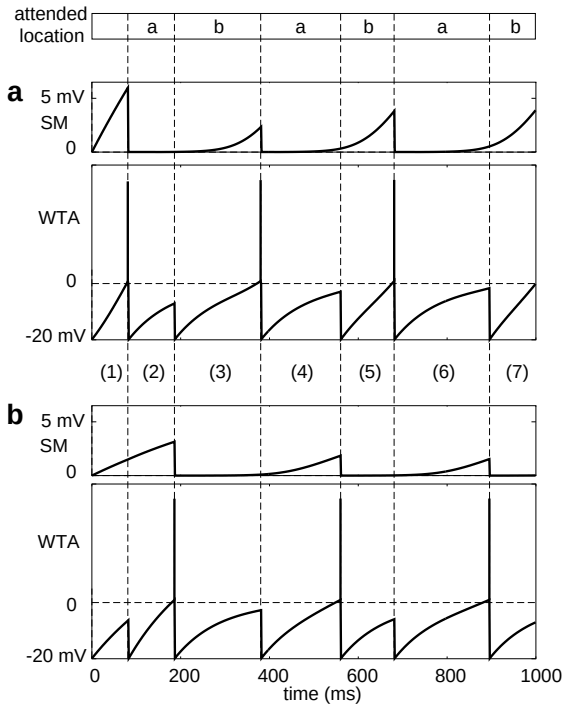

FIGURE 3. Dynamical evolution of the potential of some simulated neurons in the saliency map (SM) and in the winner-take-all (WTA) networks. The input contains one salient location (a), and another input of half the saliency (b); the potentials of the corresponding neurons in the SM and WTA are shown as a function of time. During period **(1)**, the potential of both SM neurons (a) and (b) increases as a result of the input. The potential in the WTA neurons, which receive inputs from the corresponding SM neurons but have much faster time constants, increases faster. The WTA neurons evolve independently of each other as long as they are not firing. At about 80ms, WTA neuron (a) reaches threshold and fires. A cascade of events follows: First, the focus of attention is shifted to (a); second, both WTA neurons are reset; third, inhibition-of-return (IOR) is triggered, and inhibits SM neuron (a) with a strength proportional to that neuron's potential (i.e., more salient locations receive more IOR, so that all attended locations will recover from IOR in approximately the same time). In period **(2)**, the potential of WTA neuron (a) rises at a much slower rate, because SM neuron (a) is strongly inhibited by IOR. WTA neuron (b) hence reaches threshold first. **(3)—(7):** In this example with only two active locations, the system alternatively attends to (a) and (b). Note how the IOR decays over time, allowing for each location to be attended several times. Also note how the amount of IOR is proportional to the SM potential when IOR is triggered (e.g., SM neuron (a) receives more IOR at the end of period (1) than at the end of period (3)). Finally, note how the SM neurons do not have an opportunity to reach threshold (at 20 mV) and to fire (their threshold is ignored in the model). Since our input images are noisy, we did not explicitly incorporate noise into the neurons' dynamics.

location takes into account the different coordinate systems and their interactions. However, recent psychophysical results [HW98] indicate that visual search strategies make use of surprisingly little information across movements of the focus of attention, thus obviating the need to keep track of different coordinate systems. If substantiated, these results clearly support our model.

3 Simulation Results

We have studied the behavior of the system with input from two large classes of simulated visual stimuli. The first are visual scenes constructed analogously to the stimuli typically presented in psychophysical studies of visual search (discussed in section 3.1). The second class of input are color images of natural and artificial environments. Results from the second class are discussed in Section 3.2.

3.1 Synthetic Stimuli

One of the simplest possible tasks is the detection of a bright spots on a dark backgrounds, or dark spots in bright backgrounds. This task is solved reliably by our model, and the focus of attention immediately jumps to such stimuli. If there is more than one such stimulus, the system scans them one-by-one, in the order of decreasing contrast from the background. The same is true for stimuli that have a color or orientation different from that of the background.

More challenging tasks are defined by the typical stimuli employed in visual search tasks, which is one of the most active fields of human psychophysics. A typical experiment consists in a speeded alternative forced-choice task in which the presence of a certain item in the presented display has to be either confirmed or denied. Salient visual features "guide" attention to the stimuli possessing these features, and these stimuli are selected efficiently. In particular, it is known that stimuli which differ from nearby stimuli in one or more feature dimensions can be easily found in visual search, typically in a time which is nearly independent of the number of other items ("distractors") in the visual scene. In contrast, search times for targets which differ from distractors by a combination of features (a so-called "conjunctive task") are typically proportional to the number of distractors.

In 1980, Treisman and Gelade [TG80] published an elegant explanation of this fact. The underlying computational principle is that the detection of a few elementary features is most economically done by massively parallel processes early in the visual hierarchy. Because the number of possible combination of features increases very rapidly with the number of features

to be combined, there are parallel, preattentive maps only for the elementary features. "Conjunctions" of the features can only be processed by a central attentional authority, which gives rise to the sequential attentional process. Although this "Feature Integration Theory" explained many properties of visual search, it was shown later that it is invalid in its simplest form [EVG84, WCF89, MYE90, Pal94].

We reproduced one of the original experiments used by Treisman and Gelade, in order to relate the performances of our model on a "simulated psychophysical experiment" to human psychophysics. Artificial stimulation images were generated by the computer. They were of three classes: (1) one red target (rectangular bar) among green distractors (also rectangular bars) with the same orientation; (2) one red target among red distractors with orthogonal orientation; and (3) one red target among green distractors with the same orientation and red distractors with orthogonal orientation. In order not to artifactually favor any particular direction, the orientation of the target was chosen randomly for every image generated. Also, in order not to obtain ceiling performance in the first two tasks (100% pop-out), we added strong orientation noise to the stimuli (between -17 and +17 degrees with uniform probability) and strong speckle noise to the images (each pixel had a 15% probability, drawn from a uniform distribution, to become a maximally bright red, green, blue, cyan, purple, yellow or white). The positioning of the stimuli along a uniform grid was also randomized (by up to $\pm$ 40% of the spacing between stimuli, in the horizontal and vertical directions), to eliminate any possible influence of our discrete image representations (pixels) on the system. Twenty images were computed for a total number of stimuli per image varying between 4 and 36, yielding the evaluation of a total of 540 images.

