Visual Attention and Retinal Fixation: Preprocessing Modules for Enhanced Performance in Real-Time Vision

Arunava Banerjee
Department of Computer Science
Rutgers University
New Brunswick, NJ 08903
arunava@cs.rutgers.edu

Abstract

In systems that operate in real time, the manner in which resources are allocated often determines the success of a system or the failure thereof. The human visual system manages to operate in real time by repeating in order, processes of Fixation, Attention and Recognition. In essence, it chooses regions of interest in the visual scene and subsequently concentrates its computational resources on them. In this paper, we attempt to emulate this characteristic of the human visual system. We present operators for Visual Attention and Retinal Fixation, neural architectures that implement them, and simulation results that confirm the effectiveness of the underlying mechanisms. We also discuss the biological plausibility of our model in light of some of the available data on the functional characteristics of the visual system.

1 Introduction

Systematic study of human behavior has given us insights into numerous complex problems that humans solve almost effortlessly. An analogous approach to understanding the nature of vision is certain to yield results, given that our knowledge of the principles involved remains rudimentary.

In spite of the incessant impingement of massive amounts of data upon the retina, not only does the human visual system recognize objects with uncanny accuracy, but it also does so with considerable tolerance to noise. The visual system manages to operate in real time, as indicated, by choosing conspicuous regions in the visual space and allocating its computational resources to them. The process involving the selection of an appropriate region in the visual field and the subsequent centering of the retina on it, is called *Retinal Fixation*. The ensuing process of creating a window around the region and masking the remaining data is called *Visual Attention*. Attention essentially enhances the performance of the system by eliminating information that is not relevant to the task at hand.

The phenomena of Visual Attention and Fixation have been the subject of extensive psycho-physiological research for over two decades. Over the years, researchers have painstakingly compiled an enormous wealth of facts pertaining to the characteristics of these processes [Engel (1971); Eriksen, and Hoffman (1972); Beck, and Ambler (1973); Shulman, Remington, and McLean (1979); Bashinski, and Bacharach (1980); Posner, Snyder, and Davidson (1980); Treisman, and Gelade (1980); Julesz (1981); Bergen, and Julesz (1983); Sagi, and Julesz (1986); Downing (1988); Krose, and Julesz (1989)]. More recently, there have been a few proposals on neuro-biological models for Visual Attention [Koch, and Ullman (1985); Posner, and Petersen (1990); Crick, and Koch (1990a); Crick, and Koch (1990b); Olshausen, Andersen, and Van Essen (1993); Niebur, Koch, and Rosin (1993)]. The elaborate mechanisms and precise neural circuitry suggested by most models, however, render their physical realization problematic.

In this paper, we attempt to emulate the characteristic behavior of the human visual system with respect to Fixation and Attention. We propose relatively simple operators for each of these processes. The construction of these operators and their subsequent evaluation are guided by frequent references to some of the available data on the functional characteristics of the visual system.

We consider Attention to be a spatial filter and not a segmentation technique. Segmentation of any real world image requires extensive domain knowledge. Our operators, on the contrary, assume a minimal knowledge of the visual environment which we enumerate as we proceed. As a consequence, the region extracted does not correspond precisely to the image of the object. However, we assume that the error that originates as a result of this process is remedied by a robust, error tolerant recognition system.

The paper is structured as follows. In section 2, we take a fresh look at Vision. We examine the hierarchical nature of perceptual organization and define the task of vision with respect to it. In section 3, we introduce the concept of *visual information.* [1] This concept assists in the formal description of the constructs presented in the next section. In section 4, we present operators for Visual Attention and Fixation. The subsymbolic notion of visual information facilitates the design of architectures for these processes. Finally, in section 5 we discuss the adequacy of our model with respect to available psycho-physiological data.

2 The Task of Vision

Vision is one of five distinct sensory channels available to the human being. On the one hand, it lets the agent be affected by his environment, and on the other, it facilitates an active interaction between the agent and his envi-

[1] Our definition of Visual information has not been motivated by information theory and should not be confused as such.

ronment. Thus, the visual system performs two very distinct activities. We must, however, consider a very significant characteristic of the conceptual system before we can explicate these activities in detail.

It is evident that the cognitive organization of the visual world is *hierarchical*. That recognizable *objects* such as the eye, nose, mouth, etc., when arranged in a specific spatial layout are recognized as the *object* face, bears ample attestation to this fact. Any mention of a recognized object therefore, makes an implicit reference to a specified *level* in this hierarchy.

We also observe that this perceptual hierarchy is a closed structure. There exist numerous objects that can not be fragmented perceptually into constituent parts that are also perceived as objects. Those objects that are not split perceptually, we name *atomic* objects, or **A**-objects. A hockey puck, the seconds hand on a wrist watch, and the sun in the sky, are examples of such objects. To the other extreme lie objects that exist on their own accord; objects that are not perceived as constituent parts of other objects. Those objects that are not clustered into other objects, we label *complex* objects, or **C**-objects. Examples here range from a dining table to the sun. All other objects that reside at *intermediate* levels, we name **I**-objects. One must note that the level in the hierarchy that an object resides in is not determined by the object, but by the *perception* of it. An ant seen up close, is a C-object; a product of organizing many A-objects like the legs, abdomen etc. Seen from afar, however, the ant is both a C-object, and an A-object. [2]

We now state the two distinct activities that vision performs.

1. It *extracts* and *recognizes* **A**-objects as representatives of classes, clusters them into **I**-objects, and finally into **C**-objects. Furthermore, it retrieves certain spatial relationships between the **C**-objects and the agent, i.e., in which direction (angle), and in what relation to other **C**-objects in that direction (relative distance) does the object lie.

