Integrating Overt and Covert Attention Using Peripheral and Central Processing Streams

Calden Wloka

A Thesis submitted to the Faculty of Graduate Studies
in partial fulfilment of the requirements
for the degree of

MASTER OF SCIENCE

Graduate Program in Computer Science
York University
Toronto, Ontario, Canada

October, 2012

Abstract

Visual search, a subproblem of vision, is computationally intractable in the general sense. Attentional methods can be used to attain a tractable approximation; the Selective Tuning (ST) model in particular provides a solution to many challenges of visual search. ST components which address the Boundary Problem, however, have previously not been fully implemented. This thesis implements the proposed visual fixation control using a set of modules including a Peripheral Priority Map, a Fixation History Map, and a History Biased Priority Map. The system is tested in both a physical and a virtual environment. The test on physical hardware demonstrates proof-of-principle performance for the proposed method of fixational control. The virtual data-set compares fixations produced by the system developed in this thesis with human fixational data; our system out-performs AIM in reproducing human-like fixation patterns. Several avenues of future research are proposed using this system as a development platform.

Acknowledgements

There were many people who helped make this thesis possible. Foremost, I would like to thank my supervisor, John Tsotsos, for his support and guidance. He is an excellent scientific role model and provided me with valuable feedback throughout the development and writing of this thesis.

I would also like to thank my thesis defense committee, Minas Spetsakis, Jeff Edmonds, and Maz Fallah, for their time and questions. Neil Bruce and Xun Shi greatly helped with my understanding and implementation of the AIM algorithm. Evgueni Simine and Albert Rothenstein provided extremely helpful programming support.

Finally, I would like to thank my family. The encouragement of my parents, sisters, and grandparents was greatly appreciated. Most of all, the patience, understanding, and support of my wife, Sarah, was invaluable, and I could not have done this without her.

List of Figures

Contents

Chapter 1

Background

1.1 Introduction

Vision is widespread in the animal kingdom and is a powerful sensory tool for interacting with the surrounding environment. Humans and other primates are able to perform a large number of visuo-cognitive tasks with an ease that belies their underlying computational difficulty. The field of computer vision seeks to understand and replicate the capabilities of biological vision, a task generally referred to as the *vision problem*. Although the vision problem itself is generally only loosely or intuitively defined, there are a number of visual sub-problems which are rigorously defined. Of these sub-problems, the *visual search problem* is of primary importance to this study.

Wolfe intuitively described visual search as "those tasks where one looks for something," [91]. Tsotsos more formally defined the problem in [81] as:

> Given a set of memory items or targets and a test display that contains several nontarget items and may or may not contain target items, measure the length of time a subject needs to detect a given

number of the targets in the display.

Using this definition, Tsotsos shows that unbounded visual search (visual search in which the target is explicitly unknown or not used) is computationally intractable. Tsotsos proposes a number of biologically plausible methods to reformulate the problem into a tractable approximation problem which uses natural parameters to optimally guide the approximation. It is these approximation techniques which Tsotsos posits constitute *visual attention* (for the detailed development of these ideas, see [81]. For an up to date summary, see also Chapter 2 of [82]).

Hierarchical processing of an image using a pyramidal representation is a powerful computational technique for increasing visual computation efficiency which also benefits from strong biological support. A detailed mapping of the macaque visual cortex conducted by Felleman and Van Essen arranges visual processing into a 14-layer visual hierarchy with the retina and lateral geniculate nucleus at the bottom and the entorhinal cortex and hippocampus at the top [25]. However, hierarchical processing also introduces a number of computational challenges which must be dealt with: Blurring, Cross-talk, Context, Boundary, Sampling, and Routing Problems.

The primary cause of these hierarchical processing challenges is that each element of any particular layer in the hierarchy connects to an ever-widening cone of influence across the layers (see Figure 1.1). The Blurring Problem refers to the loss of spatial acuity introduced by the hierarchical structures; although a spatially well-localized event may occur at the input layer, the activations it influences as one moves through the hierarchy will progressively become less spatially specific (see the left-hand connectivity cone in Figure 1.1). The converse of the Blurring Problem is the Sampling Problem (represented by the right-most connectivity cone in Figure 1.1); the top-most pyramidal layer represents information in the most abstract sense but also with the least spatial resolution. This represents a distinct computational challenge when recognition requires both a high level of abstraction as well as accurate spatial localization (such as a pedestrian recognizing and avoiding a speeding car heading toward him). When two cones of influence overlap it creates the Cross-talk Problem; the activity of any neuron within the overlapping region becomes a function of both. The Context problem refers to the fact that higher-level visual activity will necessarily be influenced not only by the target to which they are attuned but also by any other visual elements located within the spatial reach of the cone of influence. The Routing Problem is the challenge of finding the specific activation pathway through a hierarchical network which best represents a given stimulus. The Boundary Problem is displayed in Figure 1.2, and it arises due to the fact that for any spatial convolution, there will be a region half the width of the convolution around the edge of an image for which the convolution result is undefined. As one moves up through the layers of a pyramid, this undefined region compounds and grows with each level.

Tsotsos' Selective Tuning (ST) model of visual attention seeks to deal with these problems, though the proposed solution for the Boundary Problem remains untested with respect to both performance and biological plausibility (see Section 1.2.2.1 for an overview of ST) [82]. Implementing and testing the ST components proposed to handle the Boundary Problem is the focus of this work, as outlined in Section 1.3.

1.2 Review of Literature

The review of related literature will focus on three main topics related to this work: psychophysics and neurophysiology in Section 1.2.1, computational models of attention in Section 1.2.2, and active vision in Section 1.2.3. Section

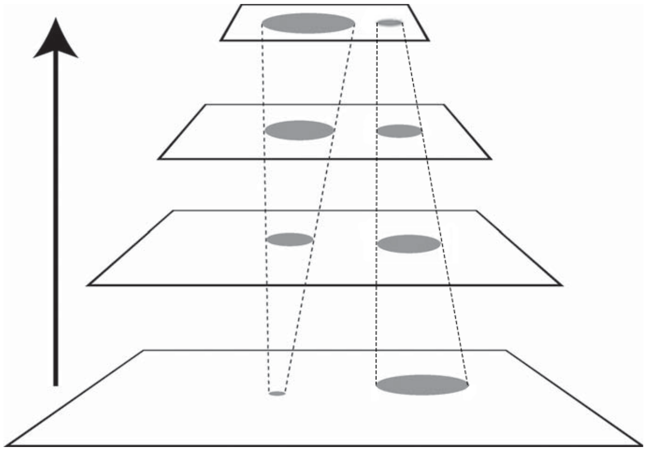


Figure 1.1 – Abstract hierarchy demonstrating the widening cone of connectivity both upward from a small region of neurons in the initial layer (on the left) and downward from a similarly small region of layers in the top-most layer (figure modified from [82]).

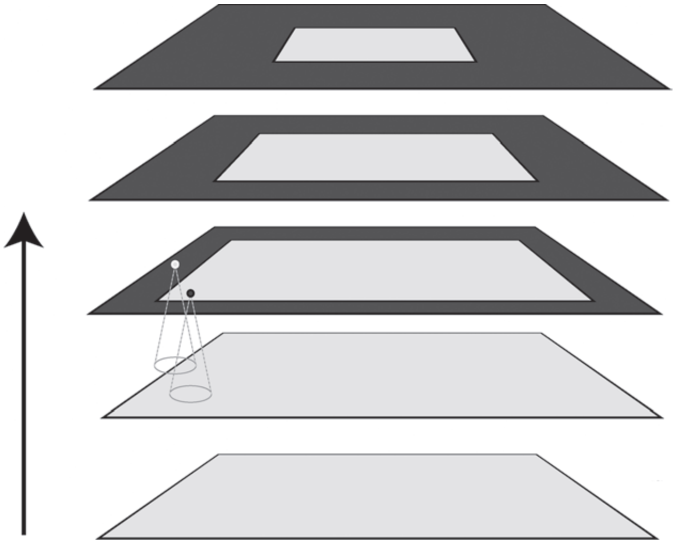


Figure 1.2 – Boundary Problem for pyramids: each progressive layer in the hierarchical process increases the area which is undefined due to boundary effects (figure taken from [82])

1.2.1 will provide an overview of biological research in visual search with a focus on research in peripheral vision, inhibition of return, and covert versus overt attention. Section 1.2.2 will concentrate primarily on saliency map models of attention, but will also provide an overview of the Selective Tuning model of attention. Section 1.2.3 will examine applications of active vision to computer vision and robotic systems, with an emphasis on visual search tasks.

1.2.1 Psychophysics and Neurophysiology

Given the impressive visual capacity apparent in many animals, it is worthwhile examining the biological literature for a more thorough understanding of the capabilities and limitations of biological visual systems. This review will concentrate primarily on human visual processing, but will occasionally make use of results from non-human animal studies, which will be explicitly noted as such.

1.2.1.1 Visual Search

As mentioned in Section 1.1, visual search is a well-researched subproblem useful for investigating the nature of visual attention. Most psychophysical studies in visual search make use of *speeded response* tasks in which subjects attempt to locate a target in an image (or, occasionally, determine its absence) as quickly as possible. Early characterizations of visual search separated the task into *parallel* and *serial* regimes. This separation was based on whether the response time of a subject to find a single target within a field of distractors stayed constant or grew linearly with increasing numbers of distractors, respectively. A highly influential model that attempted to explain this behaviour was the *feature-integration theory* of Treisman and Gelade [80]. The feature-integration theory hypothesized that there exists a set of separable visual features, such as colour and shape, and that visual search can be accomplished nearly immediately (the

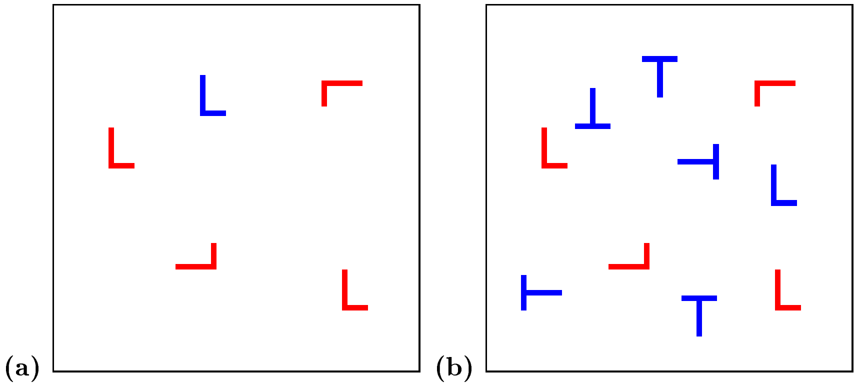


Figure 1.3 – A set of example search tasks. **(a)** displays an easy search task to locate a blue L amongst red L distractors, whereas **(b)** displays a more difficult conjunction task of finding a blue L among red L and blue T distractors.

whole image is processed in parallel) if the target can be differentiated from the distractors according to a single feature. An example of such a search would be finding a blue letter L amongst a set of red L's (see Figure 1.3(a)). When viewing such an image, the blue L leaps out at the viewer in a phenomenon known as *popout*. According to feature-integration theory, when more than one feature is required to differentiate between the target and distractors (for example, finding a blue L amongst red L's and blue T's as in Figure 1.3(b)) a serial search is employed which must investigate each element of the visual field in turn.