Results are presented in Figure 4. Clear pop-out was obtained for the first two tasks (color only and orientation only), independently of the number of distractors in the images. Slightly worse performances are found when the number of distractors is very small, which seems sensible since in these cases the distractors are nearly as conspicuous as the target itself. Evaluation of these types of stimulation images without introducing any of the distracting noises exposed above yielded 100% pop-out (target found as the first attended location) in all images. The conjunctive search task yielded a linear increase, with the number of distractors, in the number of false detections prior to the correct detection of the target. This result is in good accordance with human psychophysics. Notice that the large error bars in our results indicate that our model usually finds the target either quickly (in most cases) or after scanning a very large number of locations. The target was always found, and the system was never trapped into a recurrent cycle passing through only a limited number of locations.

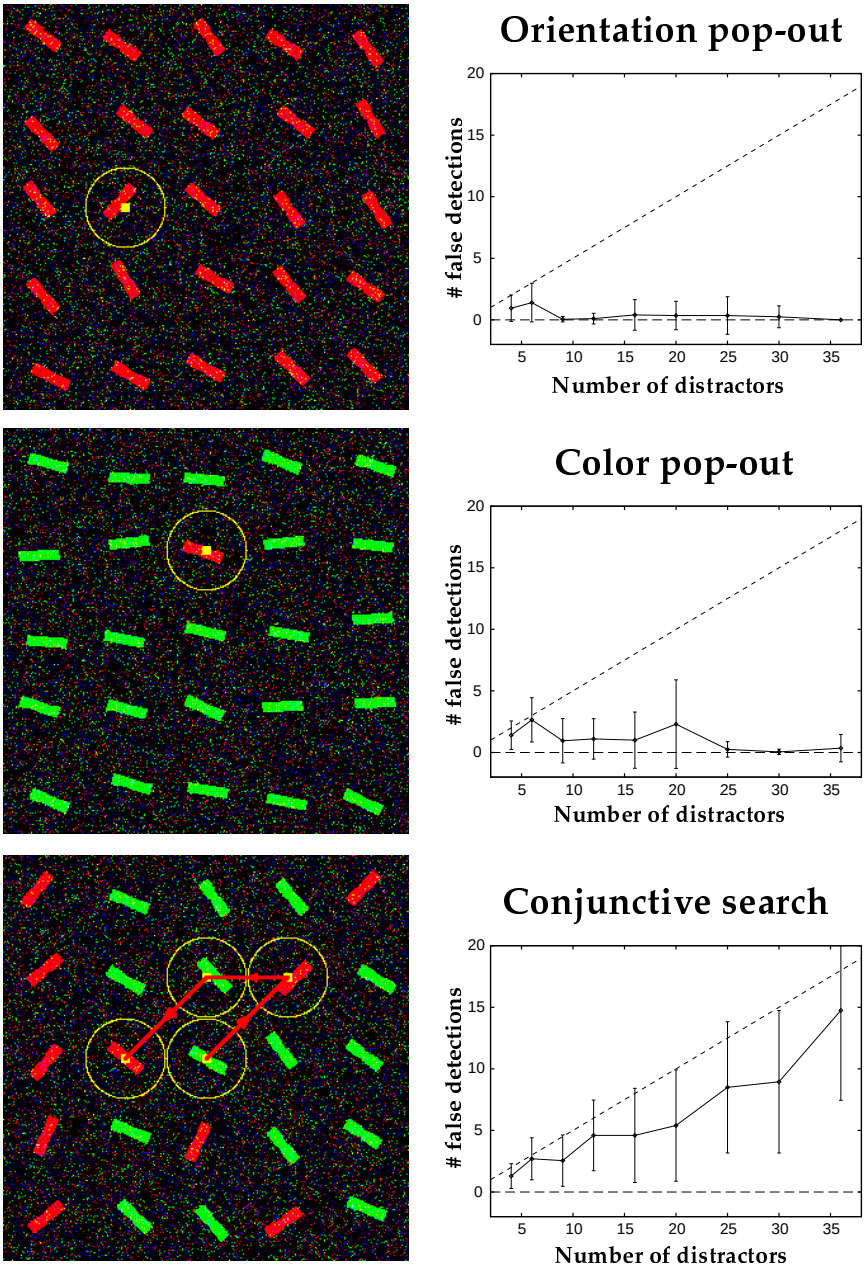

FIGURE 4. Performance of our current visual attention model on the pop-out and conjunctive tasks pioneered by A. Treisman. The typical search slopes of human observers in feature and conjunction search, respectively, are successfully reproduced by the model.

3.2 Natural Images

One of the most severe tests an artificial vision system can undergo is the evaluation of its performance when applied to natural images. We have therefore studied the behavior of our model attentional system using such images as input and we describe some of the results in this section. A substantial difficulty is, however, that it is not straightforward to establish objective criteria for the performance of the system. Unfortunately, nearly all quantitative psychophysical data on attentional control have been obtained based on synthetic stimuli similar to those discussed in section 3.1. Consequently, we are largely limited to plausibility arguments, i.e. we have to judge the performance on natural images by making arguments about the probable functional significance and usefulness of the strategy the system uses. The scan paths of overt attention (eye movements) are much better known than those of covert attention [Yar67, NS71, LF97]. It is unclear, however, to what extent these scan paths are similar to the motion of covert attention since the requirements and limitations (e.g. spatial and temporal resolutions) of the two systems are most likely quite different.

We tested our model on a wide variety of real images, ranging from natural outdoor scenes to artistic paintings. All images were in color, contained significant amounts of noise, strong local variations in illumination, shadows and reflections, large numbers of "objects" often partially obstructed, and strong textures. We present in Figures 5 and 6 some trajectories obtained without introducing any particular modification or tuning in the model. The examples shown (as well as some others we studied and which are not illustrated here) indicate that the system scans the image in an order which makes functional sense in most behavioral situations.