[2] A scheme that determines where the perception of an object resides in the hierarchy is given later on in this paper.

2. In addition to recognizing the object, it in certain cases, extracts explicit information about the *shape* of the object, so that the agent may manipulate himself to interact with it.

Note that in the first case, shape extraction is not mentioned. The rationale behind this is that while the shape information implicit in the image of an object may be exploited to *recognize* it, the task does not necessitate transforming this information into explicit shape. In this paper, we restrict our attention to the first activity mentioned above. We also assume that the input to the visual system is an array of fixed size, and that each element of the array represents the color of a unique pixel at a specific moment in time. [3]

One must note that the process of extracting the region in the visual field that pertains to a specific object is not equivalent to recognizing the object. Even though extraction as stated is essential,[4] recognition itself involves abstracting away a semantic entity that associates consistently with the agents knowledge of the world.

In what follows, we define two operators; one that characterizes objects at specific levels in the hierarchy, and another that determines the conspicuity of a point in the visual scene. Both operators are single step processes and apply on raw visual data. We also show that the application of the two operators in tandem amounts to being the front end of Visual *Attention*, and that a specific parameter in one of these operators determines its sweep.

3 Information in Objects

In the previous section, we presumed an informal understanding of entities called *objects*. We devote this section to formalizing the concept of an object

[3] In other words, we study monocular vision without object motion.

[4] The mere act of extraction bestows upon the object an independence that lies at the basis of its individuation.

and state how **A**-objects, **I**-objects, and **C**-objects may be differentiated.

The simplest object is a uniform, connected region with a closed boundary. This uniformity may be defined with respect to shading, texture, color, motion, or any other such feature. Given that we have restricted ourselves to monocular vision without motion, we choose to define uniformity solely with respect to color. One must, however, note that our techniques can be applied uniformly to all the other features. Moreover, physiological evidence against spatially distinct feature maps indicate that there exists a single map that reports a weighted integration of all features. Since a boundary is a narrow region with severe variations i.e., lack of uniformity, [5] the simplest object is a connected region with very low variations, bounded by high variations.

It is well documented that the human visual system recognizes most objects from their line-drawings. The answer to what makes this possible lies in our approach to understanding what a line-drawing is, and how it is different from the generic image of an object. Traditionally, line-drawings have been known to be a faithful replication of the important edges in the object, including of course, the defining boundary. One instantly arrives at a fallacy in this definition. Edges and boundaries are synthetic concepts. They are derived from the primary notion of variation. Whereas a visual system may not know where the important edges are, it can almost effortlessly extract the regions with high magnitudes of variation. A more accurate definition of a line-drawing is, therefore, a spatial *transformation* that preserves, in some form, the image in regions with high magnitudes of variation.

We can now see that the visual system's ability to discern objects from line drawings is because it perceives an object mostly in terms of regions of high variation. [6] In other words, variation is precisely the kind of information the visual system exploits in order to identify objects. We therefore equate variations to the *information* contained in the image of an object.

[5] From now on, any reference to variation should be read, " spatial variation with respect to color or some other feature ".

[6] Obviously then, not much is lost in the transformation.

Researchers during the past twenty five odd years have conducted numerous studies on how intensity variations in images may be detected. Through this has emerged a general agreement that the Marr-Hildreth edge detector adequately approximates the apparatus present in the visual system [Marr, and Hildereth (1980)]. [7] We use the Marr-Hildreth operator with two important changes. First, we relax all restrictions on the domain of use, i.e., we convert it into an abstract operator that can detect spatial variations with respect to any feature. Second, and more importantly, we use it as an *information detector* (defined below) instead of an edge detector.

The Marr-Hildreth operator, or in more abstract terms, the Laplacian of the gaussian (∇^2G) is presented in its radial form in equation (1).

$$\nabla^2 G(r,\sigma) = -1/\pi\sigma^4[1 - r^2/2\sigma^2]exp(-r^2/2\sigma^2) \tag{1}$$

We take the *magnitude* of the result of convoluting a feature(Q) at any point in space with the ∇^2G operator, and label it information($\mathcal{I}$). Assuming that $*$ is the convolution operator,

$$\mathcal{I}(Q,\sigma,x,y) = |\nabla^2 G * Q(x,y)| \tag{2}$$

Note that neither do we use the operator to detect zero-crossings, nor do we attach any purely semantic interpretations to $\mathcal{I}(Q,\sigma,x,y)$. On the contrary, we reserve any premier semantic interpretation for the complete object.

It is clear that $\mathcal{I}(Q,\sigma,x,y)$ in equation (2) also depends on the gaussian's standard deviation, σ. Whereas not interpreting $\mathcal{I}(Q,\sigma,x,y)$ helps us circumvent the traditional problem of spurious edges, we are now faced with a problem of spurious information. We shall come back to this problem later on in the paper. For the present, however, we assume a fixed σ.

In the context of this new framework, a simple object can be defined as a region with very low information($\mathcal{I}$), bounded by a region with high

[7] The various virtues of this isotropic operator, like spatial localization, frequency localization, and close proximity to the Wilson DoG discovered in the human visual system, may be found in most texts on vision.

information. This is because $\mathcal{I}(Q, \sigma, x, y)$ peaks just before and after an edge. In similar terms, a visual scene is a two dimensional spatial array with some (mostly non-uniform) distribution of information over it.