Although feature-integration theory did explain many of the early psychophysical results, the use of terms such as parallel and serial were perhaps premature, as they assumed a physiological processing mechanism which still has yet to be conclusively demonstrated. More troublesome, some conjunction searches were found that could be performed with a near zero slope of response time versus distractors [78]. Additionally, as more research was performed in the area of visual search, it was realized that the slope of response time versus number of distractors for single-target problems followed a continuous spectrum rather than a dichotomous parallel versus serial separation. In response,

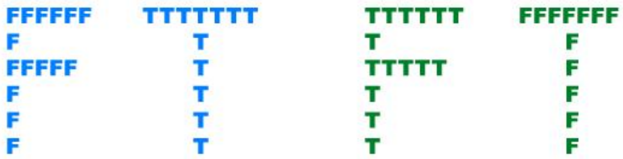


Figure 1.4 – Example of an image elements which can be perceived at both global and local scale. The large letters (global scale) are formed by smaller letters (local scale). The global and local scales can match, as with the letters on the left in blue, or contradict, as with the letters on the right in green.

Wolfe proposed shifting the vocabulary to describe the search *efficiency* [91]. Easily accomplished tasks such as the one depicted in Figure 1.3(a) would be labeled efficient, whereas more difficult tasks with a steep slope would be labeled as increasingly inefficient. Despite the change in terminology, Wolfe does not abandon the idea of a set of basic features. Wolfe defines basic features as stimuli that support both efficient search and effortless texture segmentation, from which he develops the following list of features: colour, orientation, curvature, vernier offset, size, spatial frequency, scale, motion, and shape. Of the set, Wolfe ruminates that spatial frequency and size might be the same basic feature, while scale, though related to size, is probably not identical to it and refers to the scale at which a scene is examined. For an example of differing image scales, see Figure 1.4.

Shape is also acknowledged by Wolfe to be a problematic basic feature category. Although numerous experiments exist which utilize shape features that are not reducible to orientation and curvature (for example, searching for O's amidst X's in [78]), the primitives of shape perception and the layout of a shape space is unclear. While other features such as colour lie within clear two dimensional (or three dimensional, if luminance is included) spaces with definable metrics (even if the precise axes of the space are not always agreed upon), it is not clear what sort of space might house shape.

In addition to mean response times, more recent psychophysical efforts in single-target visual search tasks have sought to characterize the distribution over response times by gathering very large data sets, [57, 92]. By using approximately 112,000 total trials from three visual search tasks (an efficient colour feature search, an intermediate colour and orientation search, and an inefficient search for a digital 2 among 5s), Wolfe, Palmer, and Horowitz are able to create a large data-set for statistical analysis. Using the shape of the distributions over the data set, they argue against the standard self-terminating model. This is based on both the extensive overlap between target present and target absent distributions and the apparent independence of normalized distribution shape to set size [92]. Palmer, Horowitz, Torralba, and Wolfe return to the same data set and attempt to fit a number of statistical distributions. They find that the ex-Gaussian, ex-Wald, and Gamma distributions[1] provide highly comparable fits to the data that are distinctly closer than those achieved by the Gaussian distribution which is usually implicitly assumed [57].

Most single-target visual search tasks yield comparable response times when the distractor and target are inverted. For example, it is generally as efficient to search for a red target amongst green distractors as it is to search for a green target amongst red distractors. However, there are some target-distractor pairs for which this is not true. Such pairs exhibit *search asymmetries*. Although it is beyond the scope of this work to make a comprehensive list of examples, Wolfe notes a few that are repeated here for illustrative purposes [91]. In colour searches, it is more efficient to find a magenta target amongst red distractors than it is to find a red target amongst magenta distractors. Likewise, in orientation searches, it is easier to find a slightly oblique line amongst vertical

[1] An ex-Gaussian distribution refers to an exponential distribution convolved with a Gaussian distribution, while an ex-Wald distribution refers to an exponential distribution convolved with a Wald (also known as an inverse Gaussian) distribution. For a more complete discussion of the formulation of these distributions and the manner in which they fit psychophysical data, see [72].

distractors than it is to find a vertical target amongst slightly oblique distractors.

Although single-target visual search tasks are common due to their methodological simplicity, Thornton and Gilden argue that the results of such studies are necessarily inconclusive regarding the serial or parallel nature of processing in visual search [79]. Instead, they propose to extend the single-target method to a generalized visual search task which can have multiple targets present. The task remains for subjects to simply respond as quickly as possible whether or not a target is present. A principle idea of multiple-target search tasks is to include occasional instances in which all objects in a visual field are targets. If processing is serial, the response time to all-target trials will obviously remain constant with respect to set size. In contrast, a limited capacity parallel process might display improved performance due to *redundancy gain*. Thornton and Gilden tested 29 different search tasks and found that most exhibited redundancy gain, but a small number of tasks exhibited the constant response rate which the authors associate with serial processing. They therefore argue that the efficient-inefficient continuum advanced by Wolfe does not exist, and the serial-parallel dichotomy originally proposed is correct (albeit with many of the tasks originally associated with seriality being operated on by a limited-capacity parallel process and with possible fuzzy boundaries between the two protocols). Thornton and Gilden do not address the possibility, however, that all the tasks are solved with a limited capacity parallel process and the redundancy gain in some of the trials is simply offset by dividing attention across multiple targets. The effects of divided attention may differ with the task, thus resulting in the spectrum of redundancy gain observed.

In addition to examining the capabilities of subjects to find targets in visual search tasks, it is also interesting to examine whether there are any targets that *must* be found. In a review of several studies which have sought to answer this question, Yantis concludes that singleton elements which are capable of producing a pop-out effect will only sometimes automatically capture attention, depending on the task [93]. When singletons are entirely irrelevant to the task, such as when searching for a specific letter amongst an array of many different distractor letters, it was found that they could be suitably ignored. This was true regardless of the nature of the singleton, whether it was a unique colour, [36], or even a form of motion, [30]. However, when subjects are themselves searching for a singleton target, it appears that unrelated singletons can automatically trigger attentional capture, disrupting the efficiency of the search. Interestingly, there appear to be asymmetries in the capacity for an unrelated singleton to preferentially capture attention. Theeuwes found that a unique-form search (searching for a green circle amongst green squares) was disrupted by a colour singleton (a red square), whereas a unique-colour search (searching for a red circle amongst green circles) was largely unaffected by a unique form singleton (a green square) [77]. For further discussion of attentional capture, see the end of Section 1.2.1.2.

Many of the psychophysical aspects of visual search reviewed here are beyond the scope of the present thesis research. Nevertheless, it is hoped that this brief overview of visual search experimentation and theory will help to frame the current work, as well as point to future extensions and experimentation of the developed system.

1.2.1.2 Peripheral Vision

A distinct characteristic of human vision is the rapid decrease in visual processing quality as one moves away from the centre of the visual field. This is partly due to the physical construction of the retina itself. As early as 1935, Østerberg established the heterogeneous distribution of photoreceptors in the human retina [72]. Cones, which come in three forms for use in chromatic vision and require bright light to function, are concentrated in high numbers near the central portions of the retina and then fall in number with increasing eccentricity. Rods, which have only a single form and therefore provide achromatic vision and which function in dim lighting, are entirely absent in a small central region of the retina and instead dominate the periphery of the retina. In addition to the distribution of the photoreceptors themselves, the ganglion cells which receive input from the photoreceptors and transmit the signals on toward the brain are themselves heterogeneously distributed through the retina with a strong central bias [17]. There are two more distinct physical regions of the retina which affect vision: the blindspot, which is a region devoid of photoreceptors in order to allow passage of blood vessels and the optic nerve located approximately 15° nasally, and the foveola, which is a region located near the centre of the retina free of blood vessels and ganglion cells in order to allow maximal light to reach the photoreceptors [88].

When discussing vision in terms of the retinotopic coordinates, it is important to define the vocabulary which will be used, as terms such as *peripheral* and *central* are notoriously ambiguous. Wandell defines the *fovea* as having a diameter of 5.2°, with a rod-free central area of 1.7°, and the *foveola* with a diameter of 1° [88]. These values are supplemented in Strasburger et al.'s review with values for the *parafovea* ($\sim 5^\circ - 9^\circ$) and the *perifovea* ($\sim 9^\circ - 17^\circ$), which altogether form the *macula* [72]. Because the fovea is often functionally defined by its rod-free component, this work will use the term *foveal vision* in the same manner as Strasburger et al. to roughly mean the central 2° of the visual field. For the purposes of this thesis, *central vision* will be defined to refer approximately to everything encompassed by the fovea and parafovea, $< 10^\circ$, while *peripheral vision* will refer to anything outside of this eccentricity rather than Strasburger et al.'s use of anything non-foveal.

In psychophysical examinations of peripheral vision, an early form of investigation of peripheral vision was through the measure of minimum discernible size, either through recognition of letters or differentiation of pairs of dots. Strasburger et al., [72], report that the earliest investigation of this kind was performed by Aubert and Foerster in 1857. They found that the minimum discernible size is proportional to the maximum eccentricity angle within the macula, and decreases at a faster rate outside of this region.

Physiological evidence suggests that this performance difference is not solely the responsibility of photoreceptor and ganglion cell distribution in the retina, but is rather accentuated throughout the visual processing stream based on the degree of cortical area devoted to a particular region of the visual field. Experimenting in monkeys, Daniel and Whitteridge developed the concept of the cortical magnification factor, M, to represent the diameter in the primary visual cortex onto which 1 deg of the visual field projects [18].

In humans, the value of M correlates well with psychophysical results for low-level tasks, but fails to capture performance differences for a number of more complex tasks such as the recognition of faces. Additionally, peripheral visual processing appears to be much more vulnerable to crowding effects, requiring a critical free space between flanking letters of approximately half the eccentricity of the target to avoid any performance deficits [72]. In a series of experiments, Engel defined a *conspicuity area* in which a target stimulus could be identified within a dense field of distractors in a single fixation [21]. Similar to visual search efficiency, Engel found that the size of the conspicuity area varied depending on the degree of 'conspicuity' between the target and the distractors. Targets which were highly similar to the distractors (a horizontal line amongst lines of various random orientations) needed to be within approximately two degrees of the fixation point, whereas highly dissimilar targets (a square amongst randomly oriented lines) could be located within a large ellipse $\sim 25° \times 15°$ in size. As is common to many aspects of visual processing, there is a distinction between the capacity to perform detection tasks versus more the more difficult task of recognition. Strasburger and Rentschler extrapolated a *maximum field of recognition* at $\sim 46° \times 32°$ beyond which stimuli could not be recognized regardless of size (though they could still be detected) [71].

