# INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

**The quality of this reproduction is dependent upon the quality of the copy submitted.** Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

# EVALUATION OF COMPUTATIONAL ATTENTION OPERATORS USING HUMAN IMAGE RECOGNITION

**Sandra Polifroni**

Department of Computer Science
McGill University, Montréal

A Thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfilment of the requirements for the degree of
Master of Science

© Sandra Polifroni, MM

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

*Your file Votre référence*

*Our file Notre référence*

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-70745-8

Canada

# ABSTRACT

This thesis presents a novel method of evaluating computational attention operators, which select locations of interest in an image, using a human image recognition task. Assuming that locations which are maximally interesting will be most useful for recognizing an image, it follows that a location selected by an attention operator will facilitate image recognition if it is of interest to a human. Since attention operators are increasingly being used to replace humans in vision tasks, it is relevant that their performance be compared to human vision.

Five different operators were evaluated. Human subjects were shown a series of black and white images in quick succession after which they were presented subimages extracted from the original image set as well as from other images. Subjects were asked to indicate whether they could recognize the subimages. The number of false-positives and true-positives associated with each operator provided information on the interest of the selected locations. Results show that the operators do not perform equally, with some selecting more recognizable image locations than others.

# RÉSUMÉ

Ce mémoire présente une méthode innovatrice pour l'évaluation d'opérateurs d'attention artificiels dont la fonction est de sélectionner les points d'intérêt d'une image. Si l'on suppose que les points les plus intéressants d'une image sont aussi les plus utiles pour identifier cette image, il s'ensuit qu'un point sélectionné par un opérateur d'attention artificiel facilitera la tâche d'identification si ce point provient d'une région d'intérêt pour un observateur humain. Les systèmes de vision artificielle utilisent de plus en plus les opérateurs d'attention pour remplacer les observateurs humains, il est donc important d'évaluer leur performance par rapport au système visuel humain.

L'évaluation de cinq opérateurs a été entreprise. L'évaluation s'est déroulée en demandant aux sujets humains de reconnaître des images noirs et blancs à partir de sous-images sélectionnées par les opérateurs; les performances des opérateurs sont mesurées en fonction du nombre et du type d'erreurs effectuées par les sujets. Les résultats démontrent que certains opérateurs sélectionnent de meilleurs point d'intérêt que d'autres.

# ACKNOWLEDGEMENTS

I thank my supervisors, Gregory Dudek and Frank Ferrie, for their guidance and support. I thank my parents for putting up with me during my thesis-induced dementia. I also thank the numerous people who helped me along the way: Ariel Tankus for supplying the code for the radial symmetry and convexity operators, helping me understand the operators and being an interesting pen-pal; Eric Bourque and Laurent Itti for providing me with the code for their interest operators; Tony Marley and Fred Kingdom, for helping me design my experiment and for answering my questions regarding psychology and psychophysics; Ron Rensink for answering my innumerable questions and supplying me with articles; Jeremy Wolfe and Daniel Simons, also for providing me with papers and citations and answering various questions. A special thank you goes out to all my friends who put up with my incessant mood swings and who prevented me, on many occasions, from chucking it all in and running off to join the circus. And, finally, a honourable mention goes to the lovely people at Lola Rosa cafe for making the fuel upon which I burned: coffee.

# TABLE OF CONTENTS

# LIST OF FIGURES

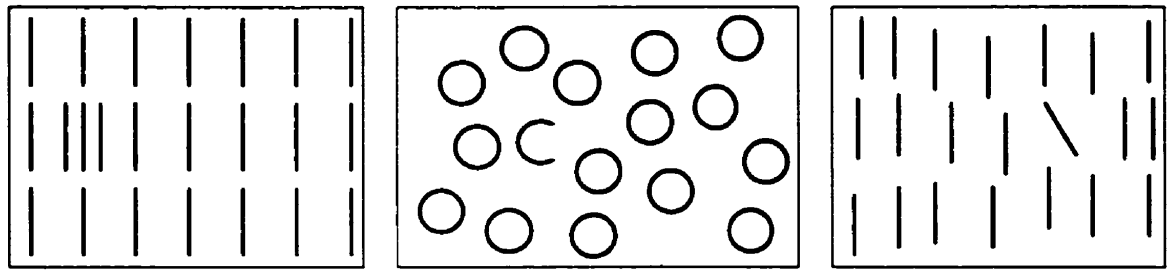
# CHAPTER 1

# Introduction

This thesis deals with attention in the context of both biological and computational vision. The particular form of attention considered is the selective observation of subparts of an image. This has been demonstrated to be crucial to human vision and is becoming a critical approach for machine vision. Several models of how to allocate attention have been proposed in the computational and biological literature. In this thesis, implementations of computational models of attention, refered to as computational attention operators, are evaluated using a new experimental paradigm based on a human image recognition task. To evaluate the operators, human subjects are asked to recognize whole images using only the locations selected by the operators. Assuming that humans will remember locations which are particularly interesting in an image[**20, 31**], subjects should have higher recognition rates if the operators select locations from interesting regions. Thus the purpose of this thesis is twofold: first to show that human image recognition *can* be used to evaluate computational operators and second to determine which operators best facilitate this task.

## 1.1. Computational Attention Operators

Computational attention operators use low-level image features to select locations of interest in an image. These operators are often based on features such as edge density, edge orientation and contour closure (Figure 1.1) which are presumed to be



(a) (b) (c)

FIGURE 1.1. Edge density (1.1(a)), contour closure (1.1(b)) and edge orientation (1.1(c)) are presumed to be selected by human visual system at a level prior to image understanding.

selected by the human visual system at a level prior to object recognition and image understanding. These features act as cues to the visual system to attend locations for further analysis[**29, 30, 31**], reducing the amount of the image to be analysed[**32**]. Because computational attention operators are designed to select image locations which would be of interest to humans, they can be used in many tasks which are normally reserved for humans such as the selection of images for virtual tours[**2**]. However, due to the nature of these tasks, it is essential that the locations selected by the operators correspond to locations humans would mark as interesting. In order to evaluate the operators using this criterion a new experimental paradigm based on human vision, human image recognition and human visual memory has been designed.

Experimental evidence implies that features selected by the human visual system at a low-level, prior to object selection and image understanding, are associated with areas of an image which are interesting to the viewer. In change blindness experiments, subjects are required to identify a change between an image and a modified copy of itself. It has been observed that less effort and time are needed to identify changes if they occur in regions containing features which would be used to describe an image (regions of central interest)[**20**]. This suggests that a location which is interesting at a low-level will also be interesting at a higher, context-dependent level. It should therefore be easier for humans to recognize an image which has been presented

for a short length of time using a location of central interest rather than a location which is not of central interest. The experimental paradigm described in this thesis is based on the assumption that an operator which performs well will select regions which are descriptive of an image and which facilitate image recognition.

## 1.2. Method

Evaluation of the operators consists of showing subjects several images, followed by subimages selected by computational interest operators and asking the subjects if they recognized the subimages. This method provides a means of measuring the representativeness of the image locations selected by the operators. If the location selected by an operator is representative of the image, a human subject will be more likely to recognize the original image from it. If, on the other hand, the location is *not* representative of the image, the subject would be more likely to incorrectly identify it. Thus by recording the number of correct and incorrect answers and by examining the types of errors committed by the subjects, an overview of operator performance can be determined.

If the results of the experiment show a clear, statistically significant result of one operator either performing exceptionally well or exceptionally poorly with respect to the other operators, it will be clear that this method of evaluation is sound and that certain operators show superior functionality than others.

## 1.3. Outline

This thesis will be separated into the following chapters:

**Chapter2:Background** will examine background work in human vision as well as computational models of attention and a description of the operators used in this research.

**Chapter3:Approach** will discuss the experimental methodology.

**Chapter4:Experimental Results** will present results as well as a discussion of the merits of each operator.

**Chapter5:Conclusions and Future Work** will recap the major findings of this thesis and present an overview of future work.

# CHAPTER 2

# Background

Discovering the locations to which one pays attention when initially viewing a scene, as well as what one recalls after viewing images[20] are some of the subjects studied in the field of attention. When an image is viewed, the human eye naturally fixates on certain locations, and some information from these locations is retained in memory. Psychophysical studies have attempted to discover why attention is focused on certain areas and what kind of information is retrieved by the viewer at a *preattentive* stage, prior to image understanding and object recognition. Computational attention operators can base their criteria for selecting locations of interest on these psychophysical studies. In this chapter, the psychophysical research on which some operators are based is described, followed by an overview of work on attention and a description of the operators used in this thesis.

## 2.1. Preattention and Attention

The human preattentive visual system operates at a level prior to image understanding and object recognition to direct visual attention to areas of interest for further processing[29]. Neurophysiological studies have shown that the human visual cortex, which is one of the parts of the brain responsible for visual processing, contains visual neurons which act as parallel filters, computing features, such as line segments, colour and luminance, represented in a 2-D retinotopic map. This suggests

that human preattentive vision is governed by searches for feature primitives which are quickly and easily recovered from a scene. In the 1980's psychophysical work by Treisman lent support to this theory by showing that certain feature primitives are preattentively extracted from a scene. Treisman asked human subjects to locate a feature target among feature distractors (the feature distractors being unlike the target). In some cases, the target could be located rapidly, in a time independent of the number of distractors, suggesting that the search was done in parallel, i.e. the whole scene was processed at once. This form of preattentive rapid visual search is referred to as *pop-out*[30](Figure 2.1). In cases where the target *did not* pop-out and search time was dependent on the number of distractors, it was postulated that search was performed serially, i.e. every object in the scene was examined sequentially and individually. Some of the features which exhibited pop-out were line orientation, edge density and colour. These features appear to correspond to the visual neurons in the visual cortex. Treisman found, however, that search for conjunctions of feature primitives, such as a search for a red **X** target among red **O** and green **X** distractors, was not conducted in parallel. Since shapes exhibit pop-out (**X** versus **O**) as do colours (red versus green), this lent further support to the notion that preattentive visual search is guided by feature primitives. If search is performed for individual feature primitives, then the presence or absence of these features will be signalled; combinations of these features would not be signalled preattentively since this would require that each item be attended individually to determine the presence or absence of the relative features.

An effect related to feature pop-out is *texture segregation.* Textures are said to segregate when the boundary between them is easily recognizable. In the 1970s, Julesz proposed a theory of texture segregation using *texture statistics*[13]. The term "texture statistic" refers to a local measurement of some statistical property of the image, such as edge density (a first-order statistic). Using grey level textures, Julesz found that textures could not segregate if they were identical in their first and second order statistics. Thus there had to be at least a difference in first order statistics for

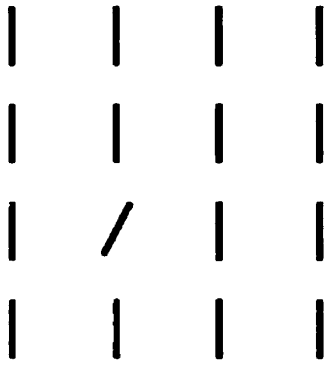


FIGURE 2.1. An example of feature pop-out. The slanted line is immediately discernible in the field of vertical lines. Individual elements need not be examined to identify the slanted line, making search for the slanted line independent of the number of vertical lines.

two textures to perceptually segregate since agreement in one order implied agreement in the orders below it. Treisman's studies on the discrimination of texture boundaries showed that if textures differ in certain primitive features, such as colour, curvature or line orientation, they will segregate[30]. Both the work of Treisman and Julesz suggest that the same features which are responsible for feature pop-out are also responsible for texture segregation (Figure 2.2).

A characteristic of texture segregation is *preattentive grouping.* Preattentive grouping refers to the tendency of the human visual system to group like objects which are not necessarily adjacent together in a rapid manner, independent of the number of objects in the scene. As more work was done on the subject of preattentive vision and rapid visual search, it was found that high level features not only exhibited preattentive grouping but also exhibited the same type of rapid visual search as feature primitives. Apparent convexity and concavity are examples of higher-level features which exhibit pop-out. Ramachandran showed that grey level circles shaded

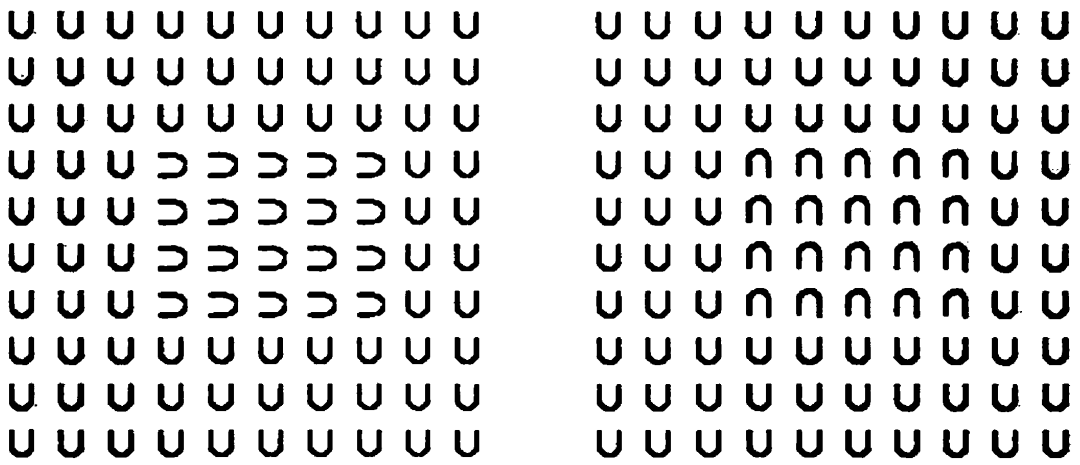
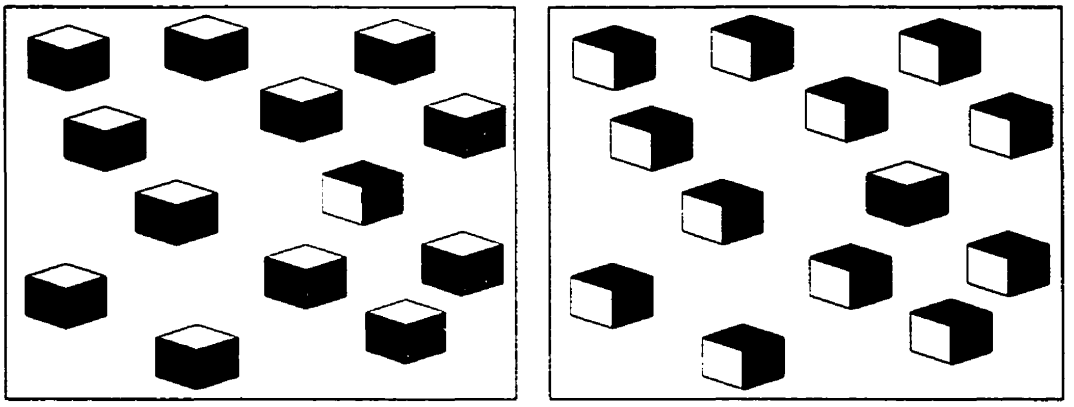


FIGURE 2.2. The textures on the left segregate more easily than the textures on the right. The textures on the left consist of two very different features: horizontal lines versus vertical lines. The textures on the right, however, are both made up of predominantly vertical lines.

in such a way as to appear concave or convex (assuming illumination from the top) will exhibit preattentive grouping[18]. Concavity and convexity are not considered feature primitives since the human visual system must extract shape and lighting information in order to extract the target from the distractors, thus requiring a higher level of processing. If the human visual system assumes that lighting comes from a single overhead light source, however, the task of extracting lighting information becomes trivial[18].