One particularly interesting application is to use the model as a target detector in complex natural scenes. Although our algorithm will focus on salient objects irrespectively of their nature, man-made objects usually are fairly salient in natural environments and are quickly detected by the system. Examples of such applications which we have investigated include the detection of traffic signs on low-resolution color video frames (Figure 5), or the detection of very small military vehicles (Figure 7) in very large digitized photographs (6144×4096 pixels) [TBKV98]. Further details about these applications may be found on our World-Wide-Web site, at `http://www.klab.caltech.edu/~itti/attention/`.

4 Discussion

4.1 Psychophysical and physiological basis of the model

We present in this report a prototype for a system mimicking the control of visual selective attention. The model identifies the most salient points in

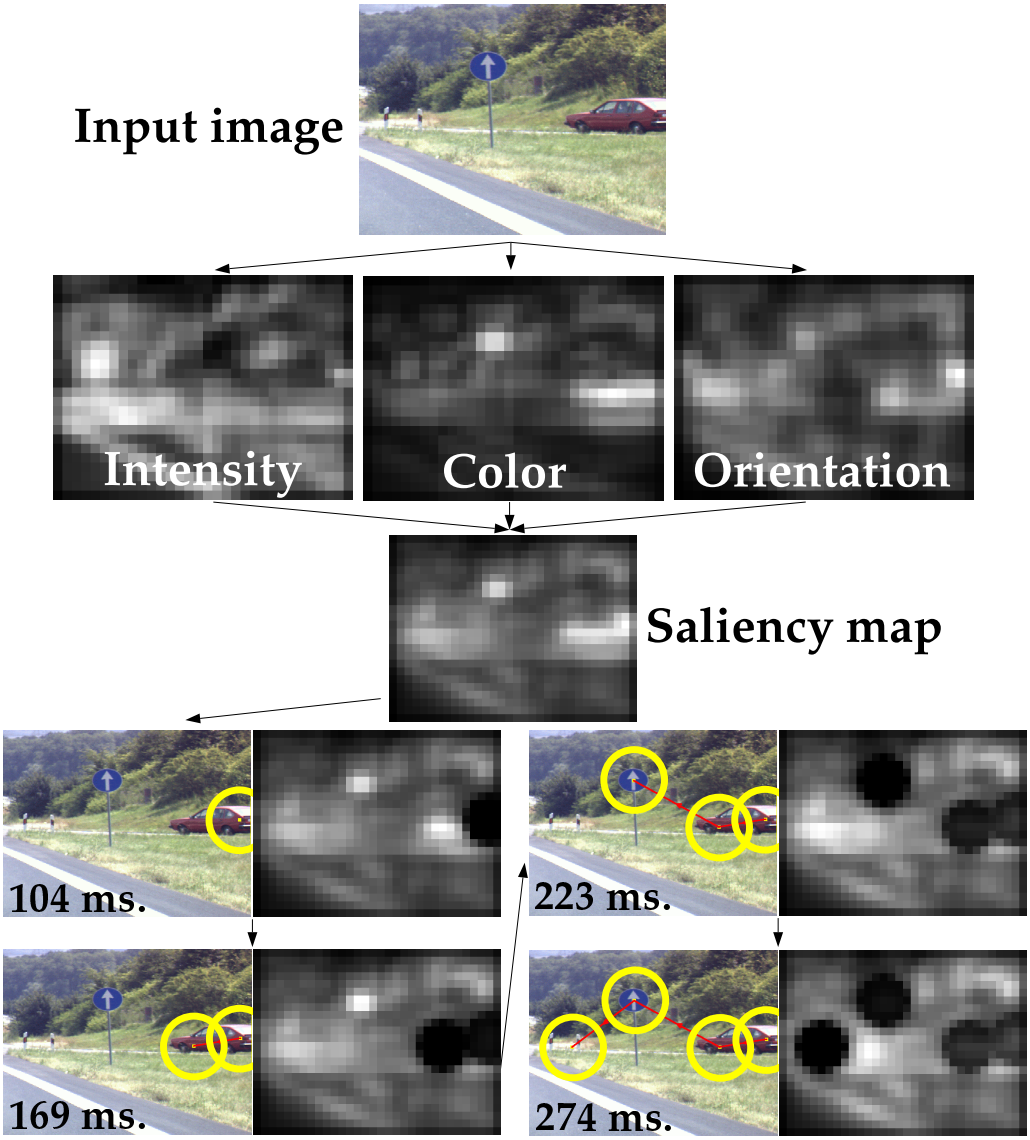

FIGURE 5. Example of the global working of our model. Feature maps extracted from the input image at several scales are combined into the saliency map. A winner-take-all neural network then successively selects, in order of decreasing saliency, the attended locations. Once a location has been attended, it is transiently suppressed by the inhibition of return mechanism. Note how the inhibited locations recover over time (e.g. the first attended location has regained some activity at 274 ms).

FIGURE 6. More examples of the application of the model to real-world images. Shown are only the trajectories of the focus of attention.

a visual scenes one-by-one and scans the scene autonomously in the order of decreasing saliency. This allows the control of a subsequently activated processor which is specialized for detailed object recognition.

The model is formulated in terms of interacting neuronal populations. The elements of the model are integrate-and-fire neurons, a crude but not unreasonable approximation for many neurons in the nervous system. Our model is compatible with the known anatomy and physiology of the primate visual system, and the way its different parts communicate by signals which are neurally plausible. In particular, there is evidence for a representation of visual information in terms of feature maps, for the generation of the feature maps in terms of center-surround operations etc; for a recent example of physiological support, see [NGVE99].