We now define the information content of a simple object(R), to be the total information contained in it.

$$\mathcal{I}(Q, \sigma, R) = \int\int_R \mathcal{I}(Q, \sigma, x, y) dx dy \tag{3}$$

We have already observed that almost all the information in a simple object is restricted to its boundary. By empirically evaluating the magnitude of information that generates a perceived boundary (lower bound), and noting the maximum possible value that information takes at a boundary with respect to a fixed σ (upper bound), we can confine the information in an unit length of boundary within an interval $[\eta(Q, \sigma) \pm \delta]$.

We define the *circumferential information density* of a simple object as

$$i(Q, \sigma, R) = \mathcal{I}(Q, \sigma, R)/Circ(R) \tag{4}$$

where Circ(R) is the total length of the defining boundary of the simple object. Consequently, for a simple object, we have,

$$i(Q, \sigma, R) \approx \eta(Q, \sigma) \tag{5}$$

We now see that an **A**-object is any object that is *perceived* as a simple object. In other words, it is any object that, because of the finite resolution of the system, is perceived as an object whose information lies primarily at the boundary.[8] For an **A**-object then, $i(R) \approx \eta$. On the other hand, **I**-objects and **C**-objects have $i(R) \gg \eta$, simply because information is also perceived within the object. *Circumferential information density* (i(R)), therefore, formally determines where a perceived object lies in the conceptual hierarchy.

[8]This is precisely why the object can not be disintegrated into constituent parts.

4 Fixation and Attention

Visual Fixation is the process wherein the eye [9] shifts and *fixates* on certain conspicuous objects in the visual field. It is a component of the greater process of Visual Attention that enables the agent to *attend* selectively to individual objects present in the visual field.

The factors that influence the process of selection have traditionally been divided into subject and object factors. Visual conspicuity, an object factor, has been the subject of numerous psychological studies. Through these studies has emerged a general notion of the saliency map. In this section, we formalize the notion of conspicuity, and present a precise mathematical formulation for the saliency map.

First, however, we address a problem that concerns the "well-formedness" of the desired construct. Assuming that the visual system has already chosen an object for Fixation, where exactly on the object does the eye fixate? In order to solve this problem, we invoke a subsidiary concept called the *center of information* of an object. Given the image of an object(R) in cartesian co-ordinates, we define the center of information $< \bar{x}, \bar{y} >$ to be,

$$\bar{x} = \frac{\iint_R x.\mathcal{I}(Q,\sigma,x,y)dxdy}{\iint_R \mathcal{I}(Q,\sigma,x,y)dxdy} \tag{6}$$

$$\bar{y} = \frac{\iint_R y.\mathcal{I}(Q,\sigma,x,y)dxdy}{\iint_R \mathcal{I}(Q,\sigma,x,y)dxdy} \tag{7}$$

The center of information of an object, like its analogous concept in mechanics, the center of mass, has numerous useful properties. It is a general, object centric point of reference, with respect to which operations such as scaling and rotation may be performed. We are therefore motivated to choose this as the point of Fixation. When we proceed to compute it, however, we run into problems. Note that the definite integrals in equations (6) and (7) are bounded by the object perimeter. At this early stage in the process, when the

[9] We shall refer to any generic input device as the "eye".

object boundary has not yet been retrieved, one is obviously not equipped with sufficient information to evaluate the equations. At a later point in this section, we present an alternative method that circumvents this problem. First, however, we enumerate certain general characteristics of conspicuous objects, and note how they translate onto our framework based on the notion of *information*.

Human attention is frequently drawn by bright objects in dull backgrounds. The size of the perceived object also influences its conspicuity. Given a visual field with two equally bright objects, an agent is predisposed to attend to the object that is large enough to be noticed, and at the same time is small enough to be recognized in one step. In essence, there seems to exist an optimal size that maximizes the conspicuity of an object. In our framework this amounts to the notion that regions with a *higher* average information density are more conspicuous than regions with a lower average. Note that $\mathcal{I}(Q,\sigma,R)/Area(R)$ increases as $Area(R)$ decreases, assuming that information accumulates at the boundary. At the same time, however, very small objects ($Area(R) \approx \sigma^2$) have a low information density because of a rapid drop in $\mathcal{I}(Q,\sigma,R)$. We now proceed to define the saliency of a *point* in the visual field, and show that not only does it grow with an increase in the average information density of the region in its close vicinity, but it also peaks at the center of information of isolated objects. The definition also accounts for the observation regarding optimal size. The saliency of a point (x,y) in a visual field is given by,

$$\mathcal{S}(Q,\sigma,\bar{\sigma},x,y) = \frac{(G(\bar{\sigma}) * \mathcal{I}(Q,\sigma,x,y))^2}{(r^2.G(\bar{\sigma})) * \mathcal{I}(Q,\sigma,x,y)} \tag{8}$$

where $*$ denotes the convolution operator, G is a gaussian filter with standard deviation $\bar{\sigma}$, and the operator $(r^2.G)$ is the result of multiplying G(r,$\bar{\sigma}$) with r^2.

We demonstrate that $\mathcal{S}(Q,\sigma,\bar{\sigma},x,y)$ does in fact behave as stated by considering three separate cases. Let w denote the perceived radius of the

image of an object(R) where, by perceived radius we mean the radius of the smallest circle that circumscribes the object.