## 2.2. Complex Attentional Effects

The assumption of a single, overhead light source was used by Enns and Rensink to help explain the phenomenon of *search asymmetry*[8]. Search asymmetry refers to the effect of a target popping-out of a field of distractors, but the same not occurring when the roles of target and distractor are reversed. While search asymmetry is



(a) (b)

FIGURE 2.3. Enns and Rensink[8] found that a bottom-lit cube among a field of top-lit cubes (2.3(a)) will be easier to identify than a top-lit cube in a field of bottom-lit cubes (2.3(b)). Since the human brain generally assumes that a scene is illuminated from above by a single light source, a bottom-lit cube in a field of top-lit cubes will deviate from what the brain expects to see and will pop-out.

a very important phenomenon in human visual attention, very few computational attention operators take this or other more complex attentional effects into account. This can be a drawback since these complex attentional effects can greatly influence the locations which humans attend in an image.

In the 1990s, Enns and Rensink showed that a bottom-lit cube will pop-out from a field of top-lit cubes, however a top-lit cube did not pop-out from a field of bottom-lit cubes (Figure 2.3)[8]. Treisman had previously observed search asymmetry between slanted and vertical lines and attributed the effect to the target differing from the distractor in the *presence* of some feature primitive, supporting the neurophysiological evidence for feature detectors in the brain (Figure 2.4). Since the feature detectors only register the presence of a feature, if the target *lacked* a certain feature with respect to the distractor, then no activity would be recorded at that location[31]. The observed search asymmetry with lighting, though, supports a second postulate advanced by Treisman: that the features which do not pop-out in a case of search asymmetry are features which the visual system codes as "normal" and the features

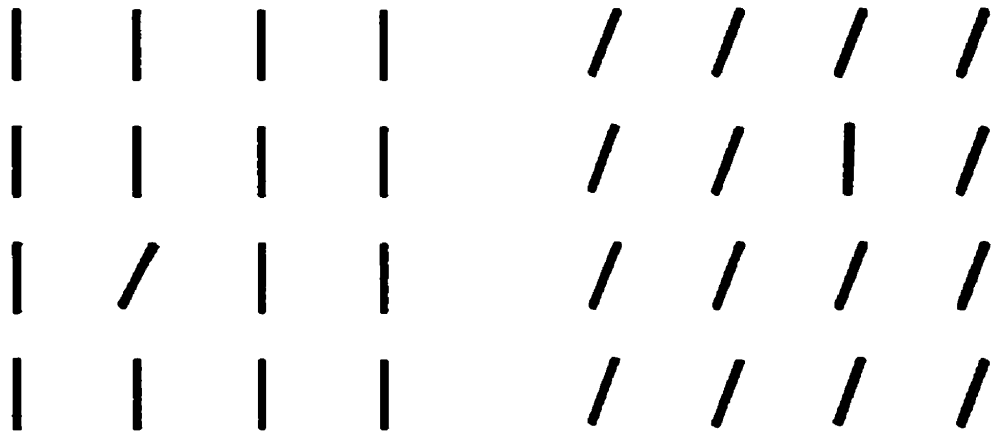


FIGURE 2.4. An example of search asymmetry. Search for the slanted line in the left-hand diagram is faster and easier than search for the vertical line in the diagram at the right.

which *do* are considered deviants from that normal. Since the brain generally assumes a single, overhead light source, as is the case in nature, a bottom-lit cube in a field of top-lit cubes will deviate from what the brain expects to see and will pop-out.

Enns and Rensink's work on lighting also showed that apparently three-dimensional information can be recovered preattentively[8]. Extracting three-dimensional information requires scene interpretation, suggesting that other factors aside from filtering for certain feature primitives influence preattentive rapid visual search. Further investigation revealed that feature primitives can be grouped together into wholes by the preattentive system. In work done in 1995, Enns and Rensink found that rapid visual search using the Mueller-Lyer configuration was guided by the overall length of the patterns rather than the length of the internal line segment (Figure 2.5)[9]. If rapid visual search is governed only by feature primitives detected by visual neurons in the visual cortex, search would be guided by the length of the central line segment. Thus, some type of high-level grouping mechanism must be used by the preattentive system. This mechanism is refered to as *preemptive grouping*[9], since it

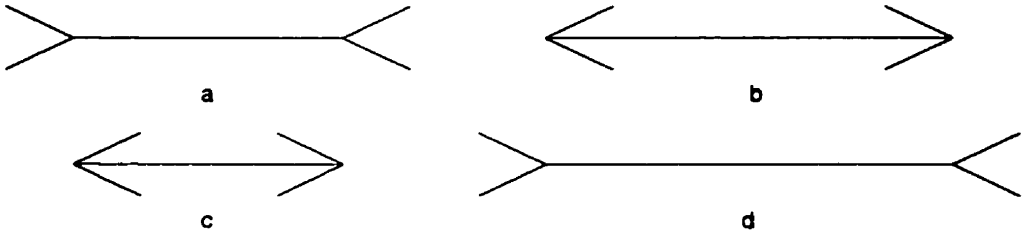



FIGURE 2.5. In the Mueller-Lyer configuration, the overall length of the figure is used in rapid visual search rather than the length of the inner segment. The inner segments of figures **a** and **c**, and figures **b** and **d** are equal, however it is figures **a** and **b** which are perceived as being of equal length (the *overall* length of the figures is equal). If rapid visual search were governed only by spatial filtering for feature primitives, this would not be the case.

preempts lower-level processes. Therefore the configuration of feature primitives can also govern human preattentive visual search.

Evidence suggests that locations which are selected by the human preattentive visual system to be attended for further processing are those locations which generate the most activity in several simple, filter-like processes typically associated with low-level vision. However, these locations might also be most interesting and contain the most information at a higher level. Since storing a complete, detailed representation of the world in the brain would be computationally intractable[**32**], only parts of the representation of the world can be stored. Furthermore, since the fovea, the most sensitive part of the retina, only subtends an angle the size of a thumb at arm's length, a complete, high-resolution representation of the visual world would be impossible[**3**]. By understanding the types of image locations the visual system attends to, the nature of human image understanding can be uncovered.

Rensink, O'Regan and Clark found that locations in an image which are attended are preferentially remembered[**20**]. Their hypothesis, which is similar to the claim made by Tsotsos, is that humans do not build a complete representation of the world around them and then modify this representation as changes are perceived, but rather only store in memory certain locations, most likely those upon which they have focused attention. Through the *flicker paradigm*, they showed that attention must be focused on a location for a change at that location to be perceived. The flicker

paradigm consists of showing two images, an original and a modified version, in quick alternation and asking subjects to indicate when they have perceived a change. The authors noticed that if the change was not located in a region of central interest of the image, the change was not immediately noticed by the subjects, even if the change was substantial. This effect is known as *change blindness*. Regions of central interest were defined as locations which were used by independent observers to describe the image[**20**]. In fact, the more peripheral the change was to the meaning of the image, the more time the subjects took to perceive the change. Thus, the complete image is not stored in visual memory, but only certain parts of the image, most likely those locations which were attended, are stored.

More recent studies by Simons show that change blindness can be induced without flicker. Simons showed subjects images which varied smoothly and asked them to indicate when they perceived a change. Again, the more peripheral the change was to the meaning of the image, the more time it took the subjects to perceive the change[**26**]. The suggestion that locations which are attended are more likely to be stored in visual memory can be used to evaluate computational attention operators which are designed to select locations of interest in an image. Since the locations which are attended by humans tend to be of a higher interest and since attended locations are more likely stored in memory, it follows that an image can be recognized by a subpart if the subpart contains a location of central interest. Using this assumption, a computational attention operator can be evaluated by asking humans to recognize an image based on a subpart containing the location selected by the operator.

## 2.3. Models of Attention

Computational attention operators can use different models to select locations of interest in an image. These models can be based on psychophysical studies, on neurological frameworks of human vision or on image structures. Psychophysically motivated models generally rely on one feature which has been shown to be preattentively selected by the human visual system. Computational attention operators

based on such models act as spatial filters for these features. Locations in the image are assigned interest values based on the degree to which they exhibit the feature used for selection (degree of convexity[28], degree of symmetry[19, 14, 23, 1]) or are based on the differences between feature characteristics at a particular location and the global feature characteristics (differences in edge orientation[2], differences in edge density[2], curvature variation[24]). Methods relying on image structures examine the features present in the image and assign values of interest based on certain properties of these features. These methods can be based on signal processing[34] and scale-space models[16, 11], among others. Neurophysiologically-based methods[10, 4, 33] attempt to model the mechanisms controlling human visual attention. One such model was postulated by Koch and Ullman[15] and will be discussed in this section.

In Koch and Ullman's model, separate feature maps are created for each elementary feature. These maps, which preserve spatial relations between locations, single out locations of high conspicuity. The feature maps are combined into a topological saliency map which contains global conspicuity information. The saliency map and the feature maps are refered to as the *early representation.* A *central representation* acting as a global feature detector also exists. The central representation does not conserve spatial relationships but only signals the presence or absence of certain features. Thus conjunctive search tasks cannot be done in parallel since the central representation can indicate the presence of two or more features but cannot indicate whether these features occurred in the same location. To perform a conjunctive search task, only information pertaining to the location needing analysis should be routed to the central representation. The saliency map provides information about which location is the most conspicuous and thus requires further analysis. The information from this location is then copied into the central representation.

The mechanism Koch and Ullman propose for the selection of salient locations is the *winner-take-all* (WTA) network. The WTA network is neurophysiologically-based and acts as a maximum-finding network. The WTA network consists of two pyramidal structures. The first pyramid finds the maximum while the second pyramid finds the

*location* of the maximum. The maximum is found by finding the maxima of small input sets of the saliency map; the "winners" of these sets are in turn compared to find the maxima among them. This process is repeated until a global maximum is found. The second pyramid, meanwhile, "marks" the path taken by the maximum in the first pyramid in order to find the location from where it originated in the saliency map. The features at this location are then copied into the central representation. Attention is then shifted to the next salient location, concentrating on different positions in the visual field in a serial manner. Thus the WTA network acts to direct an attention "spotlight" illuminating parts of the visual field to be analysed.

Shifting attention is modeled by inhibiting or decaying the signal of the attended location and making the WTA network respond to the new configuration. The signal is inhibited for a certain time period in order to prevent attention from being centered at the same location repeatedly. Proximity and similarity conditions are also imposed to guide search for a new location to attend. The proximity condition is implemented by enhancing the signal of locations which are in the neighbourhood of the attended location, while the similarity condition is modeled by enhancing the signal of locations which exhibit the same features as the attended location.

Computational attention operators which are motivated by neurophysiological models such as that of Koch and Ullman differ from the psychophysical models of attention in the way they select interest locations. Models motivated by psychophysics will typically assign interest values to locations in an image based on the degree to which they exhibit the feature used by the operator. Generally, no cohesive model for attention or attentional shifts is proposed by these operators, and the selection process is driven by a spatial filtering for image characteristics rather than by a simulation of the underlying neural processes of human visual attention.

## 2.4. Computational Attention Operators

Several interest operators exist which attempt to model certain aspects of human preattentive vision. Most of these operators concentrate on analysing one or two

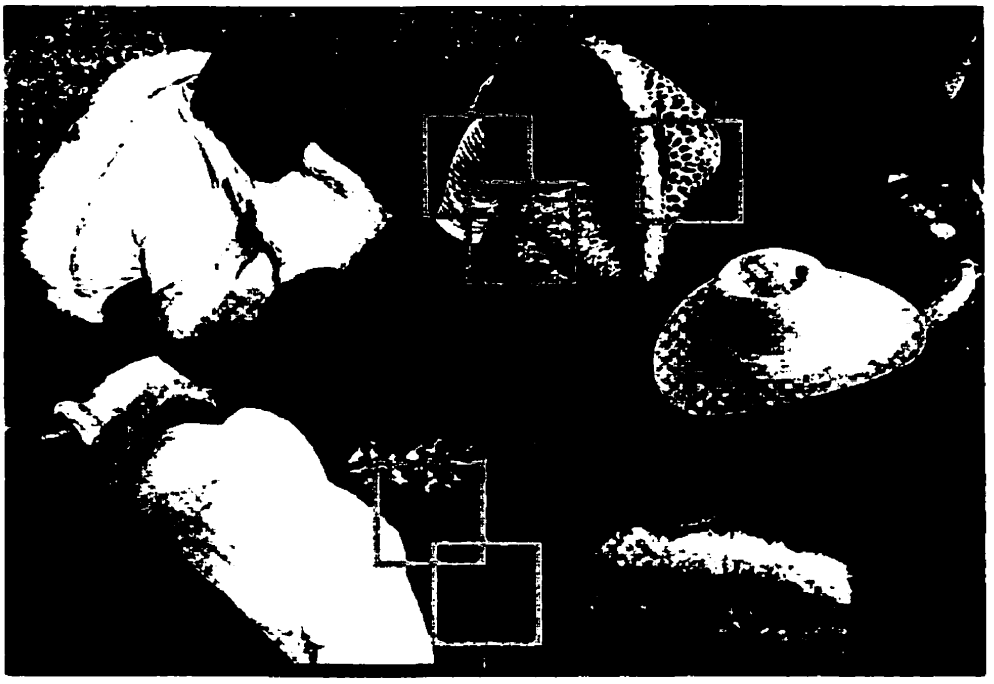


FIGURE 2.6. Top five selections of the edge density operator.

features which have been shown in the psychophysical literature to be preattentively selected by the human visual system. In this section, the operators used for this research will be described. The operators are an edge density operator[**2**], an edge orientation operator[**2**], a convexity operator[**28**], a radial symmetry operator[**19**] and a combination luminance/edge orientation/colour operator[**10**] (for the sake of brevity, this last operator will be refered to as the Caltech operator for the remainder of this thesis).