Interestingly, although cortical magnification of the central visual field is consistent through most visual processing regions, there is at least one physiological region in which this is not the case. The visual area V6, located on the anterior bank of the parieto-occipital sulcus, instead strongly emphasizes the peripheral visual field [22]. Originally identified as area PO in the macaque, Colby *et al.* claimed that the central visual field was entirely lacking in representation within the region [13]. More modern studies conducted in both macaques and humans suggest that the area identified as PO was actually a subsection of both area V6A, a visuomotor area superior to V6, and V6 such that the full V6 region does have marginal central field representation (see Figure 1.5) [22]. Area V6 is predominantly connected to areas within the so-called "dorsal pathway" of visual processing, and appears to respond most strongly to motion stimuli. Although current understanding of the dorsal and ventral processing streams shows that the two are not wholly distinct, the visual processing in regions along the dorsal pathway is nevertheless associated more with fast responses to visual information involved with muscular action. This would be consistent with the task of active visual orienting, and it therefore makes sense that such an area would place a heavy emphasis on processing the peripheral visual field. Although current research exploring lesional deficits in the area of V6 have concentrated on disruption to patients' capacity for motion characterization, I hypothesize that

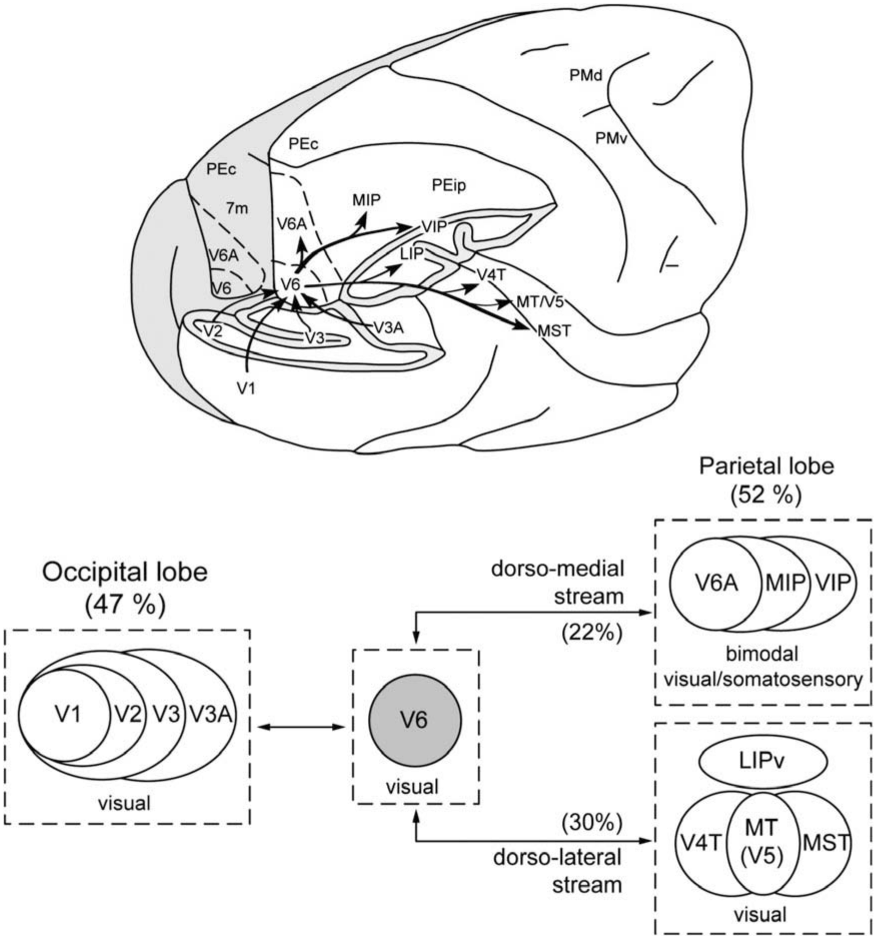


Figure 1.5 – Diagram of V6 connectivity in the macaque brain. The upper image summarizes the cortical connections, with the occipital pole and part of the inferior parietal lobule removed from the right hemisphere to show the cortex region hidden in the parieto-occipital and intraparietal sulci. The left hemisphere is shown in grey without any regions removed. The block diagram on the bottom summarizes the weights of the cortical connectivity to area V6 (note that the occipital lobe connections are also comprised of neurons predominantly part of the dorsal stream). Image reproduced from [22].

the ability to orient attention based on exogenous peripheral cues (both dynamic and static, see below) should likewise be impaired.

In Section 1.2.1.1 the question of whether certain stimuli must capture attention was addressed for salient visual elements capable of producing a pop-out effect. However, the capacity for other types of stimuli to automatically capture attention was not addressed. John Jonides performed an early set of experiments into this question using the qualities of capacity demands, resistance to suppression, and sensitivity to expectation in order to determine how automatically stimuli captured attention [35]. Jonides used two types of stimuli in his experiments: stimuli which appeared outside the central fixation area (*exogenous* cues), and arrows which appeared at the central fixation and pointed toward a target area of the visual scene (*endogenous* cues). Jonides found that exogenous cues reliably captured attention even when subjects were actively told to ignore them, whereas arrows endogenously presented at the central fixation point could be capably ignored. This finding was extended by Yantis and Jonides, who showed that the abrupt onset of stimuli preferentially captures attention, and is capable of doing so even when actively suppressed [94]. Interestingly, it was noted that the automatic capture of attention occurred over a very fast time course, whereas attention allocated by endogenous cues was allocated more slowly. This difference in time course suggests the possibility of multiple attentional forms depending on the nature of the signal, and will be discussed more fully in the next section.

1.2.1.3 Overt Versus Covert Attention

Although visual fixation is often used as a surrogate for the allocation of visual attention (see Section 1.2.2.3 for further discussion on this matter), evidence for attention independent of eye movements nevertheless has been around since Helmholtz's comprehensive psychophysical experiments in vision in the early twentieth century. Helmholtz demonstrated that when he fixated his eyes on a particular visual location in a dark room, he could nevertheless explore the contents of an alternative region of his visual field following a brief flash of light [86]. When he attempted to continue exploration of his memory of the flashed stimulus, however, he was unable to reconstruct detail outside of the attended region, suggesting that there was something special about his focus on that region. The orientation of attention independent of physical eye movement has subsequently become known as *covert attention*, while attention through eye fixation is known as *overt attention*. The existence of covert attention obviously poses a difficulty when trying to evaluate a model's performance based on comparison to eye tracking data, and it is worth asking whether these two forms of attentional assignment are two aspects of the same mechanism or whether they are distinct.

Hunt and Kingstone, [32], attempted to determine the independence of covert and overt attention through the performance of subjects on an eye movement and a discrimination with fixed gaze task. By splitting subjects into two groups and asking them to concentrate on either the eye movement or discrimination task, Hunt and Kingstone found that subjects were able to display increased speed and accuracy in validly cued trials only for the task being concentrated on, and thus concluded that the two attentional mechanisms operate independently. In contrast, Hafed and Clark, [29], posit a connection between the motor systems and covert attention by using microsaccades to predict covert attentional shifts. It should be noted that Hunt and Kingstone provided endogenous auditory cues for both trial types in their study, whereas Hafed and Clark used exogenous peripheral flashes of light. Thus, it may be that covert and overt attentional mechanisms are distinct under top-down direction, but the motoric control of the eyes is reflexively activated by exogenous cues (see the end of the previous section). In fact, subjects in Hafed and Clark's study had a tendency when first performing the task to make overt saccades to peripheral cues, and had to actively learn to suppress foveating saccades, suggesting the possibility of automatic oculomotor recruitment.

Although Hafed and Clark's results might be explained by reflexive recruitment of the oculomotor system, van der Stigchel and Theeuwes also found evidence that endogenously cued covert attention allocation could influence subsequent unrelated saccades [84]. In their experiment, subjects fixated on a central location and were cued via a central arrow to attend covertly to a location in either the upper left or upper right portion of the screen. Following the cue, either an E or an S would appear at that location, which the subjects were directed to differentiate with a button press. On a small subset of the trials, however, an audio cue was played following the stimulus display to inform the subjects to not complete the differentiation task, but to instead execute a saccade straight up. It was found that subject saccades had a significant deviation away from the cued stimulus, suggesting that the endogenous allocation of attention influenced the later motor execution despite the spatial and perceptual independence of the two locations.

It is possible that overt and covert attention are not two different systems, but rather a manifestation of the default attentional focus at the foveal region. Attention can be consciously decoupled from this location and allocated elsewhere (albeit with diminishing returns as one moves farther peripherally, see Section 1.2.1.2), but this requires active suppression of the motor system (as can be seen by the initial challenge of subjects in Hafed and Clark's study to avoid making overt saccades to peripheral cues). Thus, inhibition of the motor system would lead to the deviation seen in van der Stigchel and Theeuwes, but would leave the oculomotor system and attentional system apparently decoupled during excitatory tasks (as seen in Hunt and Kingstone). The work here provides a first approximation of this interpretation, with a central region in which attention can be allocated (with default allocation after a saccade in the centre) and automatic recruitment of the motor system for peripheral targeting. Future extensions could allow for greater top-down control over whether or not a saccade is performed.

1.2.1.4 Premotor Theory of Attention

An early attempt at describing the purpose and action of biological attention has come to be known as the *premotor theory of attention.* Originally put forth by Klein as the *oculomotor readiness hypothesis*, [39], it suggests that attention is an outcome of preparing to move the eyes to foveate a target region. Therefore, the primary mechanism of attention is the motoric component, and the enhanced processing capacity is an outcome of preparing eye movements. Klein himself rejected the hypothesis after failing to find either facilitation of detection of stimuli at a location to which a saccade had been readied, or the facilitation of eye movements to a covertly attended location [39]. Rizzolatti *et al.* revived Klein's hypothesis under the name premotor theory of attention, after rejecting Klein's findings based on the dual nature of the tasks Klein's experimental participants engaged in (which, it was argued, forced participants to wait until the task was cued to prepare motor commands, preventing the expected facilitation) [64]. The primary evidence presented for this initial revival was an unexplained increase in the response time to stimuli which required a crossing of the horizontal or vertical meridian of the visual field. Using the meridian effect, which does require explanation, as evidence for the premotor theory of attention, however, is based primarily around the assumption that saccadic programming changes which involve crossing a meridian line are more time consuming than those which do not due to the recruitment of a different set of muscle groups. While this may in fact be a valid assumption, preliminary psychophysical evidence does not support it; uncertainty in the direction of a saccade but knowledge of the expected amplitude led to virtually identical saccade latencies as for those in which the direction was known but the amplitude was unknown [3]. Also puzzling with regards to the meridian effect is that, while it is consistently found with endogenous attentional cuing, it was not obtained in studies using exogenous cuing [16, 63].