1 $[w \ll \bar{\sigma}]$ Assuming that there are no other objects in the close vicinity of R, $\mathcal{S}(Q, \sigma, \bar{\sigma}, x, y)$ approximates to $(\pi\bar{\sigma}^2/2)(\sum_R \mathcal{I}(Q, \sigma, x, y))^2/(\sum_R r^2.\mathcal{I}(Q, \sigma, x, y))$ for any point in and around R. It follows that $(\sum_R r^2.\mathcal{I}(Q, \sigma, x, y))$ is minimized, and therefore, $\mathcal{S}(Q, \sigma, \bar{\sigma}, x, y)$ is maximized at the center of information of the object. This accounts for the eye fixating on the center of information of objects that are small enough to be recognized in one step. It also explains why the conspicuity of some small objects are affected by their immediate background.

2 $[w \gg \bar{\sigma}]$ Since the object happens to be much larger than the gaussian window, $\mathcal{S}(Q, \sigma, \bar{\sigma}, x, y)$ near the center of information remains low. It, however, peaks in the proximity of convex fragments of the boundaries of objects. This explains why human attention is drawn towards the boundaries of large objects.

3 $[w \approx \bar{\sigma}]$ This case has a qualitative propinquity to (2). Here, once again, the saliency peaks somewhere in between the center of information and the boundary of the object. In addition, the exact location of the peak in relation to the center and the boundary, depends on the information distribution over the object.

Note that $\mathcal{S}(Q, \sigma, \bar{\sigma}, x, y)$ is a product of two terms, $(G * \mathcal{I})$, and $\frac{(G*\mathcal{I})}{(r^2.G)*\mathcal{I}}$. Whereas the value of the second term depends entirely on the *distribution* of information, the value of the first term depends on the total amount of information that fits into the gaussian window. Assuming that information does collect at the object boundary, it follows that the conspicuity of an object rises if it fits into the gaussian window. This explains the observation regarding optimal size. We propose that the visual system chooses a local maxima of $\mathcal{S}$ for Fixation. Although the machinery involved in maneuvering

the eye is not of concern to us, we must point out that any such mechanism will be intimately associated with the saliency map.

Real world images of humans, a bug, and two islands are presented in figures 1(a), 2(a), and 3(a). $\bar{\sigma}$ was fixed at 5 pixel lengths, and the saliency maps for the images were computed using equation (8). Figures 1(b), 2(b) and 3(b) depict the corresponding contour maps. In figure 1(b), the peaks, in decreasing order of height, stand on the woman's face, on the man's face, and at various points on the vehicle. In figure 2(b), the peaks in the mentioned order, stand on each of the six legs and the two antennas (in pairs, on four of them), and on the head. In figure 3(b), the peaks stands on the temple, on the island that seats the temple, and on the other island. Other smaller peaks stand at various points on the sea. Note that the values chosen for σ and $\bar{\sigma}$ have a crucial bearing on the locations of these maxima. The behavior of the visual system faced with diverse scenes will be determined by the specific values chosen for these parameters.

Constructing a neural architecture for the stated operator is straightforward. There is evidence that ∇^2G is a close approximation of the DoG computed by the on-centric and off-centric ganglion cells in the retina [Marr, and Hildereth (1980)]. Given that the two kinds of cells are architecturally identical, and that their local distributions on the retina are the same, we propose the existence of one of each kind for every pixel. The sum of the activations of the two cells gives us the information $\mathcal{I}(Q, \sigma, x, y)$ for each pixel. $\mathcal{S}(Q, \sigma, \bar{\sigma}, x, y)$ can be retrieved similarly, using local, radially symmetric connections that model the gaussian filter(G) and the filter $(r^2.G)$. The proposed architecture assumes a basic global connection pattern, and is therefore biologically plausible.

We now proceed to address the last problem, i.e., Visual Attention at Fixation. Let us assume that through the process described above, a conspicuous object has already been chosen and that the eye is presently fixated

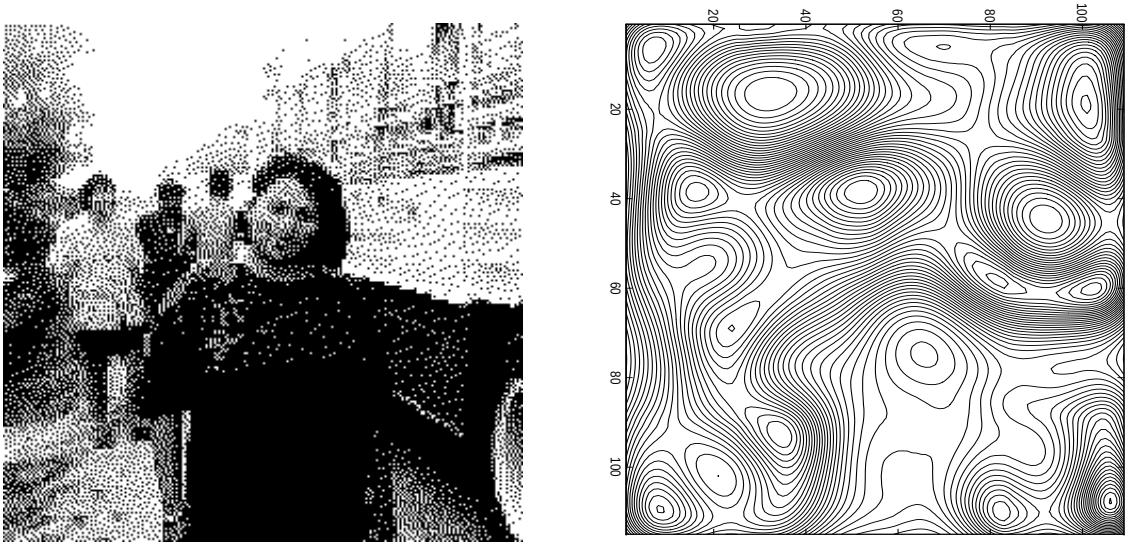

Figure 1: Image of human beings and the corresponding contour map for S.