**2.4.1. Edge Density Operator[2].** The edge density operator was developed along with the edge orientation operator for a robotic mapping task involving the assembly of images from an environment to be used in a virtual tour. This operator is motivated by work by Treisman which showed that edge density is one of the feature primitives preattentively selected by human vision[**31**]. The operator works by selecting regions in the image which deviate the most from the mean density of the whole image. An edge map is created where each element is assigned an intensity

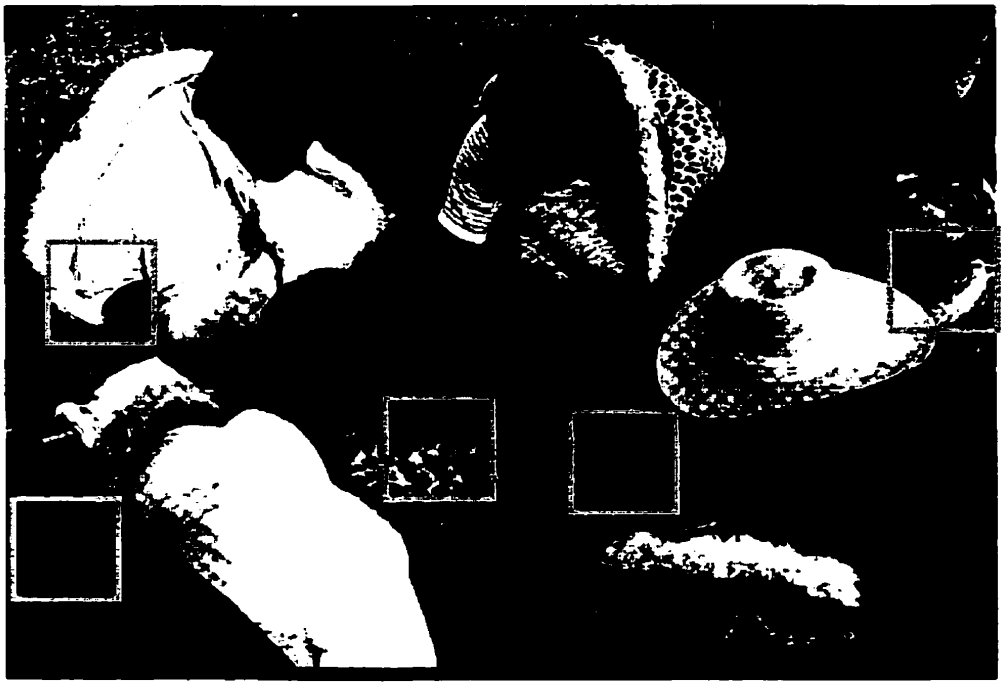


FIGURE 2.7. Top five selections of the edge orientation operator.

corresponding to the strength of its associated edge. This edge map is convolved with a Gaussian windowing operator to give an edge density map. The location of maximum interest is then defined as the location where the local density varies maximally from the mean density over the whole image (see Figure 2.6).

**2.4.2. Edge Orientation Operator[2].** This operator, which is also motivated by observations of human performance[**31**], selects regions of interest based on their deviation from the most prevalent orientation of the image. Using an edge map of the image, an orientation map of the edge elements is constructed. For each neighbourhood in the image, the number of edge elements for a particular orientation is computed and the maximum of the smoothed distribution for each of these neighbourhoods is found.

Given a function $\Phi(k, i, j)$, $k \in [0, \pi)$, which returns the number of edge elements with orientation $k$ in the neighbourhood of $(i, j)$, the most prevalent orientation in the neighbourhood of $(i, j)$ is defined as:

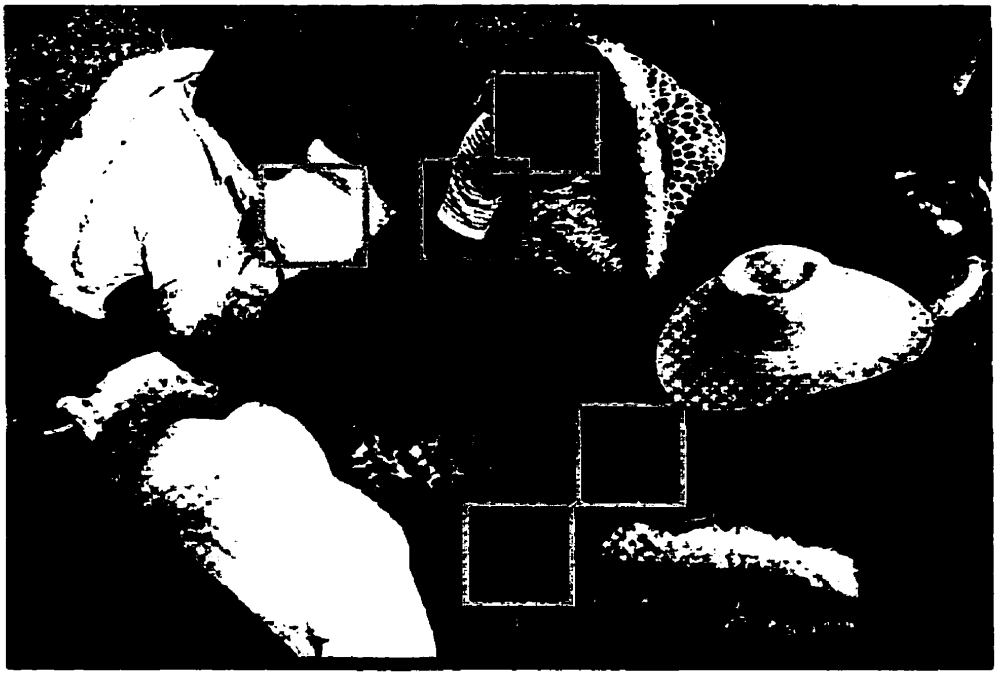


FIGURE 2.8. Top five selections of the radial symmetry operator.

$$O(i,j) = \max_{q \in [0,\pi)} \int_{q-\frac{\omega}{2}}^{q+\frac{\omega}{2}} \Phi^*(k,i,j)dk \quad \omega \in (0, \frac{\pi}{2})$$

where

$$\Phi^*(k,i,j) = \Phi(k \bmod \pi, i, j) \quad k \in \mathrm{R}$$

and $\omega$ is the subsection of the orientation distribution being considered. The neighbourhood whose maximum orientation has the greatest deviation from the robust maximum over the whole image is considered the most interesting (see Figure 2.7).

**2.4.3. Radial Symmetry Operator[19].** This operator was developed as a low-level mechanism for guiding gaze control in an active vision system. Whereas the orientation and density operators are based on edges, the radial symmetry operator is based on intensity gradients of the image. The radial symmetry operator does not attempt to select regions which are perfectly symmetric, but attempts to locate

regions which are approximately symmetric. Furthermore, since it is not edge-based, even symmetric regions which do not have smooth, uninterrupted contours will be selected.

For a point $p$ in the image intensity map, the gradient of the intensity at that point is denoted by

$$\nabla p = (\frac{\partial p}{\partial x}, \frac{\partial p}{\partial y}).$$

A magnitude term, $r$, and phase term, $\theta$ for the gradient at $p$ are defined as

$$r = \log(1 + \|\nabla p\|)$$

$$\theta = \arctan(\frac{\partial p}{\partial y} / \frac{\partial p}{\partial x}).$$

The angle $\alpha_{ij}$ denotes the angle that the line through points $p_i$ and $p_j$ makes with the horizontal. For any point $p$ in the image, radial symmetry contributions come from pairs of points which have $p$ as a midpoint. The symmetry contribution made by each pair is the product of the magnitudes of the intensities at each point weighted by a distance term, $D_\sigma$, and a phase term, $P$:

$$D_\sigma(i,j) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\|p_i - p_j\|}{2\sigma}}$$

$$P(i,j) = [1 - \cos(\theta_i + \theta_j - 2\alpha_{ij})][1 - \cos(\theta_i - \theta_j)].$$

The direction of the radial symmetry contribution is defined as the average of the phases of the two contributing points. The radial symmetry for any point is the sum of all contributions weighted by the difference between the phase of the point and the phase where contribution is the greatest. Locations of interest are defined as points with the greatest radial symmetry (see Figure 2.8).

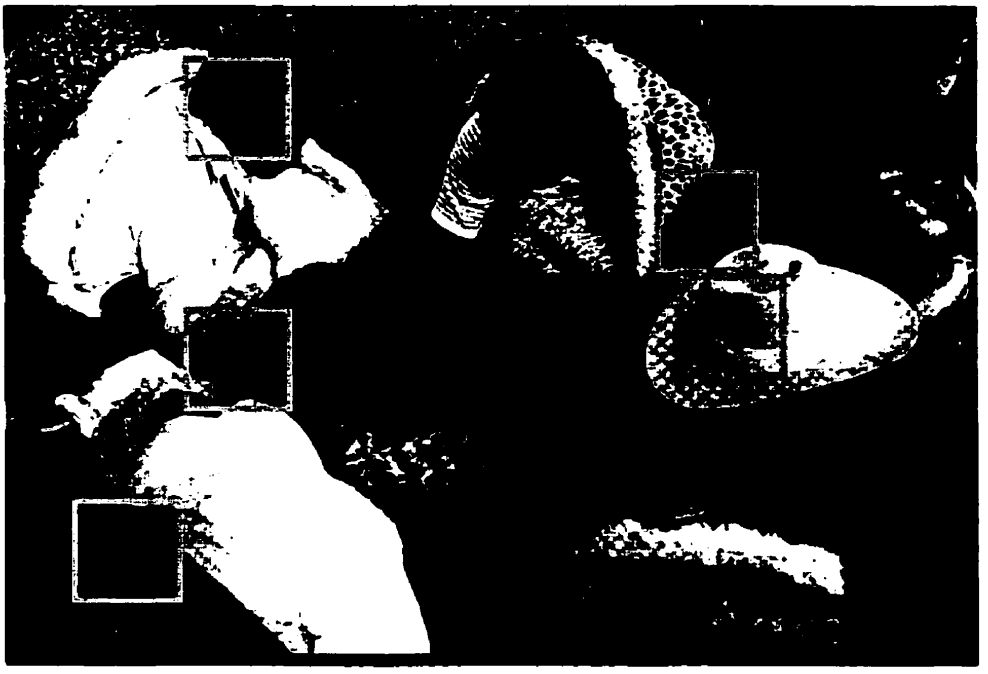


FIGURE 2.9. Top five selections of the convexity operator.

**2.4.4. Convexity Operator[28].** The convexity operator was designed to cut through noisy images and camouflaged regions by responding to three-dimensional convex or concave regions. As with the radial symmetry operator, the convexity operator is based on the intensity gradient map of an image. The operator searches for zero crossings of the phase of each point in the $y$ direction, using the derivative of the phase in the $y$-direction:

$$\frac{\partial \theta(x,y)}{\partial y} \approx [G_\sigma(x) D_\sigma(y)] * \theta(x,y)$$

where $G_\sigma(t)$ is the one-dimensional Gaussian with zero mean and standard deviation $\sigma$, $D_\sigma(t)$ is the derivative of the Gaussian and $\theta(x,y)$ is the phase at point $(x,y)$. The derivative of a zero-crossing will tend to infinity and thus produce a large response in the operator. In order for the operator to be isotropic, the image is rotated and the derivatives of the phase taken at these different orientations are summed

FIGURE 2.10. Top five selections of the Caltech operator.

together. The locations with largest response are those with the greatest interest (see Figure 2.9).

**2.4.5. Caltech Operator[10].** This operator is based on the winner-take-all model of attention by Koch and Ullman[**15**] (see Section 2.3). This operator uses multi-scale saliency maps and three interest features: colour, luminance contrast and orientation. Features are extracted using a center-surround method implemented as the difference between fine and coarse scale responses. The feature maps created using the center-surround method are fed into conspicuity maps. The colour features are extracted using red-green and blue-yellow center surrounds based on colour pop-out. Four Gaussian pyramids are created: one each for red surround and green center, green surround and red center, blue surround and yellow center and yellow surround and blue center. The luminance contrast map is made using dark center and light

surround as well as light center and dark surround. Gabor pyramids[1] are used to find locations of orientation contrast between the center and the surround.

The three conspicuity maps are then normalized and summed to give the saliency map, which is then fed through a winner-take-all network of inhibition and return so that the global salient points can be located without selecting the same points twice (see Figure 2.10).

## 2.5. Random Subimages

In addition to the computational attention operators which have been described, a "random operator" was used as a control. The random operator simply selects random image coordinates[2] and uses these as the centers of subimages. The random subimages were used as a control in order to observe whether human subjects would have significantly better image recognition using selections from computational operators than using random subimages. If recognition scores were not significantly better with the computational operators it would imply that the locations returned by the operators were no better at describing the image than random image regions.

In the next chapter, we show how these operators are evaluated using human image recognition.

[1] A Gabor filter consists of a sinusoidal grating with a Gaussian envelope.
[2] The random numbers are generated using a perl package (Math::Random) which returns integers which are deviates from a uniform distribution.

# CHAPTER 3

# Approach

Computational attention operators can be used for a variety of different tasks which would typically be performed by humans. A computational operator must therefore select regions in an image which would be interesting to a human observer. In this chapter, a method of evaluating operators with respect to human vision is described. The method is motivated by evidence from change blindness and preattentive vision suggesting that locations which are interesting to a viewer at a low level will also be interesting at higher levels. This chapter also includes a description of a method for extracting interest points from an image using an interest map.

## 3.1. Uses for Computational Attention Operators

Computational attention operators can be used to extract features for a variety of applications such as object recognition[**17, 12, 6**], image matching[**21**] and pose estimation[**25**]. The scope of this thesis, however, centers on the applications of interest locations in tasks which would ideally be performed by humans. There exist computational methods for evaluating attention operators, such as the method proposed by Schmid, Mohr and Bauckhage based on the information content of the operator's selections as well as on the repeatability of operator selections under image transformations[**22**]; however, the different applications for computational attention

operators, which will be outlined next, demonstrate why it is important that the operators be evaluated in terms of human image recognition.

**3.1.1. Navigation and Exploration.** Computational attention operators can be used for both navigation and exploration tasks. In a navigation task, visual data and attention operators can be used to select landmarks a robot can use for localization in an environment[**25**]. In exploration tasks, attention operators can be used to suggest locations for further investigation. The suggestions returned by the operators should be locations which would be interesting to humans since robots are surrogates for humans in places where it is either too costly, too dangerous or too impractical to send a person. Therefore it is important that the attention operators used return locations which are interesting to humans.

**3.1.2. Image Identification.** When there are many images in a database, it is not unusual that they are represented by thumbnails for the convenience of users. However, it is often the case that thumbnails are of very low quality and cannot be easily recognized. A different approach to image identification would be to use computational attention operators to select representative subimages of the images. The subimages would be of the same quality and resolution as the image but would only be a fraction of the size. For this method of image identification to be practical, however, the operators must return image locations of central interest to the viewer.