Two other lines of evidence are somewhat more promising for the premotor theory of attention: saccadic deviations and attentional deficits that mirror oculomotor deficits. Sheliga *et al.* conducted a series of experiments in which subjects allocated attention to the left or the right of a fixation point, and then were asked to make a saccade either upward or downward [68]. It was found that saccades deviated away from a locus of attention both when attention had been attracted to that point exogenously as well as endogenously. Stigchel and Theeuwes replicated this finding with endogenous cuing, extending the finding to show that saccadic deviation occurred away from the locus of covert attention even in the case of an invalid cue [84]. Although these studies do suggest a link between attentional allocation and motoric output, they do not provide any evidence as to the direction of that link; it is entirely possible that the motoric affect is a result of the attentional allocation.

Craighero *et al.* performed an experiment with subjects suffering from peripheral VI nerve palsy leading to palsy in the rectus lateralis muscle of either the right or left eye, disrupting horizontal saccades with that eye [14]. Subjects were evaluated in monocular trials and given endogenous cues to either left or right target positions which were accurate in 70% of trials, followed by a target onset in one of the two target positions. It was found that when either healthy subjects or those suffering from palsy used their healthy eyes, the standard difference in reaction times for valid and invalid trials was obtained. However, when subjects performed the task with their paretic[2] eyes, there was no statistical difference in the reaction times between valid and invalid trials. Curiously, both valid and invalid trials yielded reaction times comparable to the reaction times of valid trials for healthy eyes. A later paper by Craighero *et al.* sought to replicate these findings, this time in an entirely healthy cohort in which motor control was instead constrained by forcing the starting fixation to occur in an already eccentric temporal ocular position [15]. In this case, reaction times to temporal stimuli for the eccentric starting position were not statistically different for valid and invalid trials, and were on the whole comparable to the invalid response times to stimuli in the other experimental conditions. Although both studies do indicate that a motoric restriction affects response times, it is odd that one study found all response times to be comparible to validly cued trials, suggesting an *increase* in response time due to motoric disruption, while the other found all response times to be comparible to invalidly cued trials, suggesting the more expected result based on the premotor theory of attention of a disruption to attentional allocation following a disruption in motoric control.

Thus, although it is clear that there is a definite connection between some attentional phenomena and motoric control of gaze, it does not seem that attention can be entirely explained as simply a result of motoric preparation. In fact, when Klein and Pontefract modified Klein's original methods which rejected the oculomotor readiness hypothesis to take into account Rizzolatti *et al.*'s criticisms, they again found evidence to disconfirm the hypothesis [42].

1.2.1.5 Inhibition of Return

First demonstrated by Posner and Cohen in cue-fixation tasks in 1984, [60], Inhibition of Return (IOR) is a delay in response to a previously attended lo-

[2] Paresis refers to a condition of partial paralysis.

cation of the visual field. Klein later extended these findings to show that IOR occurs following attention allocation in visual search tasks as well, concluding that IOR functions as a facilitator for foraging behaviour [40]. Although Klein's interpretation was initially challenged by subsequent studies, these challenges have themselves since been rebutted (see Box 2 of [41] for a full discussion), indicating that IOR is indeed an important component in the cognitive toolkit used to tackle the visual search problem. As will be discussed further in Section 1.2.2, a simple IOR mechanism was an important component in early saliency models of attention in order to keep the system from becoming permanently fixated on a single locus of maximal salience or oscillating between two maxima. Psychophysical studies have demonstrated, however, that IOR behaviour is highly variable depending on the specific task factors involved. While it is beyond the scope of this thesis to model all aspects of IOR, they are nevertheless reviewed here for completeness and possible future extensions. For further discussion of how IOR has been factored into this implementation, see Sections 1.3.2 and 2.3.

Posner and Cohen's initial findings showed that a fixational cue improved subject reaction times when the interval between the cue onset and target onset (stimulus-onset asynchrony) was short, and that this facilitation decayed until it crossed over into an inhibitory effect after a stimulus-onset asynchrony of between 200 and 300 ms, [60]. When Posner and Cohen replaced the exogenous peripheral cues with an endogenous central cue (an arrow at the central fixation which points in the direction of the target location), however, they did not elicit the inhibitory effect, generating some confusion as to the source of the inhibition. Subsequent work was seen to confirm and refine this finding, demonstrating that it was enough for a saccade to be prepared to elicit IOR, even if central fixation was maintained throughout the cuing process, but when an endogenous cue was displayed without saccade preparation IOR was not seen, [62]. The conclusion of these early studies into the nature of IOR was therefore that it was primarily a motoric component of eye control.

The motoric interpretation of IOR was additionally bolstered by perceptual experiments in which subjects were presented with a peripheral cue, a return cue at fixation, and then two peripheral targets in rapid succession. When subjects were instructed to make a saccade to whichever target location felt “more comfortable”, they had a significantly increased tendency to direct their gaze to the uncued location [61]. In order to check whether this bias represented a perceptual or motoric delay, subjects were additionally asked to make a temporal order judgment of which target was the first to appear. Despite the bias against fixation of the cued target, there was no effect on the subjects’ ability to determine the order in which the targets appeared [61]. Further study by Terry *et al.* confirmed IOR in simple detection and localization tasks, but failed to find IOR in non-spatial discrimination tasks which relied on target form, colour, or size [76].

A non-motoric and distinctly attentional manifestation of IOR was finally discovered when the time course under investigation was expanded. Lupiáñez *et al.* found that IOR in response to a non-spatial discrimination task appeared in the $700 - 1000$ ms range, much later than the IOR response previously demonstrated in detection and spatial localization tasks [48]. These results were replicated in other studies and compiled by Klein (see Figure 1.6), displaying that task difficulty, as represented by response time, was linearly proportional to the onset time of IOR [41]. Klein additionally points out that there are a large number of other psychophysical factors which can impact experimental results, stating:

> Thus, when IOR is not obtained it could be that it was: (1)

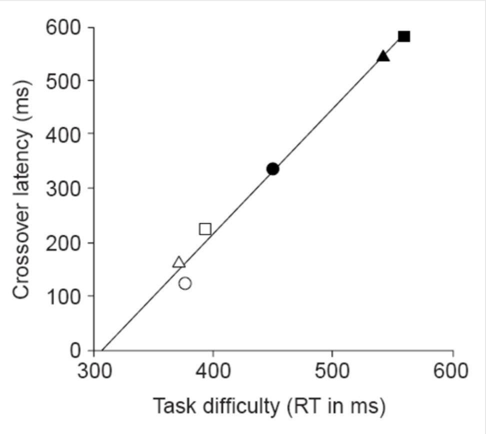


Figure 1.6 – Plot of the cross-over points, measured in stimulus onset asynchrony, at which facilitation turns to inhibition verses reaction time to the target with data from three experimental setups of paired tasks. The figure is reproduced from [41]. Circles represent data for localization tasks from [7]*, with open circles representing saccadic localization and filled circles manual localization. Triangles represent data from [47]*, with open triangles representing manual detection and filled triangles discrimination. Likewise, squares represent manual detection and discrimination (for open and filled, respectively) data from [48].

> present but the task used to measure it was not sensitive to the inhibition; (2) present and measurable, but obscured by an accompanying effect in the opposite direction [such as facilitation or a response-repitition advantage]; or (3) not present in the first place. Discriminating amongst these alternatives is one challenge to researchers of IOR.

An important aspect of IOR is not only when it begins, but also the duration of the inhibitory effect. Evidence suggests that the duration of IOR is not a straightforward property, and is instead affected not only by temporal components but also by the nature of the task and the number of inhibited regions. In the previously mentioned experiment by Lupiáñez *et al.* in which non-motoric IOR was identified, it is suggested that although the non-spatial discrimination task had a much later onset of inhibition than the simpler tasks of spatial detection and localization, the cessation of the inhibitory effect occurred began at nearly the same point for all tasks and thus the more difficult task had a shorter overall duration of inhibition [48]. When Wang and Klein compiled temporal duration data more systematically across several studies, they again found a more complex story [89]. In a complicated virtual reality task in which each "fixation" took approximately 1700 ms, significant IOR was seen for up to two previous fixations, suggesting a duration on the order of 3400 ms. In contrast, in a scene search task with average inter-saccade time of 254 ms, IOR was seen at positions two fixations previously and four fixations previously, but not at six, thus suggesting inhibition in scene-search is either limited to maximal duration of less than 1524 ms, or is limited to fewer than six inhibited regions. The fact that IOR is capable of being applied to more than one previous fixation provides evidence for an environmental encoding of IOR location rather than a retinal encoding, as a retinal encoding would shift inhibitory localities with a saccade. More recent evidence suggests that IOR is not only encoded in environmental coordinates, but may in fact be scene- or object-dependent. In several covert probe-following search experiments reviewed by Wang and Klein, it is reported that the effects of IOR are only seen if the search scene is maintained, but disappear if the scene is removed between fixations [89]. One such study in particular, [56], demonstrated an IOR effect attached to objects moving smoothly on the visual display. Importantly, these scene- and object-based IOR effects were only noticed in difficult search tasks, and in fact the object-based IOR was not seen when the visual search task was efficient.

Though stability of a visual scene appears necessary for the persistence of IOR, the question nevertheless arises as to whether IOR persists in a stable environment for fixations which move beyond the bounds of the visual field through subsequent saccades and head movements. Although I am not aware of a study showing such persistence of IOR explicitly, there is nevertheless evidence for the persistence of visuospatial information outside the bounds of the visual field which indicates that this should in principle be possible. Tark and Curtis provided aural cues to subjects using microphones placed in their ear canals, with inter-aural sound differences designed to control cue spatial location [73]. Functional magnetic resonance imaging examined the activity of the frontal eye fields (FEF), and found that subjects displayed persistent FEF activity in response to an aural cue even when that cue was located behind the head, suggesting that the FEF represents both retinal and extra-retinal space.

In addition to characterizing the behavioural effects of IOR, there have been numerous studies that have sought neurophysiological evidence for the source of the inhibitory signal. As with the psychophysical characterization of IOR, initial neurological investigation focused on a mid-brain motoric involvement. Individuals with damage to the superior colliculus were found to have drastically reduced or no IOR, whereas a hemianopic[3] patient with cortical damage but an intact colliculus exhibited IOR to cues located in his blind field [41]. However, subsequent studies showed patients with parietal damage (both with and without spatial neglect) were found to exhibit unaltered IOR on the contralesional side, but reduced IOR or even facilitation on the ipsilateral side. Similarly, object-based IOR was found in split-brain patients provided the object remain within the same hemifield but was eliminated once the object crossed the midfield boundary [46]. In light of the sometimes contradictory evidence for the nature of IOR, Taylor and Klein postulated that IOR comes in two distinct "flavours" rather than consisting of a single cognitive tool: a motoric flavour when the oculomotor system is engaged and an attentional and perceptive flavour when the occulomotor system is quiescent [75]. Taylor and Klein's two flavours of IOR could alternatively be interpreted as IOR affecting overt at-

[3] Hemianopia refers to defective vision or blindness in half of the visual field.

tention and IOR affecting covert attention, though to my knowledge this interpretation has not been explicitly explored. Houghton and Tipper, in contrast, do not limit themselves to two flavours, but rather postulate that inhibition following temporally from a neurological event is a general cognitive control strategy which happens to manifest in visual search as IOR [31]. Whether such a general interpretation of inhibitory feedback is warranted remains to be seen, but it does appear to explain many of the complexities that arise when trying to treat IOR as a singular phenomenon and will therefore influence the designs of the IOR mechanism included in this thesis, as described in Section 2.3.