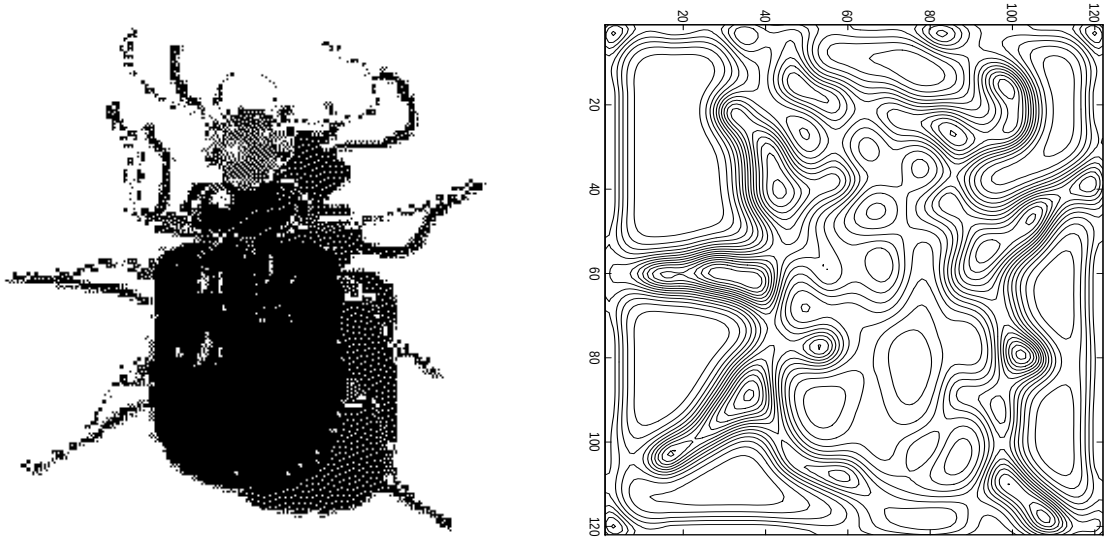

Figure 2: Image of a bug and the corresponding contour map for S.

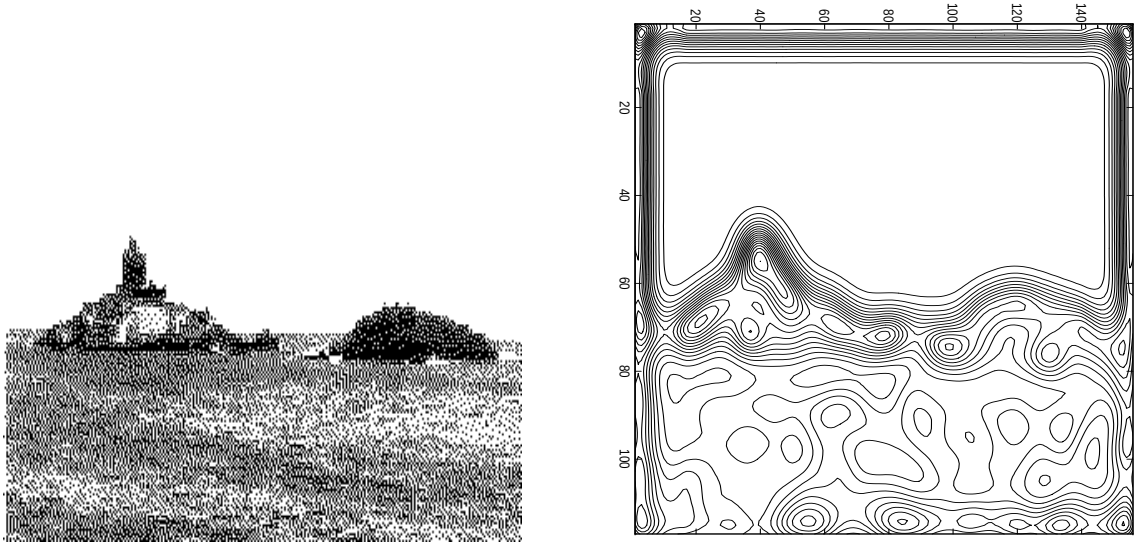

Figure 3: Image of islands and the corresponding contour map for S.

on its center of information. [10] The task is to extract from the image the **A**-object that corresponds to the peak in $\mathcal{S}(Q, \sigma, \bar{\sigma}, x, y)$, such that it may subsequently be recognized. Here, by extraction we mean a process that masks all information that does not pertain to the object.

We begin by imposing radial co-ordinates on the retina, with the center of the retina as the origin. Any point on the retina is then uniquely specified in terms of the pair $< r, \theta >$, where r represents the distance of the point from the center, and θ represents its angle with respect to the horizontal axis. Consequent to the process of Fixation, we have an object centered on the retina, i.e., the center of the retina is contained within the object perimeter. We now demonstrate how the critical constraint of equation (5) in section 3, may be used to extract the object from the scene.

[10] To be more specific, the eye fixates on a local maxima of $\mathcal{S}$(x,y). We present the case where the object is small enough to be recognized in one step. For the case wherein recognition takes more than one step, the sole difference lies in the subsequent phases of object recognition. This is to say that the analysis for the two cases are identical to the extent that they are examined here.