**3.1.3. The Vacation Snapshot Problem.** The *vacation snapshot problem*[**2**] refers to the problem of selecting interesting images from a set of images returned by a mobile robot. The problem gets its name from tourists who, while on vacation, must decide which people and places to record with their camera. Solving the vacation snapshot problem is important in various applications such as creating virtual tours of environments[**2**]; a virtual tour should contain locations and landmarks which would be interesting to viewers. This process could be automated by using a computational attention operator to select locations and images for the tour. Similarly, a computational attention operator could reduce the amount of image data

returned by a robot on a general exploration task by selecting images which are interesting. For either of these implementations to be successful, the operators would have to return images and image locations which would be interesting to humans.

## 3.2. Computational Attention Operators and Human Vision

Human preattentive vision, as seen in Chapter 2, seeks out locations in an image which are most interesting at a low level to be attended for further analysis. The level of interest is based on levels of unusualness, as determined by the assumptions made by the brain. Psychophysical studies have shown that an attended image location is more likely to be retained in memory than an unattended location[**20**]. In fact, change blindness experiments have demonstrated that a change in an image can be perceived more easily if attention is focused on the location of the change[**20**]. In other words, an image location is more likely to be perceived if it is attended.

Change blindness experiments have shown that changes which are of central interest in an image are most easily noticed than other changes(see Chapter 2). Since it is the preattentive visual system which selects locations to be attended, this implies that the preattentive visual system selects regions which will also be interesting at a higher level. Computational attention operators exploit this theory by selecting salient image locations based on low-level features such as edge density[**2**] and luminance[**10**].

Change blindness experiments suggest that regions of central interest in an image are more likely retained in memory. Since the locations where changes are most easily perceived are those which have been attended, it follows that these locations have been stored in memory. If attending locations facilitates the viewer's perception and memory of the location, and if these locations are those which are of central interest, it follows that an image should be easily recognized using only a subregion containing a location of central interest. Since regions of central interest are selected by the human preattentive visual system and since computational attention operators use low-level image features to select salient image locations, it follows that an attention

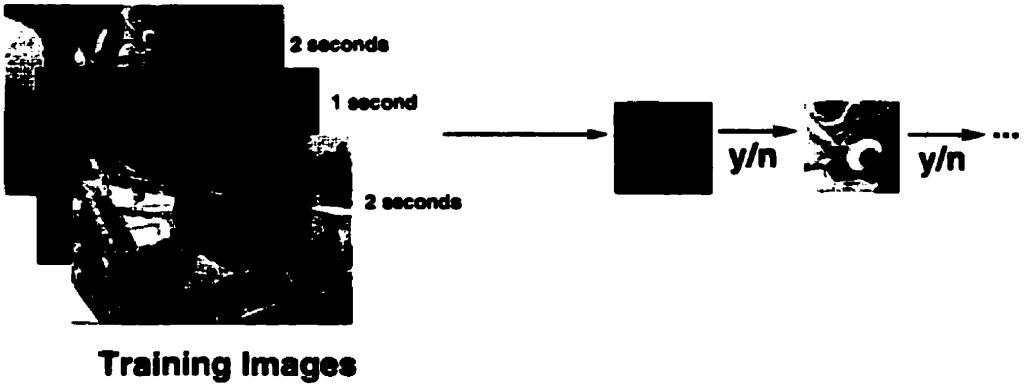



FIGURE 3.1. The experiment consisted of showing subjects 20 training images for a duration of two seconds each, with an interstimulus interval of one second. After all 20 training images were presented, subimages were shown to the subjects. The subjects' task was to say whether they could or could not recognize the subimage as originating from the training set. A new subimage would be presented only after the subject had responded.

operator which performs well will choose locations of central interest to the image. Thus testing the ability of humans to recognize an image based only on a subregion selected by the operators is good way of rating attention operators. The development of a new experimental paradigm for evaluating computational attention operators is the central idea of this thesis.

## 3.3. Methodology

Psychophysical evidence suggests that recognizing images from subregions would be facilitated if the subregion contains a location of central interest more than if it does not[**20**]. In this manner a human subject would have a greater chance of recognizing an image from a location selected by an attention operator if the location had been attended by the subject during initial viewing. This thesis uses this hypothesis to evaluate attention operators.

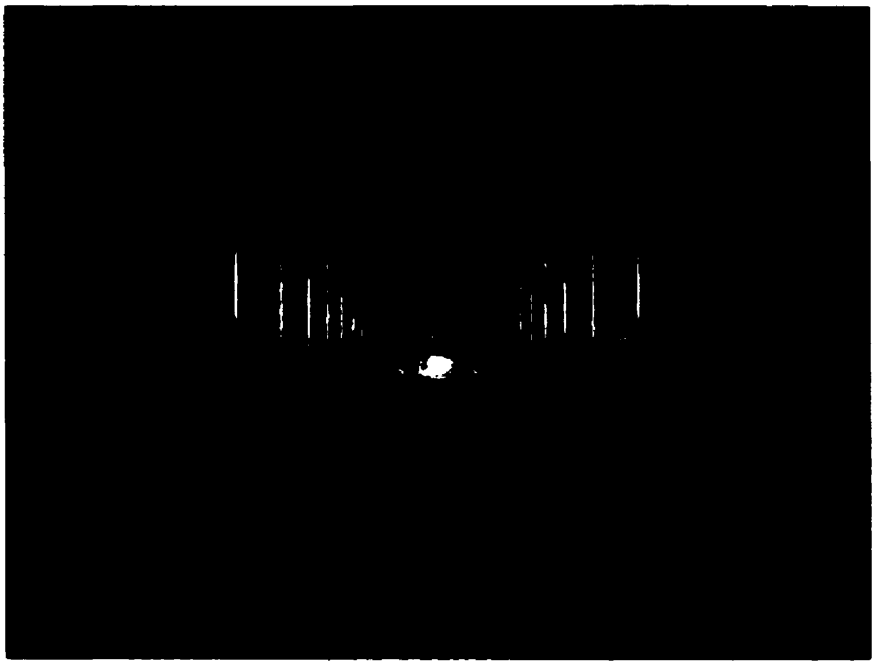
Seven naive subjects between the ages of 20 and 50, with university-level education, were shown 20 black and white, 700x500 pixel resolution images taken randomly from a 400-image database. Each image was shown for a duration of 2 seconds, with an inter-stimulus interval of 1 second. After the 20 training images had been shown,

the subjects were shown a series of subimages of resolution 75x75 pixels and were asked to say whether or not they recognized these as originating from the image set (Figure 3.1). Images and subimages were displayed on a uniform grey background using a 17 inch PS790 ViewSonic colour monitor. Viewing distance was approximately 40cm (Figure 3.2). The large images subtended and angle of approximately $28^o$ of visual angle while the subimages subtended an angle of approximately $3^o$ of visual angle. Response was in the form of a forced-choice yes/no answer signalled by a mouse-button press. Subjects were instructed to press the left mouse button (YES) if they recognized the subimage as originating from the training set and to press the right mouse button (NO) if they did not recognize the subimage. No time limit was placed on response, nor were response times recorded since the purpose of the experiment was one of general recognition. Each subimage was displayed individually and the next subimage was not displayed until a response was provided (see Figure 3.1). After every sixty subimages, there was a two minute break followed by a redisplay of the original twenty images.

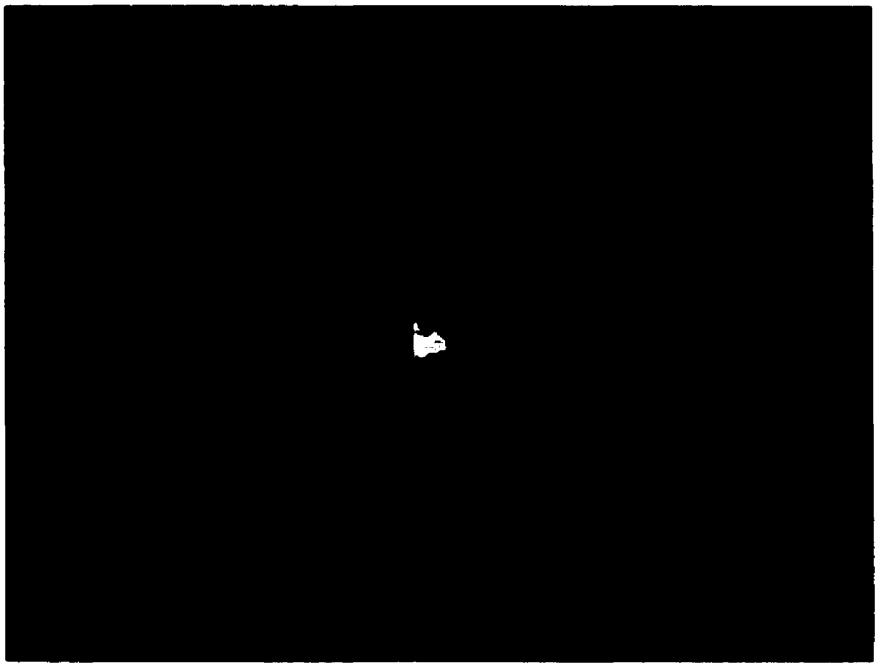
The subimages shown could originate from the twenty training images or from images in the rest of the database. In total, 180 subimages were shown; 90 originating from the original set and 90 originating from remaining images in the database. An equal number of subimages was extracted using each of the computational attention operators described in Section 2.4 and the random operator described in Section 2.5. Thus a total of 30 subimages were extracted using each operator. For 15 of the 20 original images, one subimage was extracted using each operator, for a total of 6 subimages per image. The subimages used were the first choice of the locations selected by the attention operators.

## 3.4. Images and Subimage Extraction

A database of 400 images was compiled for the experiment, some of which are shown in Figures 3.3, 3.4 and 3.5. The images were black and white (stored in pgm – portable greymap – format) with resolutions of 700x500 pixels. The black and white



(a) Screenshot of the display of a training image



(b) Screenshot of the display of a subimage

FIGURE 3.2

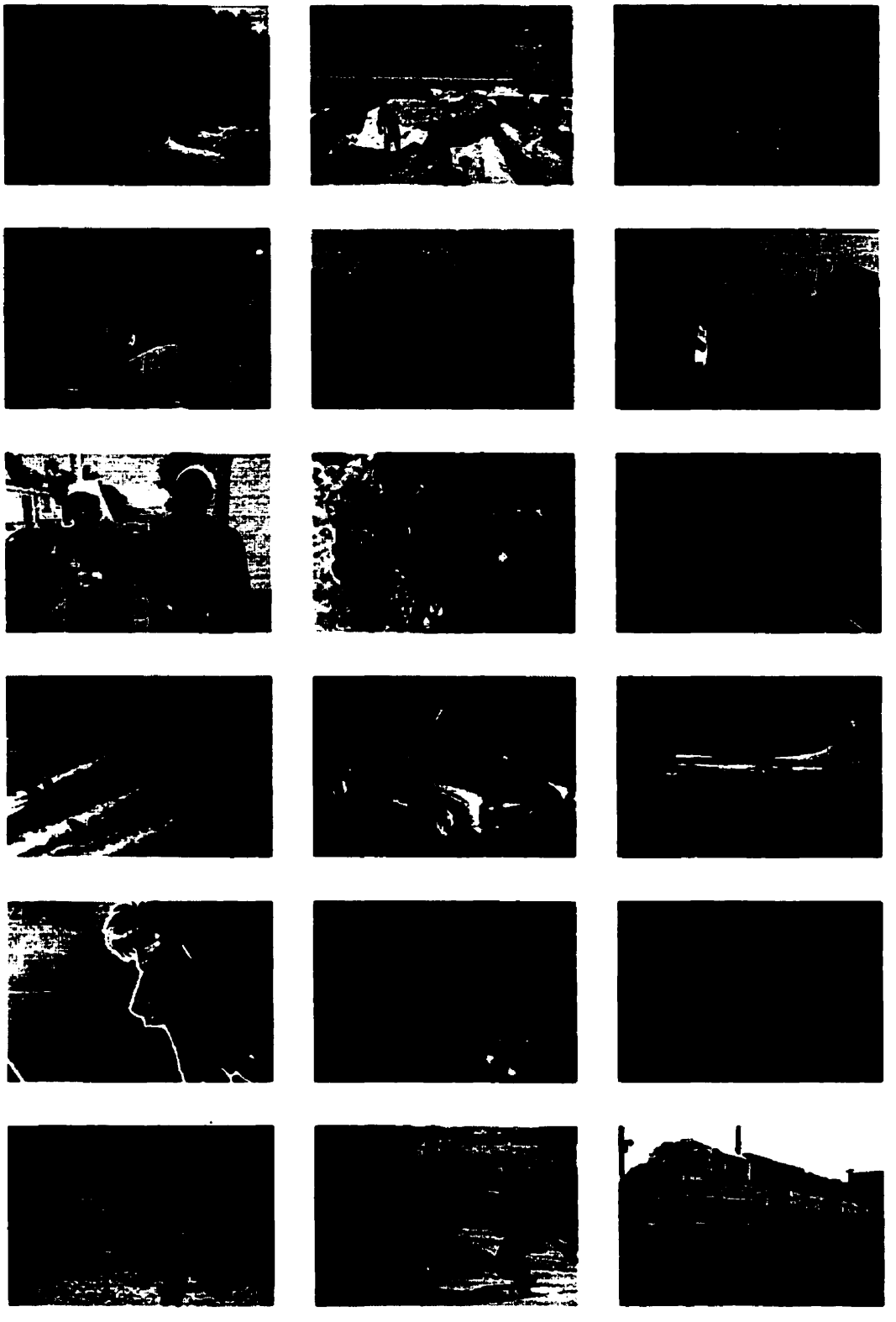


FIGURE 3.3. Selection of images from the database.

FIGURE 3.4. Selection of images from the database.