In addition to its biologically plausible structure and to the realistic input the model operates on, it also provides output which can be used immediately for information selection. The output of the model consists of elevated activity at the location representing the instantaneously attended location in visual space in a topographic representation. This is exactly the type of input required for attentional selection in the ventral pathway. In previous work [NK94], we developed a model which can use the output from the present model to implement the attentional selection process for object recognition. The same is true for a variation of that model [NKR93] except that it would require that the output of the "where" pathway has a

FIGURE 7. Example of detection of a military vehicle in a high-resolution (6144 × 4096 pixels) color photograph. This image is part of a database of 44 images for which human search times have been measured [TBKV98]. After scaling of the model's time such that it made three attentional shifts per second on average, and addition of 1.5 sec. to account for human motor reaction time, the model found the target faster than humans in 75% of the images.

periodic (repeating in time) structure, which could be generated either by intrinsically oscillating neurons or by network effects. The present model demonstrates a neurally plausible substrate that can generate the appropriate control signals which select visual targets from their background.

Sequential scanning of the visual scene requires movement of the focus of attention, analogously to eye movements[3]. Several independent sets of data indicate that the time required for one shift of attention is on the order of $30 - 50$ ms (for a direct measurement, see ref. [SJ91]) which is consistent with the behavior of the model. Longer dwell-times of the focus of attention are observed in conditions in which a stimulus is flashed, the subject attends to it, then the next stimulus is flashed and attended *etc* [DWS94]. It is possible that flashing a stimulus disrupts the attentional process, possibly by its action on the saliency map.

There is stronger evidence in favor of a *functional* saliency map than there is for a *spatially localized* saliency map. In other words, the functionality of a saliency map may be spread over different anatomical areas. Robinson and Petersen [RP92] reviewed data showing that the pulvinar nuclei of the thalamus play a significant role in the selection of visual targets. However, it seems unlikely that the pulvinar is the only location implicated in this selection process. Other candidate areas are the posterior parietal cortex [BGR81, MAM81, Mou95] and the superior colliculus [GW72]. A possible scenario is that of a saliency map distributed over two or more of these structures.

Psychophysical evidence indicates that the map may not be organized retinotopically but instead with a coordinate system relative to the visual environment of the observer [PC84] or relative to the observed objects [TDW91]. Recent data provides evidence for a neural substrate of object-centered coordinate systems, at least for overt attention [OG95] and the functional arguments in favor of coordinate systems varying with the visual scene and task requirements. On the other hand, we have already mentioned results [HW98] which indicate that surprisingly little information may be conserved across movements of the focus of attention. If these results turn out to be of general validity, our simple retinotopic coordinate system were all that is required. More experimental evidence is required before the question of the coordinate system can be answered authoritatively.

The dynamics of the saliency map are determined in our model by the interplay between the winner-take-all mechanism and the feedback provided at the location of the instantaneous winner. As a result, the system will find a winner, direct attention to its location, and then move on to the next-salient location by suppressing the activity at the previous winner. One of

[3] Analogously to the physical limitions which allow only one point of fixation for the eyes, there is strong evidence that the focus of attention is unique, i.e. there is only one focus of attention at any given time. An interesting exception are split-brain patients who may dispose of two foci, presumably one for each hemisphere [Gaz89, LHMG94]

the side effects of these dynamics is that the system will have a tendency to avoid the location of a previous winner, i.e. the previous focus of attention, for some time. The existence of this effect ("inhibition of return") is well-established in the psychophysical literature [PC84, MH85, GE94, TDW91]. Furthermore, suppression of the attended location has been observed experimentally in area 7a of rhesus monkeys performing a match-to-sample task [SCCM94, SC95]. Inhibition of return is a natural consequence of our model.

4.2 Limitations of the model

Our model does not explain every aspect of attentional control. We have already mentioned that, for the time being, we focus on stimulus-driven (bottom-up) control and make no attempt to implement top-down attentional influences in any detail. The prototypical top-down attention task is perhaps the one formalized by Posner [Pos80] although its roots go back to, at least, Hermann von Helmholtz [Hel67]. In this task, subjects are instructed to attend to one part of their visual field which is identified to them by information requiring cognitive processing or memory or both, e.g. by an arrow pointing towards the area to be attended, or by a verbal command ("attend to the upper left quadrant"). A robust increase in performance, either measured as decreased reaction time or increased accuracy, is observed for stimuli appearing in the attended area. It is likely that at least some aspects of this effect could be captured in our model by providing additional input to the saliency map

Top-down attention is not limited to space-based selection. We can attend to objects and events using non-spatial criteria. One well-controlled experimental paradigm which involves selection based on the memory of non-spatial properties is the delayed-match-to-sample paradigm (e.g, [CMLD93]), in which the animal subject has to select one of several stimuli based on a cue stimulus presented some time (typically, a few seconds) earlier and held in working memory. In work complementary to the present approach, Usher and Niebur recently developed a model for attentional selection in this condition [UN96]. Another approach to object-based attentional selection was recently modelled by Ballard and collaborators [RZHB97]. Using a representation in terms of multiple scales of oriented filters and chromatic information, very similar to that presented here and in our previous report [NK96], they modeled selection strategies for saccadic eye movements ("overt attention") by matching iconic target and scene representations. Clearly, this approach could be used without any change in the framework of the present model.

There are other phenomena that our model does not explain explicitly but which possibly could be incorporated relatively easily. One example are "express" attentional shifts, first observed by Mackeben and Nakayama [MN93]. These authors found that their subjects can generate rapid shifts

of attention by using exactly the same paradigm supposed to underlie express saccades [FR84], i.e. by turning the fixation point off and thus facilitating disengagement of attention. A related hypothesis for the interaction between attention and express saccades was put forward by Fischer and Weber [FW93]. It has been proposed that express saccades can be explained in a framework similar to that proposed in the present report as "normal" saccades with anormally short dwell times. The short dwell time was proposed to be caused by very rapid updating of the saliency map and consequently rapid issue of the motor command to execute the next eye movement [RZHB97]. A similar mechanism might be at work for covert attentional shifts rather than eye movements; if so, express attentional shifts could be explained in a similar way by the present model.