1.2.2 Computational Models of Attention

Despite the great complexity of nuance involved in biological attention of which Section 1.2.1 was only able to briefly touch upon, a number of attempts have been made at succinctly formulating attention in a well-defined theoretical framework. Some of these models will be discussed here, but, as with Section 1.2.1, the size of the field is beyond the scope of this thesis to adequately cover. For a broad historical discussion of how attention is defined such that it may be studied, see the opening chapter of [82].

1.2.2.1 Selective Tuning

Selective Tuning (ST) is an attention model developed over many years which grew out of Tsotsos' work on the complexity of the vision problem and his conclusion that an attentional mechanism was required in order to make vision tractable [81]. A key aspect of the ST model is that it takes a predominantly first principles approach to modeling attention; the model seeks first and foremost to describe in general terms what aspects of attention are required from a computational perspective. Although this section will attempt to give an overview of ST, it is a complex model which has been under development for over two decades, and thus many of the details of this development are beyond the scope of this thesis. For a comprehensive overview, see Chapter 4 of [82], and see the subsequent chapters for an up-to-date exposition of the model's details.

ST utilizes a hierarchical pyramidal processing architecture; neuronal layers increase in abstraction and complexity of selectivity as one moves higher in the pyramid. Network neurons are modulated by task considerations, and decisions (or visual problem solutions) are made in a competitive manner based on response values throughout the processing network. A distinct characteristic of ST which separates it from other population coding schemes is that the representation of a visual item (whether it is an object, scene, or event) involves the entire network pathway which is found to best respond to it, rather than just some high-level subset of neurons in the hierarchy. This allows an ST representation to easily shift between high level abstractions (such as object category) to lower level details (such as surface texture or colour).

As mentioned in Section 1.1, a number of the algorithmic details of ST are designed to address the computational challenges introduced by a pyramidal hierarchy: the Routing, Context, Blurring, Cross-talk, Sampling, and Boundary Problems (see Chapter 2 of [82] for a detailed description of these problems). ST utilizes a top-down, recursive pruning strategy for selecting the best representation within its network based on the classic branch-and-bound mechanism, [45]. Such a search strategy has tractable computational complexity and ameliorates all but the Boundary Problem. Representation at all levels of the network naturally links the highest levels of abstraction with the lowest levels of the hierarchy, thereby addressing the Routing Problem. Through the pruning process competing but less desirable representations are removed. This creates an inhibitory effect which isolates a given representation from surrounding context, and firmly localizes the representation via its activation at the lower levels of the hierarchy to solve both Sampling and Blurring issues. By allowing the system to hold only one such representation at a time (there is only one locus of attention), Cross-talk is prevented by the same inhibitory isolation of the representation from its surround.

Due to boundary constraints, the full hierarchical structure of ST may only be defined over some central portion of the visual field. Thus, the system described so far corresponds most approximately with attention operating in the central regions of the visual field; it is in principle capable of performing covert allocations of attention throughout this central region in response to endogenous as well as exogenous cues. The method with which the ST model deals with the periphery is the focus of this thesis, and thus will be dealt with more completely in Section 1.3 and Chapter 2.

1.2.2.2 Saliency Map Models

Saliency is a metric for measuring the capacity of a region in the visual scene to capture the attention of a viewer. It is often intuitively equated to "how interesting" is a given region of a visual scene. Virtually all models which make use of saliency aim to generate a single overall *saliency map*, which assigns a conspicuity value to every location within a visual scene. Such a map can then be quickly and efficiently analyzed using a Winner-Take-All (WTA) computation to generate the maximally interesting point in the image and direct the system's attention to that location. Many salience map models possess mechanisms for biasing the generation of salience values according to top-down goals (for example, favouring red targets), but they are still primarily bottom-up signal-driven methods.

The most famous saliency map model is that of Itti, Koch, and Niebur (referred to hereafter as SM for ease of reference) [34]. The model is based heavily on Koch and Ullman's architecture for attention selection, [43], which is itself closely linked to the Feature Integration Theory of Treisman and Gelade, [80]. SM takes colour images as input and runs them through a Gaussian pyramid to create a multi-scale representation of the image. Across-scale differences are taken to perform centre-surround operations, creating a set of feature maps based on different intensity scales, red-green and blue-yellow colour double-opponency, and orientation feature maps created from Gabor filters applied to the intensity pyramid. In the absence of top-down direction, each feature map is normalized by performing the following operation:

1. Normalize the values in the map to a fixed range $[0, ..., M]$
2. Find the map's global maximum M and compute the average $\bar{m}$ of all other local maxima.
3. Globally multiply the map by $(M - \bar{m})^2$

Normalization serves to smooth out feature maps as well as reduce the impact of maps which respond well to large portions of the scene while accentuating distinct feature peaks. After normalization, feature maps are combined into three conspicuity maps derived from the intensity, colour, and orientation channels. The conspicuity maps are then themselves normalized before being averaged together to form the overall saliency map.

The SM model is relatively straightforward in construction, has clear origins in psychophysics, and was one of the earliest formulated saliency map models. It is thus well-known and frequently serves as a benchmark against which the performance of other models can be compared. One of the primary drawbacks of the feature maps used in SM is that they do not yield sparse representations for most targets, thereby reducing the signal-to-noise ratio and degrading the performance of the model on complex natural images. Modern attempts have been made with some success to modify the SM model to improve performance by including elements such as task influences on attention through the use of a symbolic working memory system, [54].

An alternative to feature maps was proposed by Bruce and Tsotsos, who developed a model called Attention based on Information Maximization (AIM) which measures salience based on the information content of the image [9, 10]. In particular, AIM uses the self-information, $-\log(p)$, as the primary measure of salience (see Figure 1.7). The basic intuition behind the model is that the most salient part of an image is that which can least easily be predicted. To achieve this, AIM starts by seeking a sparse code for natural image statistics. This is created by performing Independent Component Analysis (ICA) over a large collection of randomly sampled natural image patches. The set of representative patches found by ICA is used as the feature basis used to represent the visual world, though alternative filters can be used and may even be more desirable for some tasks [8]. In the case of this thesis, log-Gabor filters were used for most tasks (see Section 2.2.1 for a discussion of this decision).

Using a basis set of ICA filters, each local neighbourhood of a test image is projected into the filter space, whereupon each pixel is converted into a vector of values corresponding to the individual contribution of each basis function. The filter space is assumed to be approximately independent, and thus the joint probability of a given pixel vector reduces to the product of the individual probabilities for observing each feature contribution. However, the likelihood of each feature component must still be estimated. Bruce and Tsotsos use a Gaussian kernel density estimate in the original formulation, though they acknowledge that this choice is somewhat arbitrary and could be changed without loss of generality (and, in fact, most subsequent implementations of AIM used the entire image as the basis for this estimate). The kernel is used to compare the

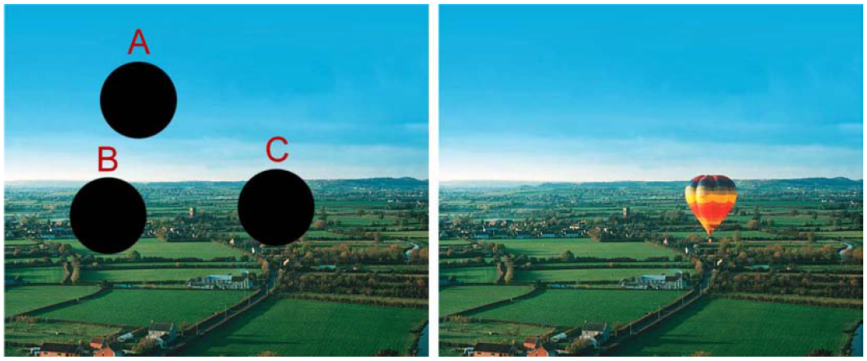


Figure 1.7 – An image displaying an intuitive sense behind the saliency measure of the AIM algorithm. On the left the black circles labeled A and B obscure regions which are visually similar to their surroundings and which would generally be predicted if one were to guess their contents; the region obscured by C, on the other hand, contains an unexpected object which is highly dissimilar compared to the rest of the image content. It is therefore both the most interesting or salient, as well as carries the highest degree of self-information (image reproduced from [11]).

value of a component's contribution for that image patch with the contribution of the same component to the surrounding image patches. Thus, a feature which contributes equally to all image patches in a region will yield a high probability, and therefore low self-information (it is easy to predict the content of that region, and thus the region is less informative). A more detailed description of AIM's operation in relation to this thesis can be found in Section 2.2.1, and of its implementation in Section 3.4.

A similar saliency approach to AIM, called saliency using natural image statistics (SUN), was developed by Zhang *et al.* [97]. SUN was motivated by a Bayesian formulation of saliency which would allow the easy incorporation of both top-down and bottom-up mechanisms. From this formulation, self-information emerged as a natural measure of bottom-up saliency, and could be combined with top-down measures to give an overall saliency measure dubbed pointwise mutual information.

The main difference of the SUN approach is that the probability distributions for each feature are not based on the test image at all, but are entirely based on prior statistics derived from natural images. Although this makes the algorithm quite efficient computationally (all intense computational steps are performed offline during learning) and actually yields results which are decently comparable to AIM for a number of test images, it causes the model to deviate greatly from psychophysical results as it creates search asymmetries for *every* pair of target and distractor. For example, if red is associated with being a more salient feature than green, the model will do a good job finding a red target amidst green distractors, but will perform extremely poorly looking for a green target amongst red distractors. Although search asymmetries do exist, as discussed in Section 1.2.1.1, they do not measurably exist for every target-distractor pair.

1.2.2.3 Saliency Based Models and Human Search Performance

All the saliency models discussed in Section 1.2.2.2 attempt to provide a solution to a visual search task from the perspective of finding what is the most salient element of an image from some objective measure. Although these models all tend to produce results which qualitatively compare to the eye fixation patterns of human observers, they nevertheless have a difficult time predicting specific fixation sequences for a given image. An alternative perspective that has motivated a number of modern saliency models is to instead focus on the prediction of human eye fixations, and worry less about the underlying reasons for those fixations. Such a pragmatic, performance-driven design lends itself to applications in advertising design, video compression, and non-photorealistic image rendering.