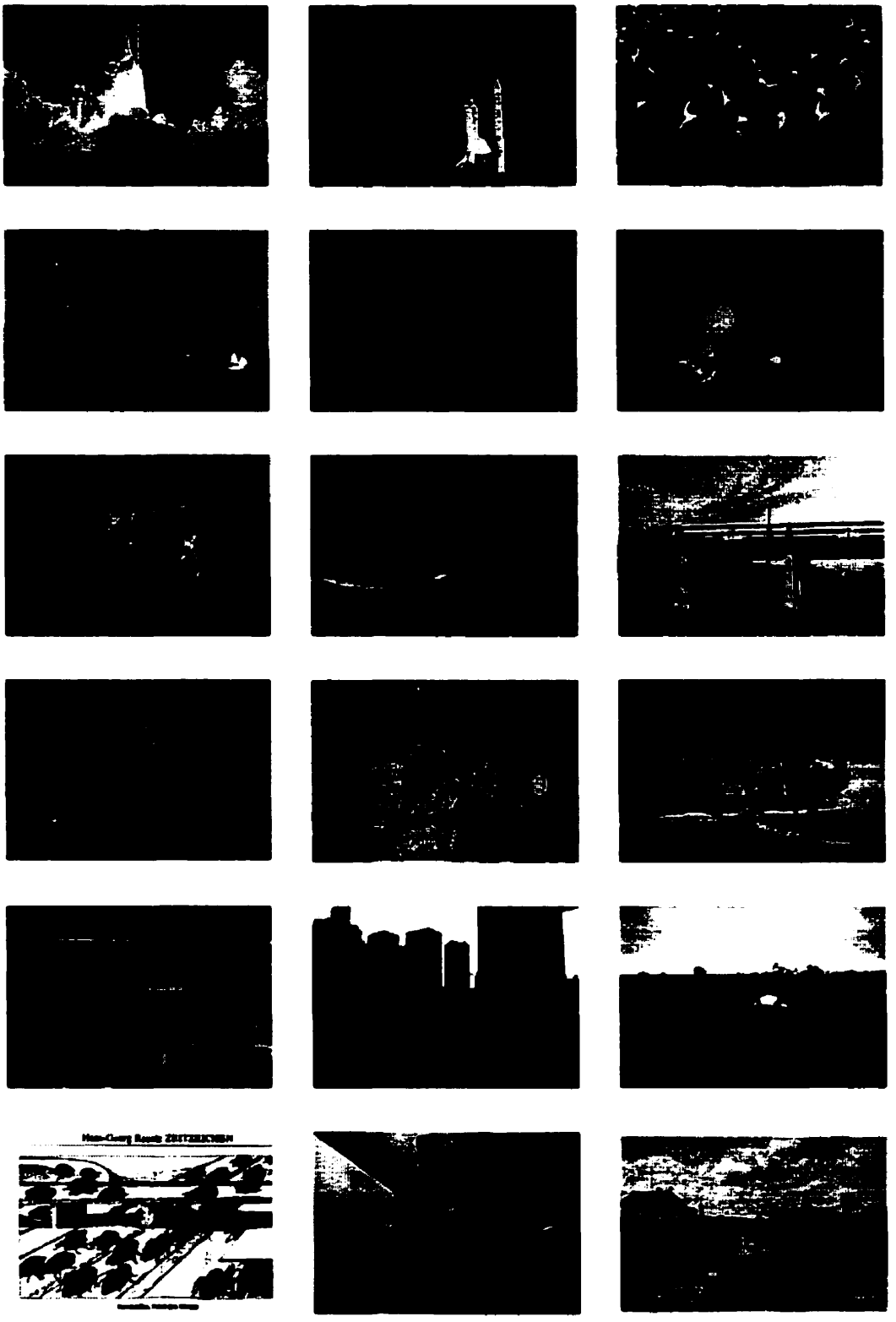


FIGURE 3.5. Selection of images from the database.

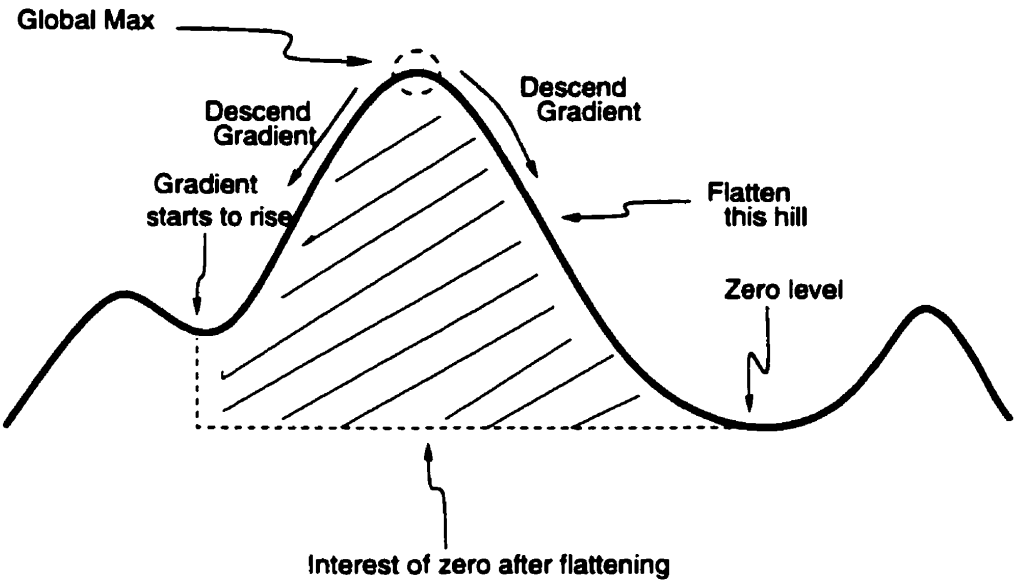
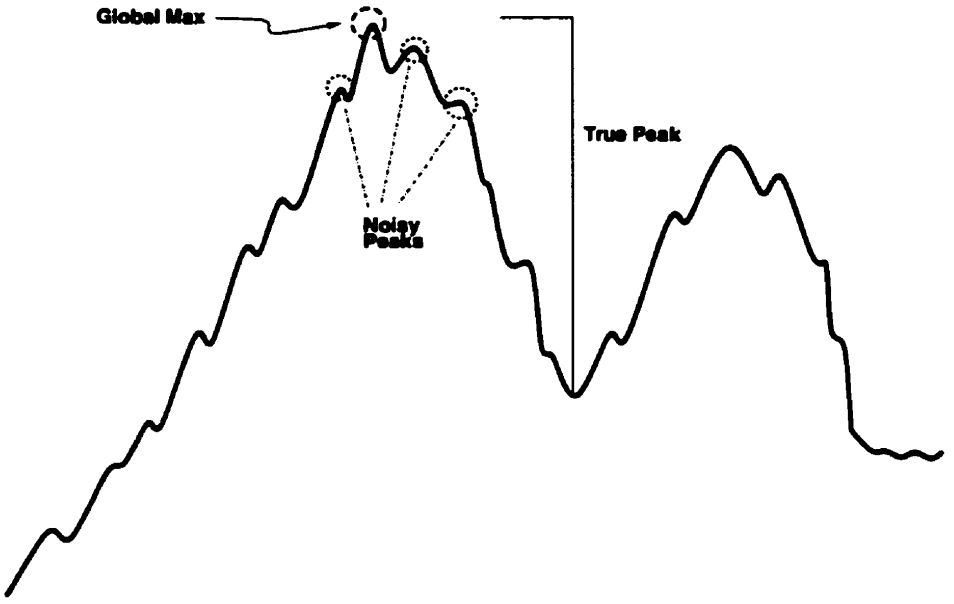


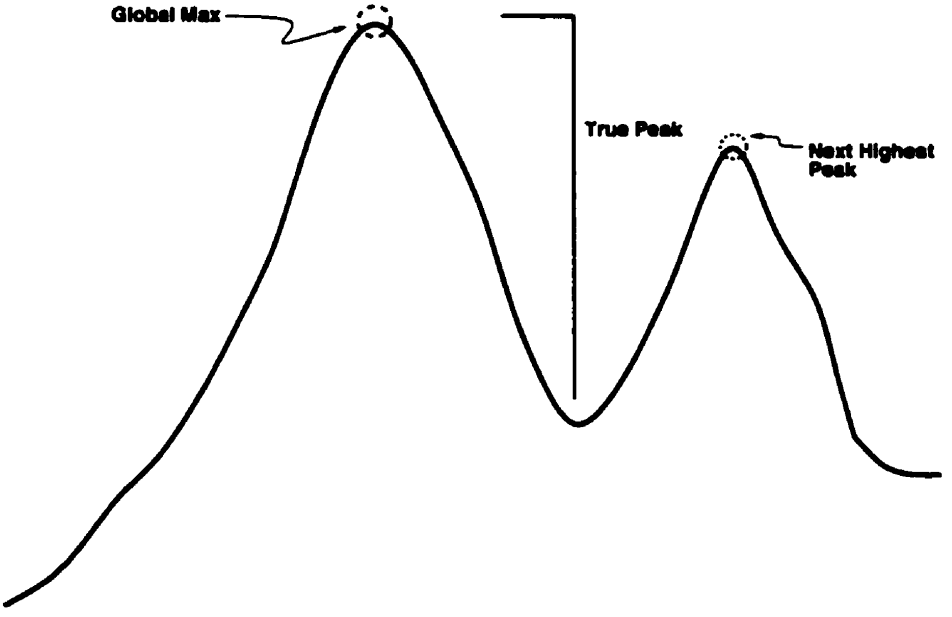

FIGURE 3.6. Non-maximum suppression can be used to extract coordinates of interest locations from an interest map. Non-maximum suppression involves finding the global maximum in the interest map and "flattening" the interest "hill" surrounding it. The procedure is repeated until the whole image has a an interest of zero.

format was chosen for practical purposes: of the five operators used in this research, four required that the image input be in black and white pgm format. Since colour is an important cue in human perception[**7, 31**], and since the majority of operators did not use this cue, it seemed unfair for humans to be able to use this very important extra cue while the operators did not use it. Since the original database compiled for this research consisted of fifty images used by Rensink, O'Regan and Clark in their change blindness experiments, the dimensions of these images was used. More images were selected from a freely available image database at the University of Wisconsin (`ftp://ftp.wustl.edu/multimedia/images/jpeg/`). These images were converted to black and white pgms and scaled to the desired size.

Subimages were extracted by cutting a 75x75 pixel square about the coordinates of the interest maxima returned by the operators for each image. Four of the five operators produced intensity maps where intensities corresponded to degree of interest (Figure 3.8(b)). The operators for edge density, edge orientation, convexity and radial




(a) When non-maximum suppression is applied to a noisy interest map, spurious results are returned. After finding the global maximum, the algorithm will find the noisy peaks, which do not contain any additional information.




(b) When non-maximum suppression is applied to a smoothed interest map, there are fewer spurious results returned.

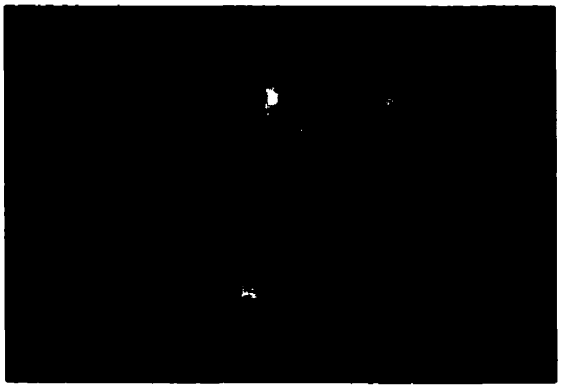
FIGURE 3.7

symmetry produced such outputs. The Caltech operator[**10**] outputs the coordinates of the locations of interest to a text file. Since only the Caltech operator returned coordinates of interest locations, a method was devised to extract coordinates of maximum interest from the interest maps of the other operators.

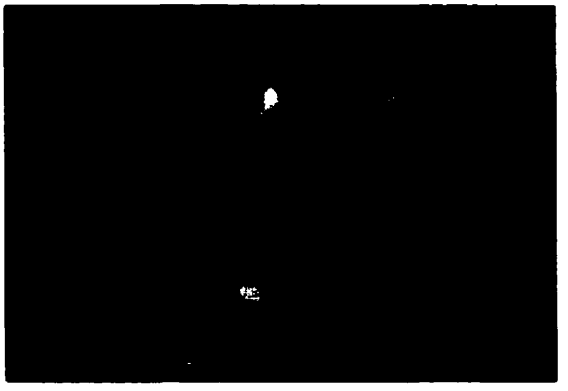
Coordinates of the interest maxima were selected using *non-maximum suppression* (Figure 3.6). A global interest value is found in the interest map and its coordinates are stored in a file. The maximum is suppressed by setting its value to zero. All non-maximum values in the maximum's neighbourhood are also suppressed by examining the nearest neighbours of the maximum and setting their values to zero if their interest value is less than or equal to the maximum value. The comparison and suppression continues recursively along the descending interest gradient until either an interest value of zero is encountered or the interest values begin to rise again. A new global maximum for the modified interest map is then found and the suppression of the maximum and all non-maximum values in its neighbourhood is executed. This process is repeated until the whole interest map has been reduced to a uniform interest of zero. This method of isolating the locations of maximum interest is desirable since it avoids duplicate responses by suppressing equi-interesting locations in the neighbourhood of the maximum. Furthermore, it does not rely on arbitrary thresholding to reduce the number of interest coordinates returned. Methods of locally inhibiting non-maximum responses are biologically inspired[**15**] and are used in various computational attention models such as the winner-take-all model[**10, 4, 33**].

In order for the non-maximum suppression to effectively extract locations of maximum interest without returning multiple responses from the neighbourhood of the maximums, the interest map used must be smoothly varying. If the interest map does not vary smoothly, noisy peaks in the map will cause spurious results to be returned by the extraction (Figure 3.7(a)). For example, Figure 3.9(a) shows the top-ten interest points for Figure 3.8(a). These points were obtained by applying non-maximum suppression to the original interest map (Figure 3.8(b)). Many of the interest points

(a) Original image.

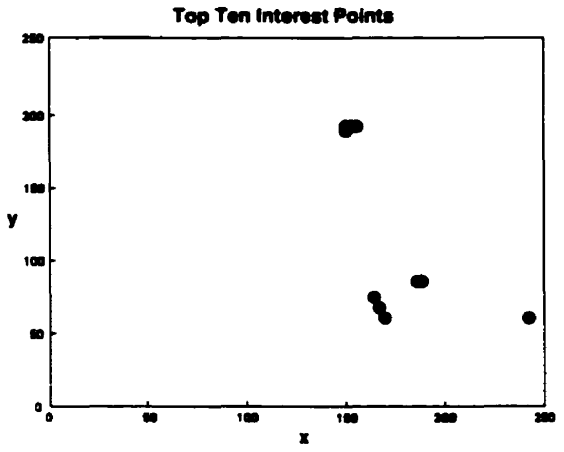


(b) The interest map generated by the edge density operator[2] for Figure 3.8(a).

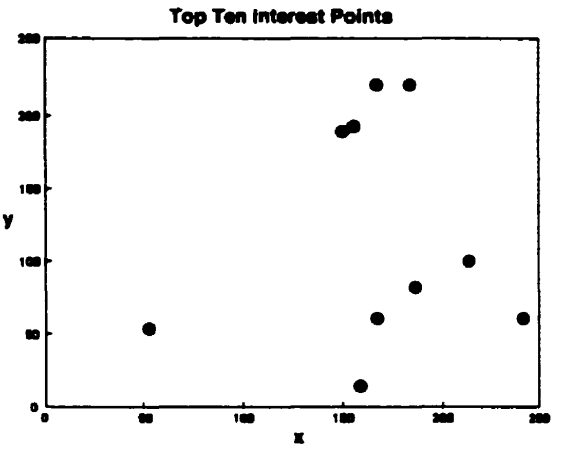


(c) The smoothed version of the interest map in Figure 3.8(b)

FIGURE 3.8. Four of the five computational attention operators used for this research generated interest maps. These maps had to be smoothed before they could be used to find locations of interest. Here we have the original (3.8(b)) and the smoothed (3.8(c)) versions of the interest map generated by the edge density operator for the image shown in 3.8(a).