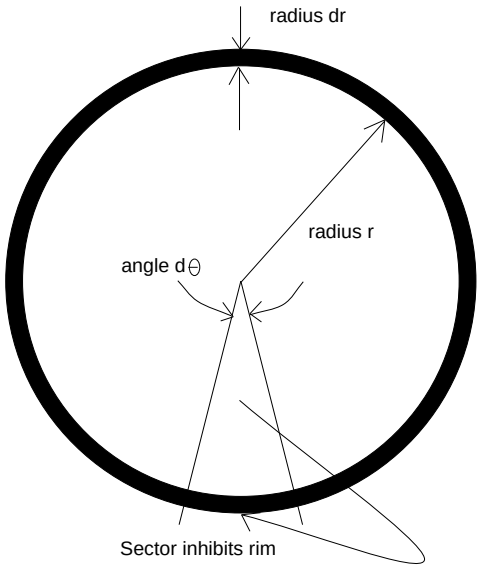

Figure 4: The inhibitory architecture.

As already noted, in the case of **A**-objects, information collects at the perimeter and is distributed rather sparsely within. Let us consider the following scheme in light of this information.

To every pixel is allocated a node that is intended to act as a gate. Each such node allows free passage of the data that originates from its corresponding pixel, unless it is inhibited to an excess of η, by a specific quantity $\mathcal{M}$. $\mathcal{M}$ is defined as follows.

Let Δ denote the node that corresponds to an arbitrary location $< \bar{r}, \bar{\theta} >$ on the retina. $\mathcal{M}$ for node Δ is defined as,

$$\mathcal{M}(\Delta) = \sum_{Z} \mathcal{I}(Q, \sigma, x, y) \tag{9}$$

where Z stands for the sector that has a radius $(\bar{r} - dr)$ and spans the angle $(\bar{\theta} \pm d\theta)$.

Constructing an architecture that computes $\mathcal{M}$ is straightforward. We overlay the layer that computes $\mathcal{I}(Q, \sigma, x, y)$ with an inhibitory connection

pattern that is radially symetric with respect to the center of the retina. The pattern is displayed in figure (4). The total information corresponding to the sector of radius r and spanning angle $d\theta$ is directed to all nodes on the rim $rdrd\theta$, for the purpose of inhibition.

Let us now consider two nodes; one just inside the boundary of the **A**-object, and one just outside. Whereas the node inside is inhibited by a signal of strength approximately equal to zero, the one just outside is inhibited by one of strength greater than η. Assuming that $d\theta$ is very small, this amounts to stating that the inhibitory architecture extracts precisely the **A**-object, and masks all other information.

Two points are noteworthy here. First, if one was willing to sacrifice retinotropy to economy on connection lengths, each sector would be reorganized into a hypercolumn that would exhibit an apparent orientation sensitivity. Second, the architecture proposed is not restricted to extracting **A**-objects. Both **I**, and **C**-objects could be extracted using values larger than η for the threshold.

Figures 5(a), 6(a) and 7(a) are magnified views of the region surrounding the peaks in $\mathcal{S}(x, y)$, corresponding to the face of the woman, the head of the bug, and the temple. Figures 5(b), 6(b), and 7(b) depict the results of applying the inhibitory architecture to the corresponding images.

It is crucial that we point out certain specific characteristics of the proposed model. First, the objects extracted are not all convex, as one would be led to believe, given the simplicity of the architecture.

Second, as indicated earlier, this is not a segmentation technique, and therefore, the regions extracted do not correspond precisely to the objects. However, precise extraction is not our objective. On the contrary, our goal is to *recognize* the object. We regard all the spurious data that is generated as a result of the imprecise extraction as a separate class of noise. We believe that the concept of an object is *inductively* learned over numerous extractions on images containing the object, and that this mechanism lends the visual system the fortitude to deal with this kind of noise.

Figure 5: Magnified view of the face and the corresponding masked image.

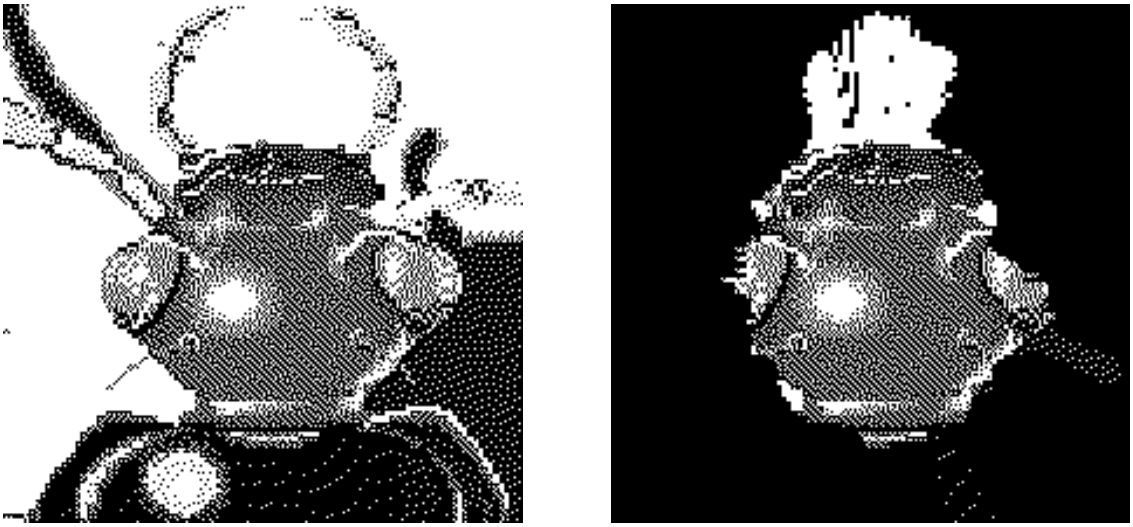

Figure 6: Magnified view of the head of the bug and the corresponding masked image.