Judd *et al.* provide an example of a saliency map approach specifically designed to reproduce human fixation patterns [38]. In order to design such a saliency model, a large database of human fixations was generated for 1003 natural images. This fixation database was then used to create a ground-truth against which the developed saliency model could be compared. The salience model itself was a classifier learned from a mixed set of low-, mid-, and high-level features, as well as a centre prior. Low-level features consist of steerable pyramid filters, the saliency channels of Itti and Koch's SM algorithm, [33], and colour channel statistics. The mid-level features were gist features trained as a horizon line detector. High-level features consisted of the Viola Jones face detector, [85], and the person detector by Felzenswalb *et al.*, [26]. Altogether, a model of visual salience was learned from these features which provided a decent approximation to human fixation records.

One element which all the saliency models of Section 1.2.2.2 fail to capture in human visual search is the temporal tuning that human search undergoes with repeated trials. In their original formulations, AIM, SUN, and SM all fail to account for temporal shifts in visual search performance over repeated trials. Psychophysical evidence suggests that repeated trials of the same search task will lead humans to tune their responses to optimize the signal-to-noise ratio of the target versus the present distractors [55]. In fact, evidence suggests that this temporal modulation of visual search, termed priming of pop-out by Maljkovic and Nakayama, occurred automatically and unconsciously, as the slowest response times were recorded when the colour of target and distractor switched every trial (something which the subjects were consciously aware of but which were nevertheless unable to prepare their visual systems for), [49]. Thus, although the SUN algorithm appears to take too long-term a view on the manner in which particular features influence saliency, it may be beneficial for future work to incorporate at least short-term temporal modulation of salience.

The topic of saliency model comparison with human search performance would be incomplete without recognition of the many distinct challenges to evaluating saliency map models directly against human visual search data. Tatler *et* *al.* provide a strong critique of this practice, identifying five major and problematic assumptions inherent in models of static scene viewing [74]. Of particular relevance here are the fact that eye tracking only tracks overt attentional shifts and provides no insight into covert shifts of attention, and the fact that human subjects find it virtually impossible to eliminate both contextual knowledge and intrinsic motivation. These valid criticisms of system evaluation solely against eye fixation data have helped guide the design of the experimental review of the thesis performance in Chapter 4.

1.2.3 Active Vision

Active vision refers to a visual system in which the image sensor moves or otherwise transforms as part of the perceptual process. Biological vision is highly active, executing an average of three saccades per second [58]. Although a large portion of computer vision research tends to focus on single, static images, active visual systems have nevertheless been explored in a number of computer vision applications. Several such studies which relate to the work of this thesis are discussed in this section.

1.2.3.1 Attentive Vision

The terminology with which to discuss and differentiate between active visual systems is not always apparent. Bajcsy was one of the first researchers to popularize the concept of active modulation of a visual system using intelligent control based in part on the visual information acquired by the system itself in what she termed *active perception* [5]. Aloimonos *et al.* showed that a number of common visual problems which are ill-posed for the passive observer become well-posed for an active observer, allowing them to be efficiently solved [4]. The work of this thesis, however, does not explicitly seek to provide additional constraints for solving a specific vision problem, but rather seeks to provide a framework by which a system can focus visual processing on the most important portion of the visual scene. Clark and Ferrier referred to this style of active vision as *attentive vision*, and provided one of the earliest formulations for such a system [12]. Their system encompasses one of the fundamental aspects of this thesis work, with its function being described as follows:

> The most salient feature is found and centered on the field of view. At this point a region of interest (ROI) processor may perform more complicated visual tasks.

Therefore, based on the formulation of Clark and Ferrier, this thesis work is fundamentally an attentive visual system. However, the focus of Clark and Ferrier was on the motoric control of a binocular visual system, whereas the focus of this work is at the determination and selection of a salient target.

1.2.3.2 Active Search Model

Active search is a subset of active vision in which an active visual system attempts to solve a visual search task. The introduction of motor capacity to the vision system enables the target to be placed outside the initial field of view, and opens up a number of new dimensions to the visual search problem beyond the traditional reaction time versus number of distractors. An implementation of an active search system which has heavily influenced this thesis work is Zaharescu's Active Visual Search Model (AVSM) [96]. The AVSM has many similarities to this thesis, including a separation of the peripheral and central visual fields as well a saccade history map which provides an IOR mechanism; the central field undergoes multi-level hierarchical processing which includes top-down task influences, while the periphery operates primarily in a bottom-up, low-level feature approach (for a brief discussion of the differences between the current work and the manner in which this thesis extends AVSM, see Section 1.4).

The scope of the AVSM was heavily concentrated on low-level visual search problems, particularly the visual search problem investigated by Motter and Belky [53]. In Motter and Belky's experiment, monkeys were tasked with identifying a red bar with a specified orientation amidst a distractor field of red bars at an alternative orientation, and green bars at an identical orientation to the target bar. It was revealed that the monkeys effectively ignored the green bars and focused their search on foveating patches of red bars to investigate further, a heuristic which was directly incorporated in the AVSM. Using a virtual implementation of the AVSM, Zaharescu was able to replicate similar search performance to that of Motter and Belky, providing a proof-of-principle justification for a biologically inspired active visual search system. This thesis aims to update, generalize, and expand upon the groundwork it laid.

1.2.3.3 Applications of Active Search

While the work in the previous section introduces motoric action to the visual search problem, it is still focused on 2D model environments in order to incorporate psychophysical findings. Vision applications to robotics typically seek to function in a 3D environment. Nevertheless, active search is a common problem posed to robot systems with applications in search and rescue, automated exploration and mapping, and object retrieval. Shubina and Tsotsos, [70], build on earlier work by Ye and Tsotsos, [95], to provide a system for object localization by a mobile robot in a 3D environment called the SYT algorithm.

A system utilizing SYT assumes knowledge of the global environment scope (for example, the size and shape of a room) but not the internal layout, and seeks to locate a target object within that region. The primary mechanism by which this is accomplished is to segment the 3D environment into an even distribution of volumetric cells, which are stored in an internal map and populated by a probability value of containing the target object (in the absence of prior knowledge, all cells are initialized to an identical probability value). The SYT system then seeks to execute a series of fixations which maximize the probability of finding the target with that fixation. The focus of this work is on navigation and camera movements within the environment, as well as recognition of internal environmental obstacles (such as furniture). However, it is pointed out that the cell probabilities could potentially be augmented with saliency information, and it is in this manner that the work of this thesis could be used to augment this type of active search task. A fish-eye lens, for example, could be used to provide a very wide angle of view, albeit with a large degree of peripheral distortion. This peripheral region could be analyzed with a rudimentary, saliency-driven peripheral mechanism which could be used to modulate the probabilities of the appropriate environmental cells, while the central region performs the actual recognition task. In this manner, a much larger set of environmental cells could be modified with each fixation, and the overall number of fixations required to complete a task should potentially be reduced.

1.3 Objective of Work

1.3.1 Problem Statement

The Boundary Problem, explored by Wal and Burt, [87], and discussed in detail by Tsotsos et al., [83], is an inherent aspect of processing with hierarchical pyramids that makes any results requiring multiple stages of processing valid only for central regions of the visual field (see Figure 1.2). Tsotsos has proposed a solution framework to combine top-down attentional direction based on higher order results in the central visual field with a bottom-up approach based on saliency in the periphery of the visual field, depicted graphically in Figure 1.8,

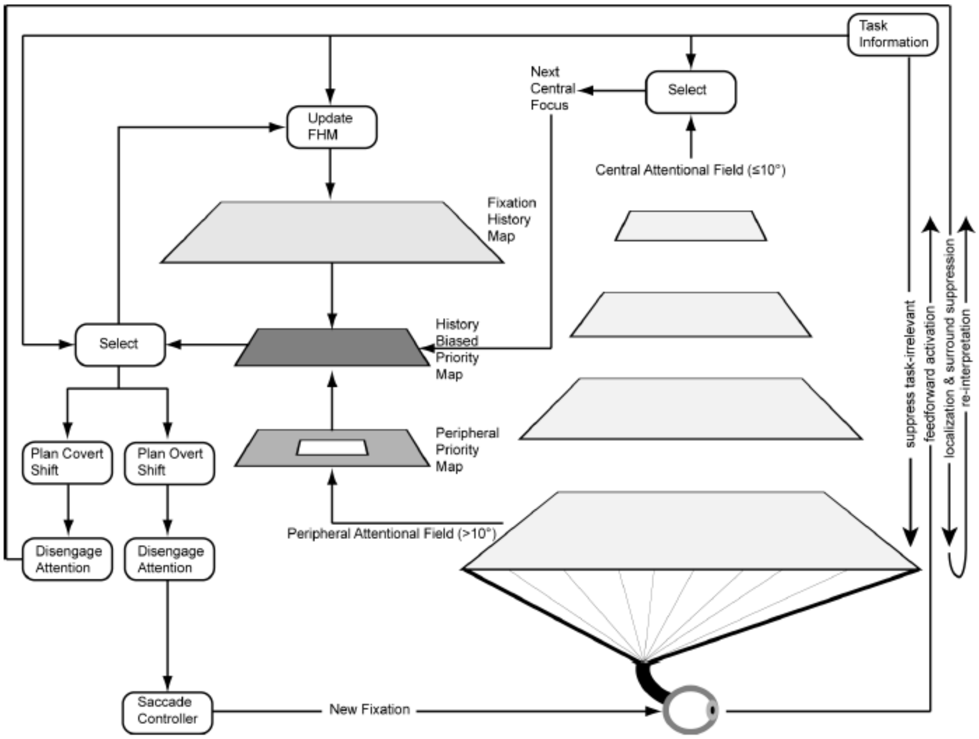


Figure 1.8 – Proposed solution for solving the Boundary Problem using a History Biased Priority Map (figure taken from [82])

[82]. Any peripherally salient items which the system identifies can be inspected in greater detail by moving the camera to bring them into the central region. The proposed framework leaves many of the specific model details unaddressed, however, and it is therefore the aim of this work to develop a functional implementation of this integration of central and peripheral processing.

1.3.2 Motivation

Psychophysical evidence suggests that attention likely operates as a collection of control mechanisms operating together rather than as a single overarching mechanism (see Section 1.2.1.3). Prior discussion of the Selective Tuning (ST) model of attention has concentrated on an attentive mechanism operating over a full visual hierarchy. This aspect of ST captures many psychophysical aspects of visual attention, and closely resembles the type of attention seen in response to endogenous cuing. This thesis seeks to implement a peripheral component of the ST algorithm to provide stimulus driven attentive direction from the visual periphery, as well as a motor control mechanism to reorient fixations according to this peripheral signal.