(a) Interest points for the image in Figure 3.8(a) using the edge density interest map in Figure 3.8(b). Applying non-maximum suppression to the original interest map results in clustered interest points.




(b) Interest points for the image in Figure 3.8(a) using the smoothed edge density interest map in Figure 3.8(c). Applying non-maximum suppression to the smoothed interest map results in more evenly distributed interest points.

FIGURE 3.9

are clustered together and therefore do not convey any additional information about the image.

Smooth interest maps produce fewer spurious interest points since fewer noisy peaks exist to be selected as global maximums (Figure 3.7(b)). Therefore, interest maps were smoothed using a Gaussian. The size of the Gaussian was determined empirically for each operator and was of a lower scale than the operator so as not to lose important information. For the edge density operator, a Gaussian with $\sigma = 5$ was used. The interest points obtained by applying non-maximum suppression to the smoothed interest map of Figure 3.8(a) (Figure 3.8(c)) were more evenly distributed than those obtained using the original interest map (see Figure 3.9(b)).

# CHAPTER 4

# Experimental Results

In this chapter, the performance of each operator is analysed independently and the overall performance of the sample population is observed. Comparisons of operator performance is also undertaken. Operator performance is based on hit rates, false alarm rates and percentage of correct answers.

## 4.1. Results

As described in Chapter 3, the evaluation of the attention operators was based on an image recognition task. After having been shown a training set of 20 images, subjects were queried with respect to their recognition of the training images. Each query consisted of a subimage extracted using one of the six operators described in Section 2.4, with a 50% probability that the subimage originated from the training set. Subjects were asked to respond YES or NO to whether they could recognize the subimage as originating from the training set. A positive query consisted of a subimage from the training set, while a negative query consisted of a subimage which did not originate from the training set.

The data collected during the experiment was analysed to determine which operator yielded subimages which were most easily recognizable by humans. The mean percentage of correct responses were calculated for each operator. The percentage of correct responses encompassed affirmative responses to positive queries (true-positives)

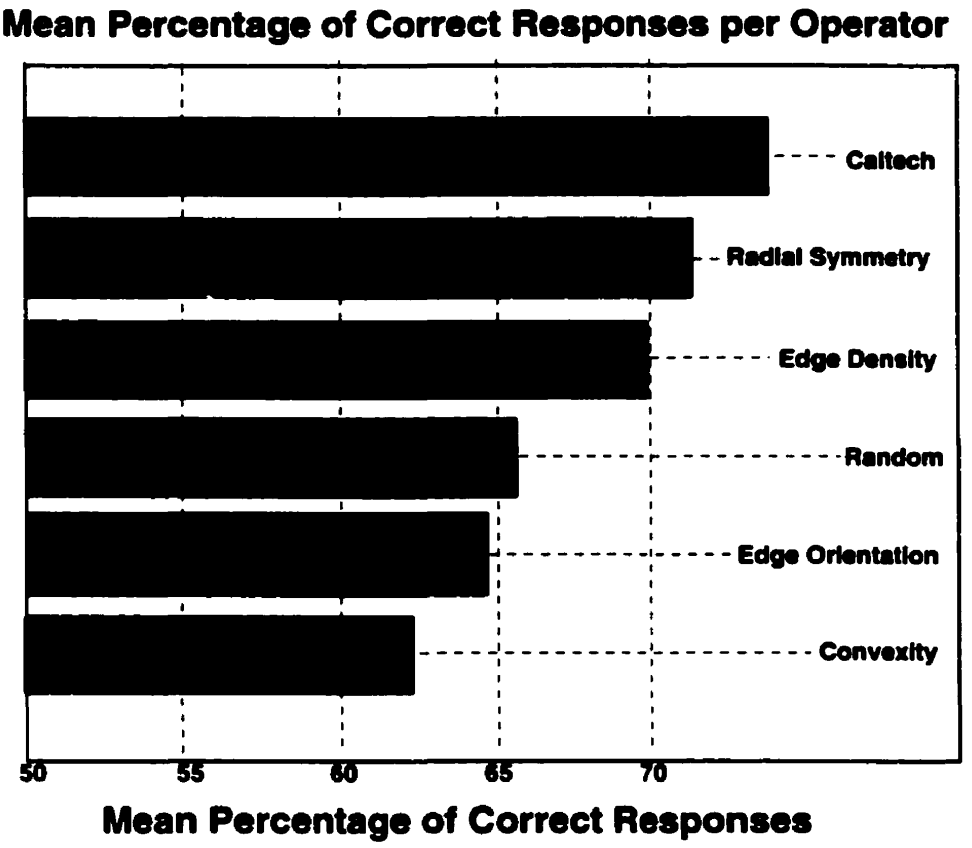



FIGURE 4.1. The percentage of correct answers over all subjects for each operator shows that subjects displayed better performances with subimages from the Caltech, radial symmetry and edge density operators. See Table 4.1

and negative responses to negative queries (correct-rejections). The results are shown in Figure 4.1 and the standard deviation for each operator are shown in Table 4.1. As can be observed in Figure 4.1, subjects had higher percentage of correct responses to queries consisting of subimages extracted using the Caltech, radial symmetry and density operators. A one-way ANOVA found that the probability that the differences in the correct response rates occurred by chance was $p = 0.138$ ($F(5, 36) = 1.798$). However, since both true-positives and correct-rejections are combined in this graph, it does not show the relationship between the types of error.

A more informative measure of how well the operators performed would be to look at which operator or operators yielded the highest rate of true-positives (hits) while maintaining the lowest rate of false-positives (false alarms). A good operator

**Mean Percentage Correct for Each Operator**

| Operator | Mean | Std. Dev. |
|---|---|---|
| Caltech | 73.8% | 8.93% |
| Radial Symmetry | 71.4% | 6.92% |
| Edge Density | 70.0% | 10.4% |
| Random | 65.7% | 8.11% |
| Edge Orientation | 62.4% | 5.67% |
| Convexity | 64.8% | 10.9% |

TABLE 4.1. Mean percentage of correct responses and standard deviation for each operator. See Figure 4.1

will maximize its hit rate while keeping its false alarm rate low. Similarly, an inferior operator will have a low hit rate and a high false alarm rate. A *discriminability measure*, $A'$, and a *bias*, $B''_D$ [**27, 5**], can be calculated for the mean hit rate and mean false alarm rate for each operator. The discriminability, $A'$, is a measure of how well a subject can differentiate between subimages they have seen in the training set and subimages from new images. This form of the discriminability measure is used when a full operator characteristic curve cannot be graphed or when there are hit rates of 1 or false alarm rates of 0[**27**]. In the case of this study only one data point was collected for each operator for each subject; a full operator characteristic curve would require several data points for each operator for each subject. Furthermore, there were occasions where subjects reported hit rates of 1 and false alarm rates of 0.

The $A'$ measure is based on the graphical analysis of Figure 4.2. Figure 4.2 represents the full range of possible responses to queries. The false-alarm rate (FA) is plotted along the $x$-axis while the hit rate (H) is plotted along the $y$-axis. The point (FA,H) represents one response and two lines are drawn through it. The first line goes through point (0,0) (negative answers to all queries), while the second line passes through point (1,1) (affirmative answers to all queries). These two lines divide the graph into four areas: **S**, **I**, $\mathbf{A}_C$ and $\mathbf{A}_L$. The area **I** contains scores where the discriminability is inferior to the data point; area **S** contains scores where discriminability is superior to the data point; areas $\mathbf{A}_C$ and $\mathbf{A}_L$ contain scores whose discriminability is

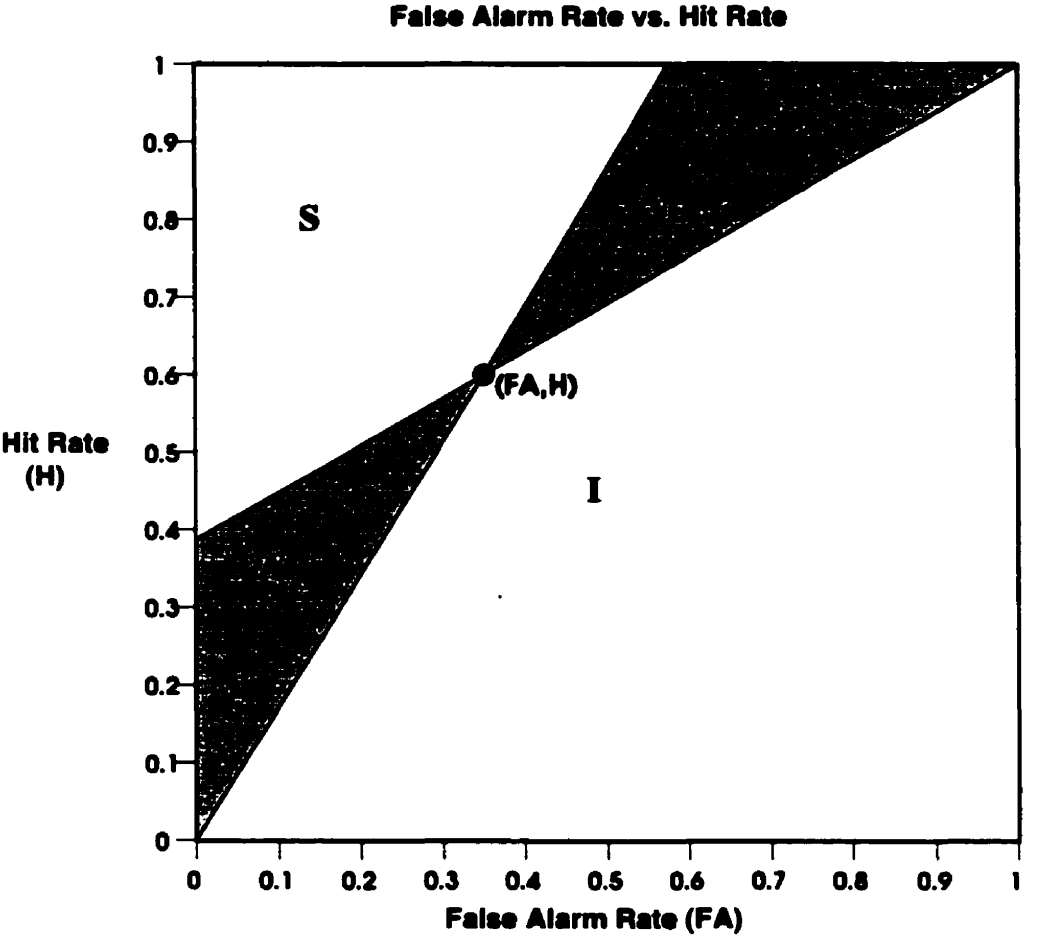



FIGURE 4.2. The discriminability measure, $A'$, can be described in terms of the areas of the false alarm rate (FA) versus hit rate (H) graph which contains all possible scores. The data point (FA,H) represents one particular score. The area **S** contains scores with discriminabilities superior to the data point; area **I** contains scores with discriminabilities inferior to the data point; areas $\mathbf{A}_C$ and $\mathbf{A}_L$ contain scores which are ambiguous as compared to the data point. This figure is a reproduction of an image originally appearing in [**5**].

ambiguous with respect to the data point. The formula for $A'$, which is an estimate of the area under the characteristic curve of the data, is based on Figure 4.2 and is given by[**5**]:

$$A' = \mathbf{I} + \frac{1}{2}(\mathbf{A}_C + \mathbf{A}_L).$$

The discriminability can also be calculated using the hit rate (H) and the false alarm rate (FA)[**27, 5**]:

$$A' = \frac{1}{2} + \frac{(\mathrm{H} - \mathrm{FA})(1 + \mathrm{H} - \mathrm{FA})}{4\mathrm{H}(1 - \mathrm{FA})}.$$

The discriminability is directly proportional to the subjects' capability of differentiating between seen and unseen images; a discriminability of $A' = 0.5$ indicates chance.

The bias, $B''_D$, is a measure of the liberalness or the conservativeness of reporting. Liberal reporting is associated with a higher likelihood of guessing that ambiguous subimages came from the training set, while conservative reporting is associated with a lower likelihood of guessing that ambiguous subimages originated from the training set. The area $\mathbf{A}_L$ in Figure 4.2 contains scores where reporting was more liberal than for the data point (FA,H). Similarly, the area $\mathbf{A}_C$ contains scores where reporting was more conservative than for the data point. The bias, $B''_D$, can be calculated from $\mathbf{A}_L$ and $\mathbf{A}_C$[**5**]:

$$B''_D = \frac{(\mathbf{A}_\mathrm{L} - \mathbf{A}_\mathrm{C})}{(\mathbf{A}_\mathrm{L} + \mathbf{A}_\mathrm{C})}.$$

The bias can also be calculated using the false alarm and hit rates[**5**]:

$$B''_D = \frac{(1 - \mathrm{H})(1 - \mathrm{FA}) - (\mathrm{H})(\mathrm{FA})}{(1 - \mathrm{H})(1 - \mathrm{FA}) + (\mathrm{H})(\mathrm{FA})}.$$

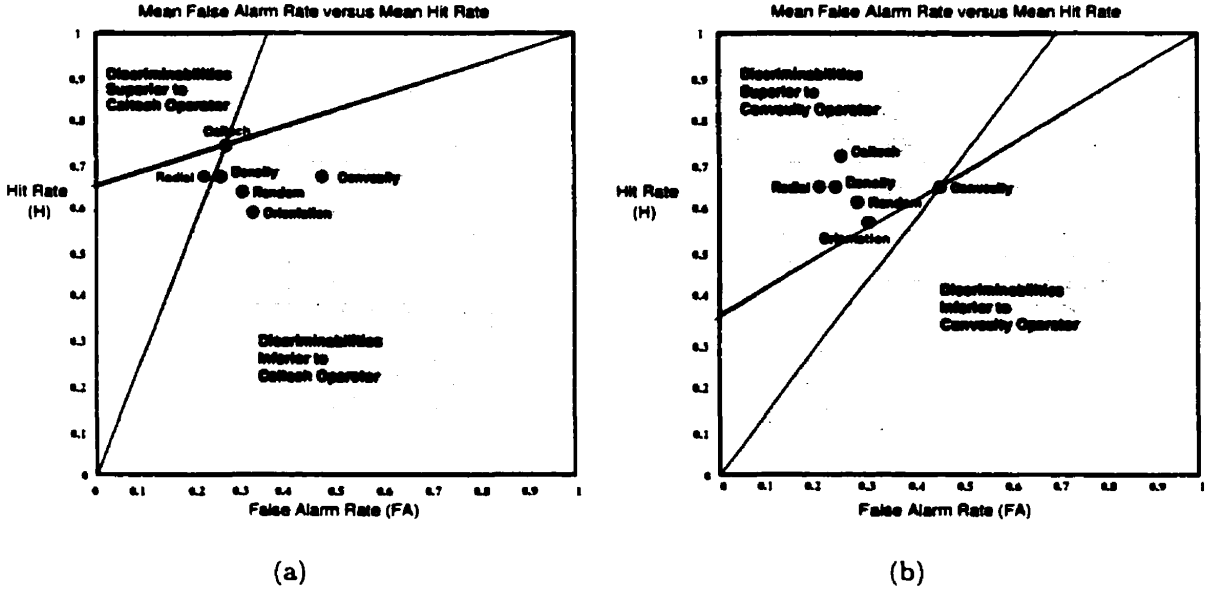
A positive bias ($B''_D > 0$) indicates conservative reporting, while a negative bias ($B''_D < 0$) represents liberal reporting. A bias of zero gives no bias information.