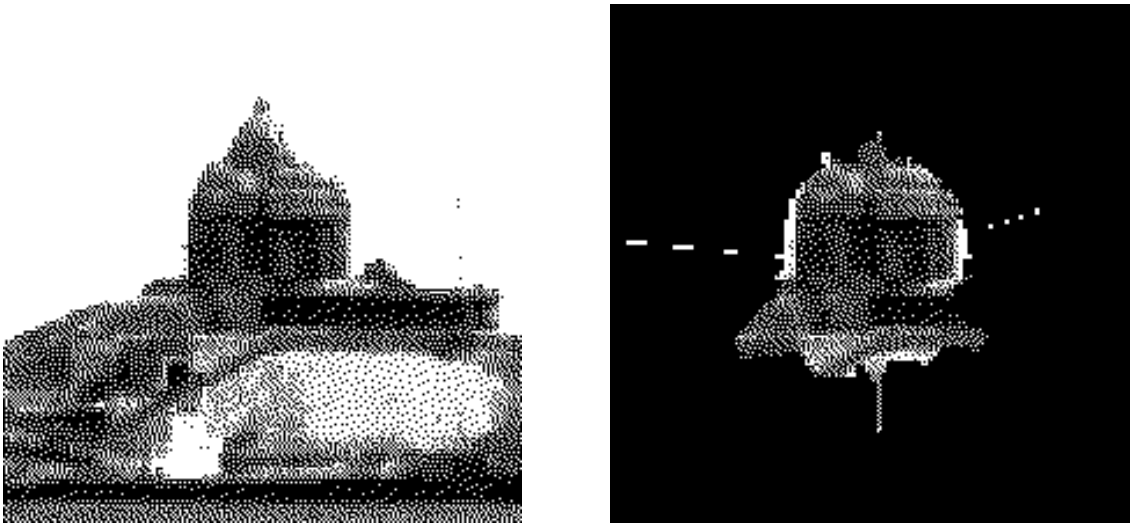

Figure 7: Magnified view of the temples and the corresponding masked image.

Third, one will observe that the objects extracted from the diverse scenes are not all **A**-objects. We managed to extract both **I**, and **C**-objects by using different values for the threshold.[11] This can, however, be achieved through a different procedure. Holding the threshold fixed at η, and modifying the focal length of the lens in the eye, yields similar results. Varying the focal length affects $\mathcal{I}(Q, \sigma, x, y)$, and in effect causes a *relative* modulation in the value of the threshold. The presence of four distinct spatial frequency channels [Wilson, and Bergen (1979)], along with a variable focal length, guarantees attention to the complete range of possible information distributions, with respect to the scale of variations in the image.

In the final section, we argue that the constructs suggested in this paper are tenable. We make the case that most of the available psychological data is consistent with our proposals. In addition, we note certain weaknesses in other models. We conclude the section with suggestions on future research.

[11]The values used were 35, 26, and 9 for the respective scenes.

5 Discussion and Future Research

In the previous sections, we formalized the notion of saliency and defined an operator that generates the saliency map. We also demonstrated how the phenomenon of Visual Attention may be realized through an inhibitory architecture. We now present some critical empirical findings that are in conformance with our model.

Our approach to dealing with Visual Attention and Retinal Fixation simultaneously is supported by results documented in [Engel (1971)]. The paper reports that even though the center of Attention and the center of Fixation may be situated at different points in space, the window of Attention always contains the point of Fixation. The nature of the window of Attention is, in other words, contingent on the results of Fixation.

Additional evidence in favor of our model comes from experiments conducted in [Desimone, Wessinger, Thomas, and Schneider (1989)]. The paper claims that the mechanism of Attention is closely linked to the oculomotor system. The recorded experiments indicate that affecting the oculomotor system influences the agent's capability to attend to certain regions in the visual field, but does not affect his ability to recognize objects. These results are consistent with our proposal. Note that our model requires that information from the saliency map and the oculomotor system (the mechanism that maneuvers the eye) be available to each other. Whereas the saliency map communicates to the oculomotor system the point of fixation, feedback information helps inhibit the current peak in $\mathcal{S}$ so that a different peak (interesting region) is chosen the next time. Affecting the oculomotor system affects the saliency map, and consequently influences the agent's ability to attend to certain regions in the visual space. The agent's ability to recognize objects is, however, not impaired, since Recognition is virtually independent of Fixation.

Experimental results reported in [Moran, and Desimone (1985)] are also consistent with our model. The authors of the paper state that they ob-

served a *contraction* of the receptive field around the attended stimulus. Figures 5(b), 6(b) and 7(b) demonstrate that our method has a characteristic proficiency in picking up, exclusively, the attended object, irrespective of its shape and size. The mask clearly wraps around the object in question.

Most other proposals on Visual Attention approach the problem independent of Retinal Fixation (Selective Routing [Koch, and Ullman (1985); Olshausen, Andersen, and Van Essen (1993)], Temporal Tagging [Crick, and Koch (1990b); Niebur, Koch, and Rosin (1993)]). We believe that our model is simpler, easier to implement into artificial vision systems, and is more biologically plausible on account of its assuming only global patterns.