We do not take into account a wealth of psychophysical results on the finer properties of stimuli and their interactions. An example is the observation of [Wol94] showing that the metrics in different feature maps may be different from those chosen in our model. For instance, he found that mirror-symmetric orientations are more similar to each other than the numerical values for the angular separation would indicate. Another example are part-whole interactions between elements of search stimuli [WFHB94] or the somewhat special role that color seems to play, insofar as it was the only feature tested by Nothdurft [Not93a] that did not require a local feature contrast for parallel detection. Again, it appears that our model should be general enough to allow such effects to be added.

Our model does not explain grouping (see ref. [HM93] for a connectionist model which focuses on grouping in visual search). There is some evidence indicating that grouping (at least of texture elements) is performed not preattentively but at a second (attentive) perceptual stage; for instance, Nothdurft found that only salient stimuli group [Not92]. This would indicate that grouping happens at a stage of processing which is beyond that modeled in the present work.

4.3 Relation to other models

Our model is a member of a much larger class of models based on the dichotomy between parallel, effortless, pre-attentive processing and sequential, effortful, attentive processing [Bro58, Nei67, Hof78, Hof79]. Variations of the saliency map concept have been used in multiple instances and with different terminologies (e.g. the "Master Map" of Treisman [Tre88] or the "Activation Map" of Wolfe and collaborators [WCF89, Wol94]). Desimone and Duncan [DD95] recently suggested that no saliency map may be required at all and selective attention is instead a consequence of interactions between feature maps only. While this is certainly a possibility, one argument in favor of a saliency map is that it provides a convenient structure for the fusion of information required to compute a single location from the data from a multitude of feature maps.

There is strong psychophysical evidence indicating that integration of information across visual dimensions takes place in attentional selection [Not93b]. Even when saliency was produced in a feature domain irrelevant for the task, targets were detected as quickly as when saliency was generated completely within the relevant feature dimension. Such integration across dimensions (and nonspecific with respect to the target properties) is exactly what the saliency map introduced in section 2.4 is based upon.

A position intermediate between the protagonists and antagonists of a saliency map is taken by Braun and Sagi [BS90, BS91] who suggested that bottom-up and top-down mechanism have different capabilities and selection criteria and that only the bottom-up control relies on a saliency map. Since we focus in this work on bottom-up influences, our model is compatible with their results.

Another model in the mentioned class was developed over the last years by Wolfe and collaborators [WCF89, Wol94]. In previous work, most notably in the earlier versions of Treisman's influential Feature Integration Theory, the activity in preattentive feature maps is observed by a cognitive process which is only capable of deciding whether a given stimulus is the target or not. In contrast, Wolfe et al. assumed that activities in several feature maps can be combined and "guide" the focus of attention towards the most promising locations (thus, "Guided Search.") This aspect is, of course, very similar to our approach and, as a consequence, many predictions made by our model are also made by Guided Search (e.g., triple conjunctions should be easier to find than simple conjunctions; this was confirmed experimentally [WCF89]). There are, however, significant differences to our model. The most important is the level of modeling. While our model is explicitly formulated in terms of neuronal populations and their interactions, Guided Search is a functional model without immediate constraints imposed by a physiological substrate.

Other models were described over the last years in the connectionist literature [FMI83, HM93, GMR94, San89, AO91, MBP91, SW89, JMP94]. In this paradigm, networks are constructed from interacting units (assumed to roughly correspond to neurons or groups of neurons) which are connected in various ways. The basic functions in many connectionist networks are, however, quite different from those of biological neurons (for instance, units may exchange informations about pointers, abstract addresses, etc.). This makes it difficult to compare predictions of connectionist architectures with physiological observations.

In contrast to these connectionist approaches, Olshausen and collaborators [OAVE93] developed a model for the neural basis of attention at a similar level of neural plausibility as the one presented in this report. Their model is based on the assumption of "shifter circuits" which switch the synaptic input to neurons in higher areas and thereby select which part of the sensory information is made accessible to higher cortical areas. Since in this model, the input to the neurons is gated while in our model,

the activity of the cells themselves is modulated, Desimone [Des92] called the model in ref [OAVE93] "input-gated" and models of the class in this report "cell-gated". Both models make clear predictions which should be experimentally verifiable. For instance, one of the defining properties of the [OAVE93] model is the conservation of spatial relations within the focus of attention. In contrast, since *all* stimuli in the focus of attention are tagged in the model in this report, the loss of spatial relationships within the focus of attention is predicted by our model.

4.4 Predictions

The mammalian visual system is characterized by a sequence of cortical areas which, despite all their differential specializations, shows a uniform trend: going from close to the sensory periphery to more central areas, neuronal responses are characterized by a simultaneous increase in feature specifity and a decrease in spatial specificity. For instance, many cells in area V1 (close to the sensory periphery) respond to *any* elongated structure of a certain orientation (close to the preferred orientation of the cell under study), provided it is in its *small* receptive field. On the other hand, cells in inferotemporal cortex (distant from the periphery) will only respond to very *specific* stimuli, e.g. a face, but they will do so in a very *large* receptive field.

One reason why this architecture evolved may be found in the combinatorial character of highly specific stimuli: it is simply not possible to provide more than a few cells which are sensitive to such stimuli, and these cells therefore have to be usable in large parts of the visual field. It was realized early on that this architecture leads, however, to a vexing problem: since the location information seems to be largely lost in higher areas, how do these neurons "know" where to attribute the different properties of two or more stimuli present simultaneously in the visual field?

Over the last few years, several groups have proposed that this so-called "binding problem" may be solved by attaching "tags" to the different parts of an object whose neuronal realization is the temporal structure of the spiketrains coding for each object [CK90, EKA$^+$92, NKR93, NK94, KELS95]. In the present model, the problem would be solved in a very simple way, or rather, it would not exist in the first place: We suggest that at any given time, only one out of the possibly many simultaneously present stimuli is selected and all other stimuli are suppressed[4] This seems to be one of the simplest methods to solve the "binding problem." Thus, attention controls access to visual awareness and we can only be aware of a

[4]The model presented in this report is only concerned with the selection of the attended stimuli, not the suppression of unattended stimuli. In earlier work [NKR93, NK94], we have shown how "tags" consisting of temporal modulation of spike trains can be used to inhibit neurons responding to non-attended stimuli.

single stimulus at a time, compatible with much psychophysical evidence.