Table 4.2 shows the discriminabilities and biases associated with each operator using mean hit rates and mean false alarm rates for each operator ($\overline{\mathrm{H}}_{\mathrm{op}} = \sum_{i \in \mathrm{subjs}} \mathrm{H}_{\mathrm{op},i}$, $\overline{\mathrm{FA}}_{\mathrm{op}} = \sum_{i \in \mathrm{subjs}} \mathrm{FA}_{\mathrm{op},i}$). The discriminabilities associated with the Caltech and radial symmetry operators surpass those associated with the other operators, while the convexity operator has an associated discriminability which is substantially lower than the other operators'. Figures 4.3 illustrates discriminabilities of Table 4.2. From Figure 4.3(a) it can be seen that reporting for the Caltech operator shows superior discriminability to all operators except to the radial symmetry operator. The response to the radial symmetry operator with respect to the Caltech operator is ambiguous,

$A'$ and $B''_D$ for the Mean Hit and False Alarm Rates for Each Operator

| Operator | $A'$ | $B''_D$ |
|---|---|---|
| Caltech | 0.82 | −0.03 |
| Radial Symmetry | 0.80 | 0.22 |
| Edge Density | 0.79 | 0.17 |
| Random | 0.74 | 0.13 |
| Edge Orientation | 0.70 | 0.19 |
| Convexity | 0.66 | −0.27 |

TABLE 4.2. Discriminability, $A' = f(\overline{\text{FA}}, \overline{\text{H}})$, and bias, $B''_D = g(\overline{\text{FA}}, \overline{\text{H}})$, for the mean hit rates ($\overline{\text{H}} = \sum_{i \in \text{subjs}} \text{H}_i$) and mean false alarm rates ($\overline{\text{FA}} = \sum_{i \in \text{subjs}} \text{FA}_i$) for each operator (see Figure 4.3).




(a) (b)

FIGURE 4.3. Discriminabilities associated with the operators using a plot of mean hit rates versus mean false alarm rates of the subjects for each operator. Figures 4.3(a) and 4.3(b) show the same plot with different analysis of the data. Figure 4.3(a) shows the relationship between the Caltech operator and the other operators. No operator tested shows superior discriminability in subject reporting to the Caltech operator. The relation between the Caltech operator and the Radial Symmetry operator is ambiguous. Figure 4.3(b) shows the relationship between the convexity operator and the other operators. The Convexity operator shows less discriminability in subject reporting than any of the other operators including the random operator.

**Mean Discriminabilities with Each Operator**

| Operator | $A'$ |
|---|---|
| Caltech | 0.820 |
| Radial Symmetry | 0.806 |
| Edge Density | 0.776 |
| Random | 0.749 |
| Edge Orientation | 0.704 |
| Convexity | 0.664 |

TABLE 4.3. Mean discriminabilities ($\overline{A'} = \sum_{i \in \text{subjs}} A'_i$) for each operator over all subjects.

lying in an area of more conservative reporting than the Caltech operator. In contrast, Figure 4.3(b) shows that subjects had the lowest discriminability when reporting for images extracted using the convexity operator.

Using a one-way ANOVA, it was found that the mean discriminabilities between the operators were statistically significant, with the probability of the differences between the mean discriminabilities occurring by chance being $p = 0.024$ ($F(5, 36) = 2.977$). [1] Note that the discriminability measures are more significant than the correct response rates; the discriminabilities indicate that there is a difference in the overall ability of the operators to select locations of interest.

## 4.2. Discussion

The results show that there is a statistically significant difference in the ability of the operators to select locations of interest, with the Caltech operator and the radial symmetry operator performing best overall and the convexity operator performing worst overall. Subjects showed 5% better discriminability with subimages obtained using the Caltech operator than with subimages extracted using the edge density operator and 9% better discriminability than with selections from the random

[1] The mean discriminabilities denoted by $\overline{A'}_{\text{op}} = \sum_{i \in \text{subjs}} A'_{\text{op},i}$ and tabulated in Table 4.3, were used instead of the discriminabilities of the mean false alarm rates and mean hit rates ($A'_{\text{op}} = f(\overline{\text{FA}}_{\text{op}}, \overline{\text{H}}_{\text{op}})$) shown in Table 4.2 for ease of calculation. By examining Table 4.2 and Table 4.3 it can be seen that these values are approximately equal.

operator. Subimages obtained using the radial symmetry operator generated a 4% better discriminability in reporting than the edge density operator and an 8% better discriminability in reporting than the random operator. By contrast, the convexity operator had an associated discriminability 11% *worse* than the random operator and 19% worse than the Caltech operator. The bias of the mean hit rates and mean false alarm rates of the subjects with the various operators (Table 4.2, Figure 4.3(a)) shows that response to the radial symmetry operator was more conservative than response to the Caltech operator. In other words, subjects were less likely to answer YES to queries testing the radial symmetry operator than for those testing the Caltech operator, implying that subjects were less likely to guess that ambiguous images extracted using the radial symmetry operator originated from the training set. This effect may simply be a statistical artifact or it may indicate that the subimages obtained using the radial symmetry operator were more ambiguous than those selected using the Caltech operator. However, the take-home message remains that the relationship between the performance of the Caltech operator and the performance of the radial symmetry operator is ambiguous, and since the difference between their associated discriminabilities is not statistically significant it can be assumed that both score equally well with respect to the discriminability measure.

The poor performance of the convexity operator, on the other hand, may be due to the selection of images which were shown to the subjects. Images with insufficient shading information, for example cartoon images (see Figure 4.4(a)), could cause the operator to return unintuitive locations. The location which shows the highest convexity response (brighter region) in the convexity interest map of Figure 4.4(b) is the region which has the most noticeable shading in the original image displayed in Figure 4.4(a). The authors of the convexity operator mention that their operator will not select cartoon drawings of faces in an image, but will instead select real faces[**28**]. This is an advantage when attempting to select faces in natural scenarios, such as in camouflage breaking tasks, which was the original purpose of the convexity operator. However, what makes the operator strong in camouflage breaking may be a

(a) One of the images used in the evaluation. Inset: the first choice of the convexity operator. Note that the first choice is in a region where shading effects are most noticeable.

(b) Response of the convexity operator to the image in Figure 4.4(a). Note that the locations with the highest response are those which show the greatest amount of shading.

FIGURE 4.4

hindrance when it comes to selecting locations of interest in an image database where many different types of images are stored. The radial symmetry operator, in contrast, does not rely on scenes containing convex and concave objects. The radial symmetry operator will select locations with high radial symmetry, regardless of shading information. Furthermore, since the human form tends to be symmetric, the radial symmetry operator is sensitive to human forms. This gives the radial symmetry operator an advantage over the edge density operator and the edge orientation operators because humans are also sensitive to human forms.

The Caltech operator's use of several different feature maps and winner-take-all model has a strong neurophysiological basis[**15**]. Furthermore, since the operator is multiscale, problems of scale-dependence are lessened and fewer spurious results are returned. In fact, the multiscale aspect of the Caltech operator makes the comparison between it and the other operators somewhat unfair; finding a unique, optimal scale for all images was not possible and therefore the scale-dependent operators performed suboptimally for some subset of the images. The combination of several different feature maps in the Caltech operator also reduces the reliance on one particular feature and when combined in the saliency maps, the strongest of the features will be taken into account. However, despite the computational differences between the Caltech operator and the radial symmetry operator, the difference in their associated discriminabilities was not statistically significant. This indicates that radial symmetry may be as important a cue for human preattentive vision as colour, luminance and edge orientation.

Responses to subimages obtained using the edge density operator showed a comparable discriminability to those recorded using the radial symmetry operator. However one of the drawbacks of the edge density operator is that it selects regions which are maximally different from the mean density over the whole image. Thus, if an image is very busy, a relatively empty area may be selected and, conversely, if the image is largely composed of areas of low edge density, an area with high edge density will be chosen. This can be a drawback in selecting interest locations in natural

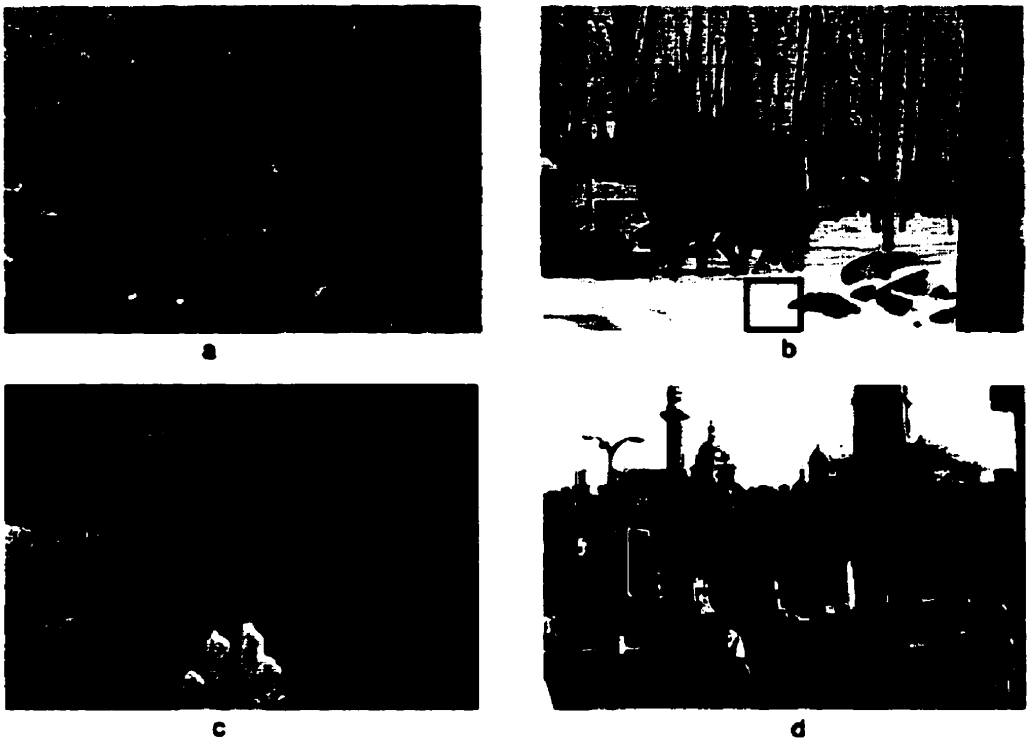
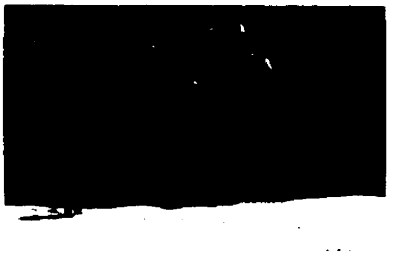



FIGURE 4.5. The edge density operator selects locations where the edge density is maximally different than the mean edge density of the entire image. This method, however, is not always optimal. Figures **a** and **b** show cases where edge density is high throughout the image, causing a region of low edge density to be selected by the operator. Figures **c** and **d** show cases where edge density is low throughout the image, causing the operator to select locations of high density which are not of central interest to the image.

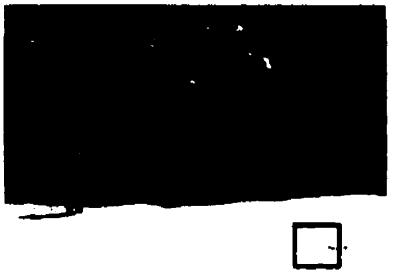
settings (Figure 4.5) since in cases where the scene has a small patch of dense edge elements, such as vegetation, but a great deal of sparser elements, such as faces, the busy elements will most likely be selected. Similarly, in cases where the whole image contains uniformly dense regions, locations where edge density is sparse, such as parts of the sky, will be selected. The difference between the discriminabilities associated with the radial symmetry operator and the edge density operator might also be due to the scale dependence of the operators. The scale for each operator must be set by the user, and the same scale is not necessarily optimal for all images. The orientation operator also suffered from the same drawbacks as the edge density operator. The

(a) Interest location selected by the radial symmetry operator

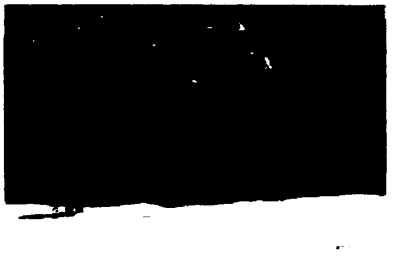
(b) Interest location selected by the convexity operator



(c) Interest location selected by the Caltech operator



(d) Interest location selected by the edge orientation operator



(e) Interest location selected by the edge density operator

(f) Close up of edge orientation operator selection

FIGURE 4.6. In this example the radial symmetry (Figure 4.6(a)), convexity (Figure 4.6(b)) and Caltech (Figure 4.6(c)) operators selected locations which are intuitively salient while the edge orientation (Figure 4.6(d)) and edge density operators (Figure 4.6(e)) selected non-intuitive locations. The edge density operator selected a region of higher density as compared to the global average of the image, while the orientation operator selected the signature in the lower right-hand corner of the image. The signature, which can be seen in more detail in Figure 4.6(f), consists of lettering with a great deal of curvature and thus has a very different orientation profile from the rest of the image.