Our reasons for using the $\nabla^2 G$ operator as an information detector instead of an edge detector are two-fold. First, extracting zero-crossings not only involves differentiating on-centric cells from off-centric cells, but also necessitates a precise connection pattern that links And-gates with these cells [Marr, and Hildereth (1980)]. We believe that this is biologically improbable. Second, the performance of this architecture deteriorates rapidly with an increase in the number of faults in the connection pattern. Information extraction is, however, tolerant to such faults.

We look forward to augmenting this research in two directions. There exist other filters in the literature that could be modified suitably to generate information detectors. Comparative studies could result in more accurate object extraction. There is also the problem of object recognition that we have not examined. Schemes that are appropriate to our system must be considered.

References

[1] Barlow, H. D. (1953) Summation and Inhibition in the Frogs Retina. Journal of Physiology, 119, pp 69-88.

[2] Bashinski, H. S., and Bacharach, V. R. (1980) Enhancement of Perceptual Sensitivity as a result of Selectively attending to Spatial Locations. Perception and Psychophysics, 28, pp 241-248.

[3] Beck, J., and Ambler, B. (1973) The Effects of Concentrated and Distributed Attention on Peripheral Accuity. Perception and Psychophysics, 14, pp 225-230.

[4] Bergen, J. R., and Julesz, B. (1983) Parallel versus Serial Processing in Rapid Pattern Discrimination. Nature(London), 303, pp 696-698.

[5] Colby, C. L. (1991) The Neuroanatomy and Neurophysiology of Attention. Journal of Child Neurology, 6, pp S90-S118.

[6] Crick, F., and Koch, C. (1990) Some Reflections on Visual Awareness. Cold Harbor Symposium on Quantitative Biology, LV, pp 953-962.

[7] Crick, F., and Koch, C. (1990) Towards a Neurobiological theory of Consciousness. Seminars in the Neurosciences, 2, pp 263-275.

[8] Desimone, R., Wessinger, M., Thomas, L., and Schneider, W. (1989) Effects of Deactivation of Lateral Pulvinar or Superior Colliculus on the Ability to Selectively Attend to Visual Stimulus. Society for Neuroscience Abstracts, 15, pp 162.

[9] Downing, C. J. (1988) Expectancy and Visual Spatial Attention: Effects on Perceptual Quality. Journal of Experimental Psychology: Human Perception and Performance, 14, pp 188-202.

[10] Driver, J., and Baylis, G. C. (1989) Movement and Visual Attention: The Spotlight Metaphor breaks down. Journal of Experimental Psychology: Human Perception and Performance, 15, pp 448-456.

[11] Engel, F. L. (1971) Visual Conspicuity, Directed Attention, and Retinal Locus. Vision Research, 11, pp 563-576.

[12] Eriksen, C. W., and Hoffman, J. E. (1972) Some Characteristics of Selective Attention in Visual Perception determined by Vocal Reaction time. Perception and Psychophysics, 11, pp 169-171.

[13] Hubel, D. H., and Weisel, T. N. (1962) Receptive fields, Binocular interaction, and Functional architecture in the Cat's Visual Cortex. Journal of Physiology, 160, pp 106-154.

[14] Hubel, D. H., and Weisel, T. N. (1968) Receptive fields, and Functional architecture of Monkey Striate Cortex. Journal of Physiology, 195, pp 215-243.

[15] Julesz, B. (1981) Textons, the Elements of Texture perception, and their Interactions. Nature(London), 290, pp 91-97.

[16] Koch, C., and Ullman, S. (1985) Shifts in Selective Visual Attention: Towards the underlying neural circuitry. Human Neurobiology, 4, pp 219-227.

[17] Krose, B. J. A., and Julesz, B. (1989) The Control and Speed of Shifts of Attention. Vision Research, 29, pp 1607-1619.

[18] Marr, D., and Hildereth, E. (1980) Theory of Edge Detection. Proc. R. Soc. London, B, 207, pp 187-217.

[19] McLeod, P., Driver, J., and Crisp, J. (1988) Visual Search for a conjunction of Movement and Form is parallel. Nature(London), 332, pp 154-155.

[20] Moran, J., and Desimone, R. (1985) Selective Attention gates Visual Processing in the Extrastriate Cortex. Science, 229, pp 782-784.

[21] Niebur, E., Koch, C., and Rosin, C. (1993) An Oscillation based model for Neuronal basis of Attention. Vision Research, 33, pp 2789-2802.

[22] Olshausen, B., Andersen, C., and Van Essen D. (1993) A Neural model of Visual Attention and Invariant Pattern Recognition. Journal of Neuroscience, 13, pp 4700-4719.

[23] Posner, M. I., Snyder, C. R. R., and Davidson, B. J. (1980) Attention and the Detection of Signals. Journal of Experimental Psychology: General, 109, pp 160-174.

[24] Posner, M. I., and Petersen, S. E. (1990) The Attention System of the Human Brain. Annual Review of Neuroscience, 13, pp 25-42.

[25] Sagi, D., and Julesz, B. (1986) Enhanced Detection in the Aperture of Focal Attention during simple Discrimination tasks. Nature(London), 321, pp 693-695.

[26] Shulman, G. L., Remington, R. W., and McLean, J. P. (1979) Moving Attention through Visual Space. Journal of Experimental Psychology: Human Perception and Performance, 5, pp 522-526.

[27] Treisman, A. M., and Gelade, G. (1980) A feature Integration theory of Attention. Cognitive Psychology, 12, pp 97-136.

[28] Wilson, H. R., and Bergen, J. R. (1979) A four mechanism model for Spatial Vision. Vision Research, 19, pp 19-32.