Support for our model and, in fact, all models which are based on the existence of an anatomically identifiable saliency map has been provided recently by Friedman-Hill and coworkers [FHRT95]. These authors identified a patient with bilateral parietal-occipital lesions who showed exactly the kind of problems to be expected in a patient who is lacking significant parts of a saliency map. In particular, this patient routinely shows evidence of "illusory conjunctions," i.e. he miscombines colors and shapes even under free viewing conditions (i.e., in the absence of high perceptual load which is required to generate illusory conjunctions in normal subjects). In the context of our theory, we would predict that the absence of a saliency map (or important parts of it) leads to the absence of the signal which is used to distinguish the (usually unique) attended object from all others, therefore leading to the miscombination of the features of several objects.

A prediction of our model is that the input to central sensory areas (e.g. the inferotemporal areas) should rapidly change under free viewing conditions in a complex environment. This prediction should be verifiable in electrophysiological experiments, but there are only few studies of cortical neuronal responses under free viewing conditions [GCDVE95, GCVE98].

Another prediction is that inhibition should be observed at the attended location in structures likely to control the position of the focus of attention, due to the inhibition underlying the scanning mechanism discussed in section 2.4. This is a particularly valuable prediction because it seems counterintuitive. Furthermore, in earlier work, *enhanced* activity was observed in neurons representing the attended location [BGR81]. However, more recent recordings in posterior parietal cortex of awake behaving monkey, using more consistent deployment of attention to identified locations in the visual field than it was the case in the earlier work, provided evidence for a substantial suppression of activity of those neurons that represent the attended location [SCCM94]. The fate of our model is at this time in the hands of experimentalists.

Acknowledgments

Work at Caltech on this project was supported by NSF, the NSF- funded Center for Neuromorphic Systems Engineering and by ONR. Work at Johns Hopkins University was supported by NSF and by the Alfred P. Sloan Foundation.

5 References

[AAB+84] E.H. Adelson, C.H. Anderson, J.R. Bergen, P.J. Burt, and J.M. Ogden. Pyramid methods in image processing. *RCA Engineer*, Nov-Dec., 1984.

[AO91] S. Ahmad and S. Omohundro. Efficient visual search: a connectionist solution. In *Proc. 13th Ann. Conf. Cog. Sci. Soc.* 1991.

[BD99] J. A. Brefczynski and E. A. DeYoe. A physiological correlate of the spotlight of visual attention. *Nature Neuroscience*, 2:370–374, 1999.

[BG83] G. Buchsbaum and A. Gottschalk. Trichomacy, opponent colour coding and optimum colour information transmission in the retina. *Proceedings of the Royal Society of London B*, 220:89, 1983.

[BGR81] M. C. Bushnell, M. E. Goldberg, and D. L. Robinson. Behavioral enhancement of visual responses in monkey cerebral cortex. I. Modulation in posterior parietal cortex related to selective visual attention. *J. Neurophysiol.*, 46:755–772, 1981.

[Bra94] J. Braun. Visual search among items of different salience: removal of visual attention mimics a lesion in extrastriate area V4. *J. Neuroscience*, 14:554–567, 1994.

[Bro58] D. E. Broadbent. *Perception and Communication.* Pergamon, London, 1958.

[BS90] J. Braun and D. Sagi. Vision outside the focus of attention. *Perception and Psychophysics*, 48:45–58, 1990.

[BS91] J. Braun and D. Sagi. Texture-based tasks are little affected by second tasks requiring peripheral or central attentive fixation. *Perception*, 20:483–500, 1991.

[BS99] A. J. Bell and T. J. Sejnowski. The independent components of natural scenes are edge filters. *Vision Research*, 37:3327–38, 1999.

[Bun91] C. Bundesen. Visual selection of features and objects: is location special? a reinterpretation of Nissen's (1985) findings. *Perception & Psychophysics*, 50:87–89, 1991.

[CH94] M. Carandini and D.J. Heeger. Summation and division by neurons in primate visual cortex. *Science*, 264:1333–1336, 1994.

[CK90] F. Crick and C. Koch. Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, 2:263–275, 1990.

[CMLD93] L. Chelazzi, E.K. Miller, A. Lueschow, and R. Desimone. Dual mechanisms of short-term memory: ventral prefrontal cortex. *Soc. for Neuroscience Abstracts*, 19:975, 1993.

[DD95] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Ann. Rev. Of Neurosci.*, 18:193–222, 1995.

[Des92] R. Desimone. Neural circuits for visual attention in the primate brain. In G. Carpenter and S. Grossberg, editors, *Neural networks for vision and image processing*. MIT Press, Cambridge, 1992.

[DKS98] S. Danziger, A. Kingstone, and J. Snyder. Inhibition of return to successively stimulated locations in a sequential visual search paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, 24(5):1467–75, Oct 1998.

[DWS94] J. Duncan, R. Ward, and K. Shapiro. Direct measurement of attentional dwell time in human vision. *Nature*, 369:313–315, 1994.

[EKA+92] A. K. Engel, P. König, Kreiter A.K., Th. B. Schillen, and W. Singer. Temporal coding in the visual system: new vistas on integration in the nervous system. *Trends in Neurociences*, 15:218–226, 1992.

[EVG84] H.E. Egeth, R.A. Virzi, and H. Garbart. Searching for conjunctively defined targets. *J. Experimental Psychology*, 10(1):32–39, 1984.

[FHRT95] S. R. Friedman-Hill, L. C. Robertson, and A. Treisman. Parietal contributions to visual feature binding: evidence from a patient with bilateral lesions. *Science*, 269:853–855, 1995.