1.4 Significance and Contributions

This thesis describes an implementation of a solution to the Boundary Problem in a biologically plausible active vision model proposed in [82]. It extends previous work on a biologically plausible active visual model by Zaharescu, [96], by providing implementations of several algorithmic components which are either entirely novel or greatly extend similar structures in Zaharescu's work: the Peripheral Priority Map, the Fixation History Map, and the History Biased Priority Map. Additionally, a novel environment control component has been created to encapsulate the input and output signals (either physical or virtual) of the saccade controller, allowing the implementation of all components to be done in a modular and independent manner. This provides faster integration of new hardware components a far more general development platform for a wide variety of visual search research. Specific examples of research areas which may use this thesis as a computational substrate on which to experiment include the currently unresolved aspects of IOR discussed in Section 1.2.1.5 and improvements in active search, and are outlined in more detail in Section 5.2.

As well as providing a general active vision platform, this thesis makes several specific contributions. A conceptual framework for combining salience based on self-information from a heterogeneous distribution of filters is presented in Section 2.2.3. An implementation of the system on a physical pan-tilt camera unit provides a proof-of-principle demonstration of the system's capabilities to solve a visual search task. An additional experiment conducted in a virtual environment on a psychophysical data-set of natural images demonstrated fixation sequences more closely aligned with human fixation patterns than those produced by the AIM algorithm.

1.5 Thesis Outline

This thesis is organized into five chapters:

- Chapter 1 provided an overview of related research in areas of psychophysics and neurophysiology, computational models of visual attention, and active vision. It describes the primary motivation and the significance and contributions of the thesis.

- Chapter 2 describes the theoretical aspects of the novel thesis components.

- Chapter 3 gives specific implementation details for all of the thesis components.

- Chapter 4 describes the experiments conducted to test the system in both a physical demonstration as well as a virtual fixation experiment, and presents their results.

- Chapter 5 discusses of the experimental results in Chapter 4 and provides concluding remarks.

Chapter 2

Bottom-up Peripheral Saliency Model

2.1 Overview

As mentioned in Section 1.2.2.1, the Selective Tuning (ST) model of visual attention requires a peripheral component which undergoes only a few layers of processing in order to handle the Boundary Problem; this will necessarily be primarily data driven as a consequence of the low levels of abstraction which can be generated by only a few layers of processing. Such a mechanism is outlined in Chapter 4 of [82], and it is the implementation and validation of this mechanism which is the primary focus of this thesis work. A schematic representation of the entire ST model can be seen in Figure 1.8. A number of aspects of the peripheral component for ST required novel implementation. The three main components for handling peripheral attention are the Peripheral Priority Map (PPM), Fixation History Map (FHM), and History Biased Priority Map (HBPM), outlined in Sections 2.2-2.4, respectively. Although this Chapter discusses the theoretical basis and design for these novel components, specific implementation details can be found in the appropriate sections of Chapter 3. The HBPM additionally provides output to a saccade controller; since the design of this controller is relatively straightforward and not a theoretical focus of this thesis, it is discussed solely in terms of its implementation in Section 3.7.

2.2 Peripheral Priority Map

2.2.1 Functional Overview

The Peripheral Priority Map (PPM) utilizes the AIM algorithm described in Section 1.2.2.2 to provides a bottom-up measure of saliency for each point in its region of operation (see Section 2.2.2 for a discussion of the spatial extent of the PPM). This bottom-up calculation stream, displayed in Figure 2.1, can be modulated by top-down task biasing in order to produce a priority signal similar in nature to that described by Fecteau and Munoz, [24]. The priority map is then sent to the History Biased Priority Map (HBPM) where it can be incorporated into attentional selection and possibly trigger an overt attentional allocation with a saccade. Although AIM is a well-developed and studied saliency map model, it has not previously been implemented in the TarzaNN architecture (see Section 3.3 for an overview), and thus had to be programmed largely from scratch and adapted to its role in the system as a whole (see Section 3.4 for a discussion of the details of this implementation).

Figure 2.1 provides an abstracted schematic of the network which generates the bottom-up portion of the PPM. Data flows through the network from the bottom of the figure to the top. The input plane represents data from the visual field over which the PPM will operate; this may either be the entire visual field or some subset (see Section 2.2.2). This visual input is passed to

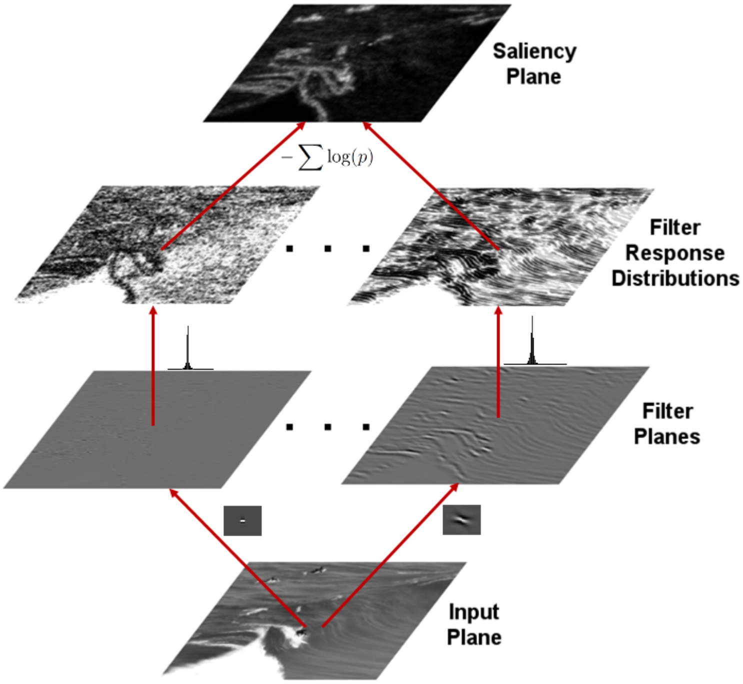


Figure 2.1 – Abstract representation of the bottom-up portion of the peripheral priority map calculation stream. An input image (bottom) passes through a bank of filters to produce a set of filtered image planes (second row). In the example shown 24 log-Gabor filters were used, but the number and type of filters can vary with task and implementation. The pixel response values of each filter plane are counted to produce a filter response density estimation, which is subsequently used to estimate the probability of the filter response value for each pixel (third row). The self-information of each pixel is summed across all the filter channels to produce a saliency map of the image (top image).

a series of filter planes; each filter plane receives an identical copy of the input data and operates independently and in parallel with the other filters. The actual filter set used can be modified to suit any particular task, environment, or implementation needs. A range of possible filter bases over which AIM could operate, including ICA patches, Gabor filters, log-Gabor filters, and Difference of Gaussians, was explored by Bruce et al. [8]. When comparing the ability of AIM to highlight coherent objects in overhead street scenes, it was found that log-Gabor filters provided the best correspondence of salient regions with those labeled by human observers. In addition to superior performance, log-Gabor filters are well-studied in computer vision with a relatively simple parametric structure, [44], and have response properties very similar to those of the early primate visual cortex, [27, 59]. Together these properties make log-Gabor filters the most attractive option for this thesis work.

Additionally, having a parametric set of basis functions should allow more transparent control over the response behaviour of the system and thereby allow extensions for short-term temporal priming of particular features, [49], or even goal-oriented feature tuning, [28, 55]. Although top-down modulation of the saliency signal is not yet implemented for this thesis, it is a clear next step in producing a biologically plausible guide for visual attention. To that end, the representation produced by the PPM is referred to as a priority map rather than simply a saliency map in order to remain consistent with the terminology used by Fecteau and Munoz, [24]. One of the aims of this thesis is to provide a computational platform on which various methods of top-down control may be tested, as the ways in which humans exert top-down control of visual search salience remains an area of ongoing research, [49, 50, 51, 23].

Once a filter response image has been calculated as the output of each filter plane, an estimate is formed for the distribution of that filter's response over a

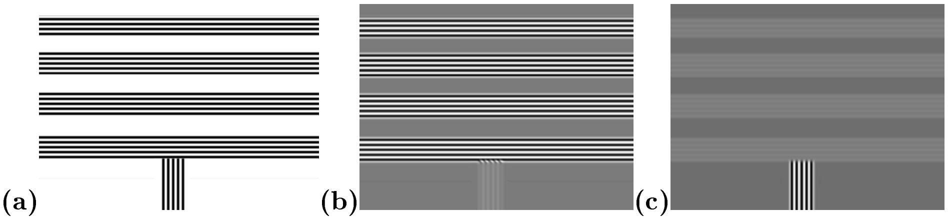


Figure 2.2 – Demonstration of oriented log-Gabor filter responses to an example image with many horizontal edges and few vertical edges. **(a)** The original image being run through the filters. **(b)** The result from a horizontally aligned log-Gabor filter. **(c)** The result from a vertically aligned log-Gabor filter. In both filter response images darker colours correspond to a strong negative response of the filter, lighter colours to a strong positive response, and grey to a zero response.

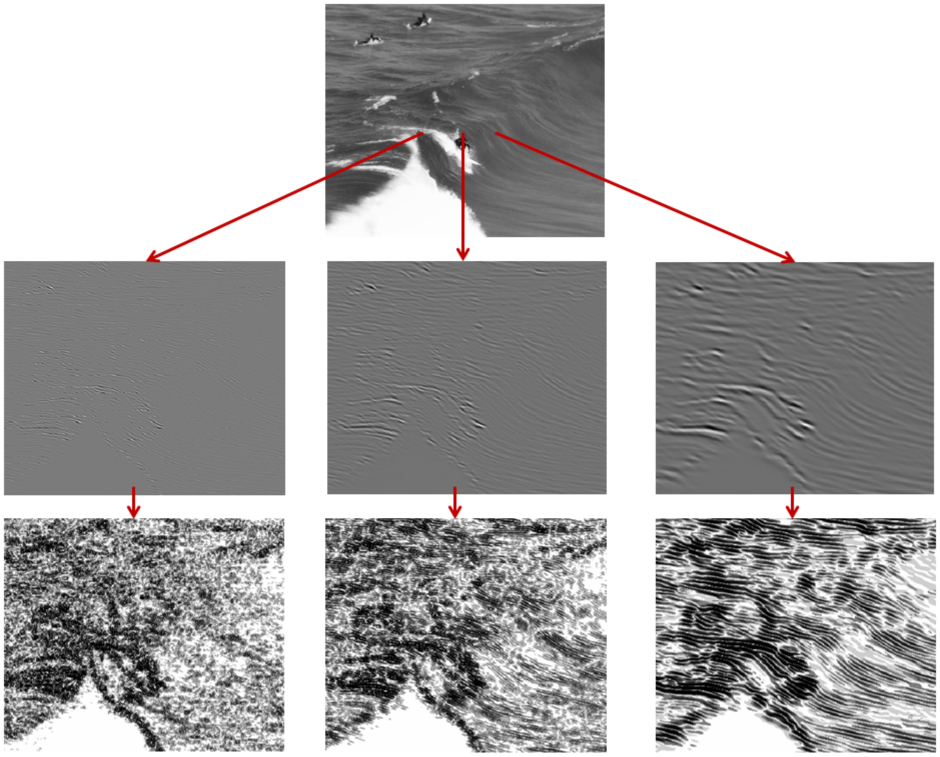


Figure 2.3 – Filter output (middle row) and the corresponding probability estimates (bottom row) for log-Gabor filters with the same orientation tuned to three different spatial scales. The filter responding to the highest spatial frequencies is on the left, and the filter responding to the lowest is on the right.

given image. For example, an image with many horizontal edges but few vertical edges when passed through a horizontally oriented log-Gabor filter would have a number of regions with a large filter response. A good estimate for the probable response of a pixel randomly selected from that image would therefore reflect this, and give a high probability score for a large response. Likewise, that same image passed through a vertically oriented log-Gabor filter would produce few regions with a large response, and therefore the probability density estimate for that filter response would be high for a low response (see Figure 2.2). Estimating the probable response of each pixel allows the self-information of that pixel to be calculated, which can in turn be used to calculate the saliency value according to the AIM algorithm (see Figure 2.3).