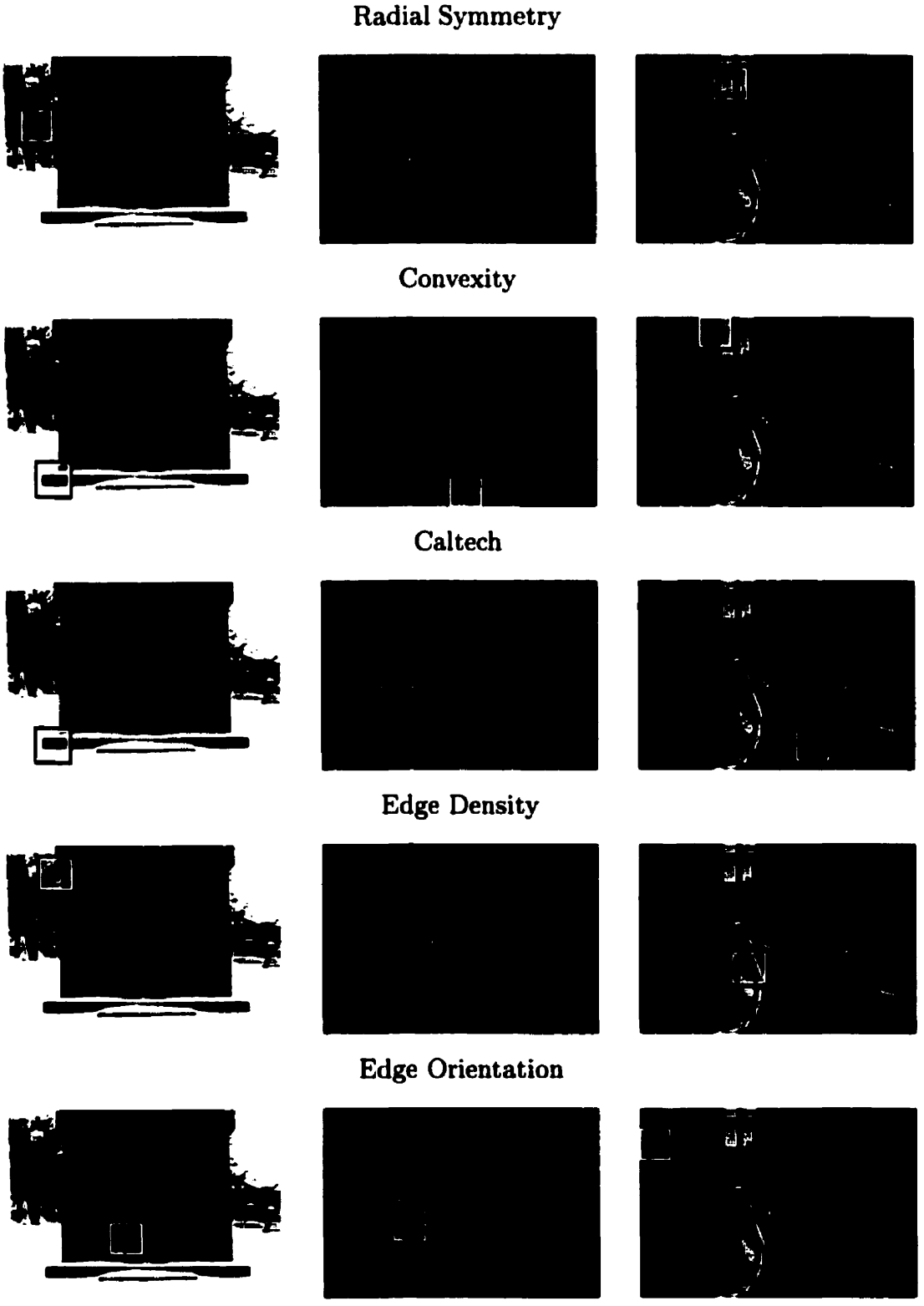



FIGURE 4.7. Response of the operators on three images.

edge orientation operator selects locations where the local predominant edge orientation is maximally different from the global predominant edge orientation. An example of the drawbacks of the edge orientation and edge density operators as compared to the other three computational operators can be seen in Figure 4.6. In the case of the image in Figure 4.6, neither the edge density operator nor the edge orientation operator return intuitive results (Figures 4.6(d) and 4.6(e)). Operator selections for other images are presented in Figure 4.7.

Though certain subimages returned were less than intuitive and even though subjects claimed to have "guessed" most of the time and felt they had done very badly, the fact that the mean discriminability over all subjects and all operators was $A' = 0.76$ ($\sigma = 0.05$) points to some implicit memory effects involved in recognizing the images. Furthermore, since many of the images had semantic peculiarities as well as texture differences, these cues could have been used by the subjects to identify subimages which did not contain locations of interest. In future, a study which would compare people's scanpaths to the locations selected by the operators could be useful since studying the scanpaths would show where the subjects attended and for how long and would factor out guessing based on the textures of the images.

# CHAPTER 5

# Conclusions and Future Work

This thesis presented a novel method for evaluating computational attention operators motivated by human visual attention, human visual memory and human image recognition. Results show that the proposed method of evaluation is indeed sound and that certain attention operators are better suited than others at selecting image locations which would be of interest to humans. Five different operators were evaluated: an edge density operator[**2**], an edge orientation operator[**2**], a radial symmetry operator[**19**], a convexity operator[**28**] and a combination luminance contrast/colour/edge orientation operator (Caltech operator)[**10**]. The Caltech operator and the radial symmetry operator rated highest in the evaluation, while the convexity operator rated lowest.

Human subjects were asked to view training images and then indicate whether they recognized subimages containing locations selected by the different attention operators. The ability of the subjects to discriminate between subimages originating from the training set and subimages from other images was used to evaluate the operators. As a control, subimages extracted using random image coordinates were also presented to the subjects. The Caltech and radial symmetry operators selected locations which facilitated discrimination more than any other operators; in contrast, subjects demonstrated the lowest discriminability with selections from the convexity operator.

The Caltech operator differed from the other operators by its use of a winner-take-all network which simulated the neurophysiology of human visual attention, as well as by its use of multiple features and scales to calculate interest. The radial symmetry operator had an advantage over the other operators because of its sensitivity to human forms. Since people are also sensitive to human forms, locations returned by the radial symmetry operator would be more likely to be considered interesting. Although the Caltech operator used a very complex model of attention, it did not fare significantly better than the radial symmetry operator. This indicates that symmetry may be an important cue in human preattentive vision and that selecting locations of interest need not require complete models of attention, but only an appropriate choice of interest feature.

An interesting result was that subjects scored better than chance discriminability (discriminability greater than 0.5) with subimages selected at random from the images. This suggests that humans require very few cues to recognize an image. The discriminability associated with the random images seems to indicate that a random subimage is reasonably effective at describing an image; however a comparison of the differences in discriminabilities associated with each operator suggests that certain subimages contain more information than others. Thus, the operators which had the highest associated discriminabilities, the Caltech operator (discriminability of 0.82) and the radial symmetry operator (discriminability of 0.80) selected locations which contained more information than the other operators. Similarly, the operators which had the lowest discriminabilities associated with them, the convexity operator (discriminability of 0.66) and the edge orientation operator (discriminability of 0.70), selected locations which contained less image information than the other operators and random subimages. Therefore, in a situation where a large number of images need to be identified, subimages obtained using the radial symmetry operator or Caltech operator will dramatically reduce the number of misclassified images.

## 5.1. Future Work

The Caltech operator had an advantage over the other operators as it used multiple scales in selecting locations of interest. Future work needs to be conducted to determine how great a factor scale is in the results for each operator. A similar study, performed with the same operators but at several different scales could determine a close-to-optimal scale for each operator. Furthermore, unlike the other operators, the Caltech operator output coordinates of interest locations rather than interest maps. A method of extracting coordinates of interest maximas from the interest maps was devised especially for this thesis (see Section 3.4). However, in may cases the interest maps contained regions of equal interest. A great deal of information may be lost by extracting the first set of maximal coordinates discovered from each region. A better method of subimage extraction could involve finding the center of a region of equal interest using a convex hull. In this manner, the center of the interest region would be used as the center of the subimage to be extracted. A more seductive idea would be to extract equi-interesting subregions from the images. Unfortunately, this would yield unequally sized subimages and could possibly skew results. Performing the experiment with subimages extracted using the centers of convex hulls could indicate whether the method of subimage extraction played a role in the results of the experiment.

The question of how to extract subimages brings to the forefront the issue of how important a role context plays in the recognition of subimages. A possible experiment could study how the different sizes of subimages and the number of subimages shown per query affects response and whether there is a point where discriminability is equal for all operators. Some preliminary studies in this direction were carried out prior to finalizing the methodology of the evaluation. The preliminary studies showed that increasing the subimage size to 100x100 pixels caused subjects to display smaller differences in performance between the operators. Displaying three subimages per query produced dramatic improvements in overall image recognition. A study comparing

subjects' scanpaths to the locations selected by the operators could be useful to factor out contextual memory effects.

In conclusion, the results of the evaluation show a clear, statistically significant result of the Caltech and radial symmetry operators performing exceptionally well with respect to the other operators and the convexity operator performing exceptionally poorly with respect to the other operators. It is therefore clear that human preattentive vision, human visual memory and human image recognition can be used as a means of evaluating computational attention operators.

# REFERENCES


[1] Yoram Bonneh, Daniel Reisfeld, and Yehezkel Yeshurun, *Quantification of local symmetry: Application for texture discrimination*, Spatial Vision **8** (1995), no. 4, 515–530, Feature Issue on Symmetry.

[2] Eric Bourque and Gregory Dudek, *Viewpoint Selection – An autonomous robotic system for virtual environment creation*, IEEE/RSJ Conference on Intelligent Robots and Systems (Victoria, Canada), vol. 1, October 1998, pp. 526–532.

[3] C.L. Colby, *The neuroanatomy and neurophysiology of attention*, Journal of Child Neurology **6** (1991), S90–S118, Supplemental.

[4] Sean M. Culhane and John K. Tsotsos, *An attention prototype for early vision*, Computer Vision – European Conference on Computer Vision '92 (Santa Margherita Ligure, Italy) (G. Sandini, ed.), Second European Conference on Computer Vision, Springer-Verlag, 1992, pp. 551–560.

[5] Wayne Donaldson, *Measuring recognition memory*, Journal of Experimental Psychology: General **121** (1992), no. 3, 275–277.

[6] G. Dudek and D. Jugessur, *Robust place recognition using local appearance based methods*, IEEE International Conference on Robotics and Automation, April 2000.

[7] M. D'Zmura, P. Lennie, and C. Tiana, *Color search and visual field segregation*, Perception & Psychophysics **59** (1997), 381–388.

[8] James T. Enns and Ronald A. Rensink, *Influence of scene-based properties on visual search*, Science **247** (1990), 721–723.

[9] _____, *Preemption effects in visual search: Evidence for low-level grouping*, Psychological Review **102** (1995), no. 1, 101–130.

[10] L. Itti, C. Koch, and E. Niebur, *A model of saliency-based visual attention for rapid scene analysis*, IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998), no. 11, 1254–1259.

[11] Martin Jägersand, *Saliency maps and attention selection in scale and spatial coordinates: An information theoretic approach*, Proceedings of the 5th International Conference on Computer Vision, 1995, pp. 195–202.

[12] D. Jugessur and G. Dudek, *Local appearance for robust object recognition*, IEEE Computer Vision and Pattern Recognition, June 2000.

[13] Bela Julesz, *Experiments in the visual perception of texture*, Scientific American **232** (1975), no. 4, 34–43.

[14] M.F. Kelly and M.D Levine, *Annular symmetry operators: A method for locating and describing objects*, ICCV '95, 1995, pp. 1016–1021.

[15] C. Koch and S. Ullman, *Shifts in selective visual attention: towards the underlying neural circuitry*, Human Neurobiology **4** (1985), 2428–2434.

[16] Tony Lindeberg, *Detecting salient blob-like image structures and their scales with a scale-space primal sketch*, International Journal of Computer Vision **11** (1993), no. 3, 238–318.

[17] David G. Lowe, *Object recognition from local scale-invariant features*, International Conference on Computer Vision (Corfu, Greece), September 1999, pp. 1150–1157.

[18] Vilayanur S. Ramachandran, *Perceiving shape from shading*, Scientific American **259** (1988), no. 2, 76–83.

[19] D. Reisfeld, H. Wolfson, and Y. Yeshurun, *Context free attentional operators: The generalized symmetry transform*, International Journal of Computer Vision **14** (1995), 119–130, Special Edition on Purposive Vision.

[20] R.A. Rensink, J.K. O'Regan, and J.J. Clark, *To see or not to see: The need for attention to perceive changes in scenes*, Psychological Science **8** (1997), 368–373.

[21] Cordelia Schmid and Roger Mohr, *Local grayvalue invariants for image retrieval*, IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997), no. 5, 530–535.

[22] Cordelia Schmid, Roger Mohr, and Christian Bauckhage, *Comparing and evaluating interest points*, International Conference on Computer Vision, 1998, pp. 230–235.

[23] Gal Sela and Martin D. Levine, *Real-time attention for robot vision*, Real-Time Imaging **3** (1997), 173–194.

[24] Amnon Sha'ashua and Shimon Ullman, *Structural saliency: The detection of globally salient structures using a locally connected network*, International Conference on Computer Vision, 1988, pp. 321–327.

[25] Robert Sim and Gregory Dudek, *Learning and evaluating visual features for pose estimation*, International Conference on Computer Vision (Kerkyra, Greece), September 1999.

[26] D.J. Simons, S.L. Franconeri, and R.L. Reimer, *Change blindness without visual disruptions*, Investigative Ophthalmology and Visual Science, vol. 41, The Association for Research in Vision and Ophthalmology, March 2000, ARVO Annual Meeting, Fort-Lauderdale, Florida, April 30-May 5 2000, p. S750.

[27] Joan Gay Snodgrass, Gail Levy-Berger, and Martin Haydon, *Human experimental psychology*, Oxford University Press, New York, 1985.

[28] A. Tankus and Y. Yeshurun, *Detection of regions of interest and camouflage breaking by direct convexity estimation*, IEEE Workshop on Visual Surveillance (Los Alamitos, CA, USA), 1998, pp. 42–8.

[29] Anne Treisman, *Perceptual grouping and attention in visual search for features and for objects*, Journal of Experimental Psychology: Human Perception and Performance **8** (1982), 194–214.

[30] ———, *Preattentive processing in vision*, Computer Vision, Graphics, and Image Processing **31** (1985), no. 2, 156–177.

[31] ———, *Features and objects in visual processing*, Scientific American **255** (1986), no. 5, 114B–125.

[32] John K. Tsotsos, *The complexity of perceptual search tasks*, International Joint Conference on Artificial Intelligence, 1989, pp. 1571–1577.

[33] John K. Tsotsos, Sean M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando J. Nuflo, *Modeling visual attention via selective tuning.*, Artificial Intelligence **78** (1995), no. 1-2, 507–545.

[34] Carl-Fredrik Westin, Carl-Joahn Westelius, Hans Knuttson, and Gosta Granlund, *Attention control for robot vision*, Computer Vision and Pattern Recognition, 1996, pp. 726–733.