[FMI83] K. Fukushima, S. Miyake, and T. Ito. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 13(5):826–834, 1983.

[FR84] B. Fischer and E. Ramsperger. Human express-saccades: extremely short reaction times of goal directed eye movements. *Experimental Brain Research*, 57:191–195, 1984.

[FW93] B. Fischer and H. Weber. Express saccades and visual attention. *Behavioral and Brain Sciences*, 16:553–610, 1993.

[Gaz89] M.S. Gazzaniga. Independent hemispheric attentional systems mediate visual search in split-brain patients. *Nature*, 342:543–545, 1989.

[GBG+94] H. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C.H. Anderson. Overcomplete steerable pyramid filters and rotation invariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1994.

[GCDVE95] J. L. Gallant, C. E. Connor, H. Drury, and D. C. Van Essen. Neural responses in monkey visual cortex during free viewing of natural scenes – mechanisms of response suppression. *Investigative Ophthalmology and Visual Science*, 36(4):S1052, 1995.

[GCVE98] J. L. Gallant, C. E. Connor, and D. C. Van Essen. Neural activity in areas v1, v2 and v4 during free viewing of natural scenes compared to controlled viewing. *Neuroreport*, 9(9):2153–8, 1998.

[GE94] B.S. Gibson and H. Egeth. Inhibition of return to object-based and environment-based locations. *Perception & Psychophysics*, 55(3):323–339, 1994.

[GMR94] S. Grossberg, E. Mingolla, and W.D. Ross. A neural theory of attentive visual search: interactions at boundary, surface, spatial and object recognition. *Psychological Review*, 101(3):470–489, 1994.

[GW72] M. E. Goldberg and R. H. Wurtz. Activity of superior colliculus in behaving monkey II: The effect of attention on neuronal responses. *J. Neurophysiology*, 35(560-574), 1972.

[Hel67] H. von Helmholtz. *Handbuch der physiologischen Optik*. Voss, Leipzig, 1867.

[HM93] G.W. Humphreys and H.J. Müller. Search via recursive rejection (SERR): A connectionist model of visual search. *Cognitive Psychology*, 25:43–110, 1993.

[Hof78] J. E. Hoffman. Search through a sequentially presented visual display. *Perception & Psychophysics*, 23(1):1–11, 1978.

[Hof79] J. E. Hoffman. A two-stage model of visual search. *Perception & Psychophysics*, 2325(4):319–327, 1979.

[HW98] T. S. Horowitz and J. M. Wolfe. Visual search has not memory. *Nature*, 394:575–577, Aug. 1998.

[INK98] L. Itti, E. Niebur, and C. Koch. A model of saliency-based fast visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, November 1998.

[JMP94] S. R. Jackson, R. Marrocco, and M. I. Posner. Networks of anatomical areas controlling visuospatial attention. *Neural Networks*, 7(6/7):925–944, 1994.

[KE92] H. Kwak and H. Egeth. Consequences of allocating attention to locations and to other attributes. *Perception & Psychophysics*, 51(5):455–464, 1992.

[KELS95] P. König, A. K. Engel, S. Löwel, and W. Singer. How precise is neuronal synchronization? *Neural Computation*, 7(3):469–485, 1995.

[KU85] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiol.*, 4:219–227, 1985.

[LF97] M. F. Land and S. Furneaux. The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society of London B*, 352(1358):1231–9, 1997.

[LHMG94] S.J. Luck, S.A. Hillyard, G.R. Mangun, and M.S. Gazzaniga. Independent attentional scanning in the separated hemispheres of split-brain patients. *J. Cog. Neurosci*, 6(1):84–91, 1994.

[LN93] A. Lüschow and H. C. Nothdurft. Pop-out of orientation but no pop-out of motion at iso-luminance. *Vision Res.*, 33:91–104, 1993.

[LPA95] M. B. Law, J. Pratt, and R. A. Abrams. Color-based inhibition of return. *Perception & Psychophysics*, 57(3):402–408, 1995.

[MAM81] V. B. Mountcastle, R. A. Andersen, and B. C. Motter. The influence of attentive fixation upon the excitability of the light-sensitive neurons of the posterior parietal cortex. *J. Neurosci.*, 1:1218–1232, 1981.

[MBP91] R. Milanese, J.M. Bost, and T. Pun. Visual indexing with an attentive system. In *Lecture notes in artificial intelligence*, pages 415–419. Springer Verlag, Berlin, 1991.

[MH85] E. A. Maylor and R. Hockey. Inhibitory compoents of externally controlled covert orienting in visual space. *Journal of Experimental Psychology: Human Perception and Performance*, 11:777–787, 1985.

[MN93] M. Mackeben and K. Nakayama. Express attentional shifts. *Vision Research*, 33:85–90, 1993.

[Moo85] I. R. Moorhead. Human colour vision and natural images. In *Colour in information technology and information displays*, number 61, page 21. Institution of Electronic and Radio Engineers, Alderman, Ipswich, 1985.

[Mou95] V. B. Mountcastle. The parietal system and some higher brain functions. *Cerebral Cortex*, 5(5):377–390, 1995.

[MPW93] R. Milanese, T Pun, and H. Wechsler. A non-linear integration process for the selection of visual information. In V. Roberto, editor, *Intelligent perceptual systems*, pages 323–336. Springer Verlag, Berlin, 1993.

[MYE90] J.T. Mordkoff, S. Yantis, and H.E. Egeth. Detecting conjunctions of color and form in parallel. *Perception & Psychophysics*, 48(2):157–168, 1990.

[Nei67] U. Neisser. *Cognitive psychology.* Appleton-Century-Crofts, New York, 1967.

[NGVE99] H. C. Nothdurft, J. L. Gallant, and D. C.N Van Essen. Response modulation by texture surround in primate area V1: correlates of "popout" under anesthesia. *Visual Neuroscience*, 16(1):15–34, Jan-Feb 1999.

[Nis85] M.J. Nissen. Accessing features and objects: is location special? In M.I. Posner and O.S.M Marin, editors, *Mechanisms of atention: Attention and Performance XI*, pages 205–219. Hillsdale, NJ, 1985.

[NK94] E. Niebur and C. Koch. A model for the neuronal implementation of selective visual attention based on temporal correlation among neurons. *Journal of Computational Neuroscience*, 1(1):141–158, 1994.

[NK96] E. Niebur and C. Koch. Control of selective visual attention: Modeling the "where" pathway. In D. S Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 802–808. MIT Press, Cambridge, MA, 1996.

[NK98] E. Niebur and C. Koch. Computational architectures for attention. In R. Parasuraman, editor, *The Attentive Brain*, chapter 9, pages 163–186. MIT Press, Cambridge, MA, 1998.

[NKR93] E. Niebur, C. Koch, and C. Rosin. An oscillation-based model for the neural basis of attention. *Vision Research*, 33:2789–2802, 1993.

[Not92] H. C. Nothdurft. Feature analysis and the role of similarity in preattentive vision. *Perception & Psychophysics*, 52:355–375, 1992.

[Not93a] H. C. Nothdurft. The role of features in preattentive vision: comparison of orientation, motion and color cues. *Vision Res.*, 33:1937–1958, 1993.

[Not93b] H. C. Nothdurft. Saliency effects across dimensions in visual search. *Vision Res.*, 33:839–844, 1993.

[NS71] D. Noton and L. Stark. Scanpaths in eye movements. *Science*, 171:308–311, 1971.

[OAVE93] B. Olshausen, C. Andersen, and D. Van Essen. A neural model of visual attention and invariant pattern recognition. *J. Neuroscience*, 13(11):4700–4719, 1993.

[OF96] B. Olshausen and D. J. Fields. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[OG95] C. Olson and S. Gettner. Object-centered direction selectivity in the macaque supplementary eye field. *Science*, 269:985–988, August 1995.

[Pal94] J. Palmer. Set-size effects in visual search: the effect of attention is independent of the stimulus for simple tasks. *Vision Res.*, 34:1703–1721, 1994.

[PC84] M. I. Posner and Y. Cohen. Components of visual orienting. In H. Bouma and D. G. Bouwhuis, editors, *Attention and Performance X*, pages 531–556. Hilldale, NJ, 1984.

[Pos80] M.I. Posner. Orienting of attention. *Quart. J. Exp. Psychol.*, 32:3–25, 1980.

[RP92] D. L. Robinson and S. E. Petersen. The pulvinar and visual salience. *Trends in Neurociences*, 15(4):127–132, 1992.

[RZHB97] R. P. N. Rao, G. J. Zelinsky, M. M. Hayhoe, and D. H. Ballard. Eye movements in visual cognition: a computational study. Technical Report 97.1, Department of Computer Science, University of Rochester, March 1997.

[San89] P.A. Sandon. An attentional hierarchy. *Behavioral and Brain Sciences*, 12:414–415, 1989.

[SC95] M. A. Steinmetz and C. Constantinidis. Neurophysiological evidence for a role of posterior parietal cortex in redirecting visual attention. *Cerebral Cortex*, 5:448–456, 1995.

[SCCM94] M. A. Steinmetz, C. E. Connor, C. Constantinidis, and J. R. McLaughlin. Covert attention suppresses neuronal responses in area 7A of the posterior parietal cortex. *J. Neurophysiology*, 72:1020–1023, 1994.

[SJ91] J. Saarinen and B. Julesz. The speed of attentional shifts in the visual field. *Proc. Nat. Acad. Sci., USA*, 88:1812–1814, 1991.

[SS93] S. Shih and G. Sperling. Visual search, visual attention and feature-based stimulus selection. *Investigative Ophthalmology and Visual Science*, 34(4):1288, 1993.

[SW89] G.W. Strong and B.A. Whitehead. A solution to the tag-assignment problem for neural networks. *Beh. Brain Sci.*, 12:381–433, 1989.

[TBKV98] A. Toet, P. Bijl, F. L. Kooi, and J. M. Valenton. *A high-resolution image dataset for testing search and detection models (TNO-TM-98-A020).* TNO Human Factors Research Institute, Soesterberg, The Netherlands, 1998.

[TDW91] S. P. Tipper, J. Driver, and B. Weaver. Short report: object-centered inhibition or return of visual attention. *Quarterly Journal of Exp. Psychology*, 43A:289–298, 1991.

[TG80] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[TL93] Y. Tsal and N. Lavie. Location dominance in attending to color and shape. *Journal of Experimental Psychology: Human Perception and Performance*, 19(1):131–139, 1993.

[Tre88] A. Treisman. Features and objects: the fourteenth Bartlett memorial lecture. *Quart. J. Exp. Psychol.*, 40A:201–237, 1988.

[UN96] M. Usher and E. Niebur. A neural model for parallel, expectation-driven attention for objects. *J. Cognitive Neuroscience*, 8(3):305–321, 1996.

[WCF89] J.M. Wolfe, K.R. Cave, and S.L. Franzel. Guided search: an alternative to the feature integration model for visual search. *J. Exp. Psychology*, 15:419–433, 1989.

[WFHB94] J.M. Wolfe, S.R. Friedman-Hill, and A.B. Bilsky. Parallel processing of part-whole information in visual search tasks. *Perception & Psychophysics*, 55:537–550, 1994.

[WH97] D. G. Watson and G. W. Humphreys. Visual marking: Prioritizing selection for new objects by top-down attentional inhibition of old objects. *Psychological Review*, 104(1):90–122, 1997.

[Wol94] J.M. Wolfe. Guided search 2.0 – a revised model of visual search. *Psychonomics Bulletin & Review*, 1(2):202–238, 1994.

[Yar67] A.L. Yarbus. *Eye Movements and Vision.* Plenum Press, New York, 1967.

[YG89] A. L. Yuille and N. M. Grzywacz. A winner-take-all mechanism based on presynaptic inhibition feedback. *Neural Computation*, 2:334–344, 1989.