Using the probability density estimate for each filter, we evaluate the probability of obtaining the value for each pixel in the filter response image. These probability estimates are then combined across all the filters to produce the salience of each pixel in the saliency output plane according to the AIM saliency equation:

$$\mathbf{SALIENCY}_x = -\log(\prod_{i \in \mathbf{F}} p_i) = -\sum_{i \in \mathbf{F}} \log(p_i) \tag{2.1}$$

where $\mathbf{SALIENCY}_x$ represents the saliency of an arbitrary pixel x, $\mathbf{F}$ is the set of all filters, and p_i is the probability estimate for the response of pixel x to the ith filter.

2.2.2 Spatial Extent

The PPM produces a priority score for a defined proportion of the visual field which defaults to eccentricity greater than ten degrees, but which can be customized to any appropriate value (including the entire image, in which case the PPM operates in much the same way as any bottom-up saliency map algorithm discussed in Section 1.2.2.2). Even though the central visual field is not given a bottom-up salience value by the PPM, it could nevertheless provide important contextual image statistics in the generation of the probability density estimates for the peripheral region in which the PPM does generate a salience value. Thus, it is important that the PPM analyzes the entire visual field with virtually the same concentration of filters (see Section 2.2.3 for a possible exception to this). This necessary representation of the entire scene can be seen physiologically in area V6 in the human visual system, which represents the whole scene in a roughly equivalent manner rather than incorporate the rapid drop-off in cortical magnification with increasing eccentricity of most other visual areas (see Section 1.2.1.2) [22]. It should be noted that the computation shown in Figure 2.1 completes the saliency calculation over the entire image, rather than just over the periphery for which the priority signal will ultimately be utilized. This was done both for display purposes and because it was convenient to mask out the central saliency values at the level of the History Biased Priority Map (see Section 2.4) rather than at the level of the PPM. It would be unlikely that a biological system would perform excess calculations which are never passed on, however, so a future system seeking better biological fidelity would most likely shift the central masking to the level of the PPM.

In addition to addressing where the PPM applies its filters, however, the visual neighbourhood from which the density estimates are created must also be examined. As mentioned in Section 1.2.2.2, the original formulation of AIM applied a Gaussian window from which it drew the filter responses to create the density estimate, effectively resulting in an individual regional density estimate centered at each individual pixel. However, published implementations up to now have, for computational simplicity, all relied on a definition for the surrounding neighbourhood in which each pixel in the entire image contributes equally to the density estimate, creating a single histogram density estimate equivalent for all pixels. Although the psychophysical evidence for spatial distribution of context in human visual search is not well explored, it is clear that humans are not constrained to just a single level of spatial context (see Figure 2.4). The TarzaNN framework is designed to handle parallel computation over a feature plane, and thus more limited local neighbourhoods should in principle become computationally feasible, thereby allowing the exploration of alternative definitions of local neighbourhood in the future. To facilitate this, the PPM was designed to be extensible in the manner with which it calculates density estimates.

2.2.3 Entropy-based Heterogeneous Filter Relation

Under the standard formulation of AIM, each pixel has its saliency determined according to equation 2.1. However, it is possible that some subset β of the filter set cannot be executed over a given image region. An example of when this might occur is for a filter set which varies in size (higher spatial frequencies can be measured with smaller filter sizes), which therefore introduces a strip of pixels around the image periphery over which only the smaller filters are defined. In the case of a pixel which only has some filters defined, a measure of salience based on self-information can still be derived from the observed filters, but it is not readily apparent how this measure can be compared to the saliency of elements over which all filters have been executed.

An obvious mechanism, and the simplest option to implement, would be to simply replace the sum of the self-information over all filters with the sum of self-information over the executed filters. Thus, equation 2.1 becomes:

$$\mathbf{SALIENCY}_x = -\sum_{i \in \alpha} \log(p_i) \tag{2.2}$$

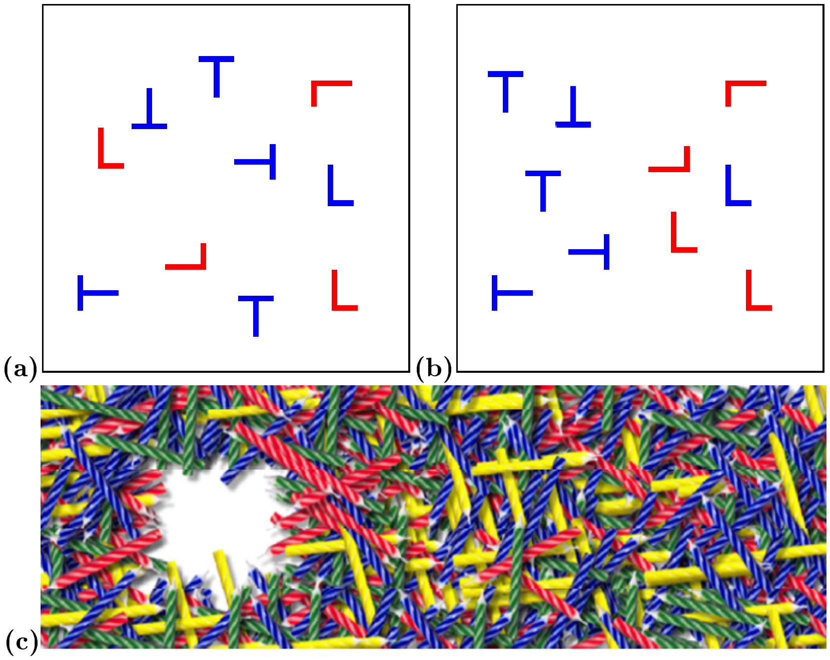


Figure 2.4 – Several demonstrations of the effect that neighbourhood size has on the saliency calculation. **(a)** Hard conjunction search reproduced from Figure 1.3. **(b)** A conjunction search with identical elements to (a), but which is made easier by the spatial grouping of objects. If only the global neighbourhood is considered in a saliency calculation, however, the target blue L will have identical salience in both (a) and (b). **(c)** An image in which the global spatial context is necessary for locating the distinctly salient hole in the candle coverage (image reproduced from [9]).

where $\alpha \subseteq \mathbf{F}$, α is the set of filters which are defined over pixel x, and all other elements of the equation are as defined in equation 2.1. This effectively penalizes the salience measure of a pixel for any missing filters by setting the self-information measure from that filter to zero. Although this is easy to implement, for any region which has a large number of undefined filters it will become quite difficult to generate the maximum salience value in the salience map even with a very high self-information score from the defined filter set.

Another relatively simple option would be to normalize the salience measure based on the number of components the measure is based upon. Thus, equation equation 2.1 becomes:

$$\mathbf{SALIENCY}_x = -\frac{1}{N}\sum_{i\in\beta}\log(p_i) \tag{2.3}$$

where N is the size of set α, $\alpha \subseteq \mathbf{F}$, and all other elements of the equation are as defined in equation 2.1. Although this ensures that regions with fewer defined filters still have a decent chance of capturing attention, it leaves those regions more vulnerable to noise and reduces the overall effect of the filters which are not defined over the whole image.

A third, novel, option is proposed here as *entropy-based heterogeneous filter relation* (EBHF). In this formulation, entropy is used as a measure of the expected self-information for the undefined filters:

$$-\sum_{k\in Y} p_{j_k}\log(p_{j_k}) \tag{2.4}$$

where Y is the full set of elements in each probability distribution $P_j(Y)$ and p_{j_k} denotes the probability $P_j(Y = k)$. Therefore, inserting equation 2.4 into equation 2.1 in place of the self-information component for all filters β which are undefined over pixel x, we get:

$$\mathbf{SALIENCY}_x = -\sum_{i \in \alpha} \log(p_i) - \sum_{j \in \beta} \sum_{k \in Y} p_{j_k} \log(p_{j_k}) \tag{2.5}$$

where $\alpha \subseteq \mathbf{F}$, $\beta \subseteq \mathbf{F}$, $\alpha \cap \beta = \emptyset$ and $\alpha \cup \beta = \mathbf{F}$.

Although the entropy calculation adds some increased computational complexity to the saliency calculation, it is a constant value over each probability distribution and therefore needs to simply be calculated once at the time of the density generation. The addition of an entropy component will allow one to essentially guess at the self-information contribution of an unresolved filter, and thereby directly compare saliency measures across an image even when the image has been run through a heterogeneous filter set. It may be useful to modify equation 2.5 with a weight term applied to the entropy elements, in order to adjust the impact of the entropy term:

$$\mathbf{SALIENCY}_x = -\sum_{i \in \alpha} \log(p_i) - c \sum_{j \in \beta} \sum_{k \in Y} p_{j_k} \log(p_{j_k}) \tag{2.6}$$

where c is a constant weight.

EBHF was not implemented as part of this thesis work, as the filters were all made equal in size. However, in the future it may be a worthwhile mechanism to explore, particularly if a larger and more diverse filter set is required for a PPM implementation.

2.3 Fixation History Map

The Fixation History Map (FHM) provides a record of prior fixations over a region greater than the current visual field. This allows IOR to extend beyond the visual field of a current fixation, and prevent an oscillatory cycle of fixations between a set of three or four widely spaced points of interest. It provides a first approximation for the psychophysical evidence reviewed in Section 1.2.1.5;

as with the PPM described in Section 2.2, the FHM implemented in this thesis currently functions in a bottom-up manner. It is meant to serve as a computational tool which can be used to test top-down control strategies in future implementations.

2.3.1 Functional Overview

The FHM provides a short-term inhibitory mechanism based on prior fixation locations. Previous fixations are stored in a retinotopic map, with spatial locations relative to the current fixation. The FHM is two dimensional; if future work extends the FHM to provide object-based IOR or other more complex behaviour seen in human psychophysics, it is unclear if this will require an extension of the FHM into three dimensions or if it will still be best served by recording two dimensional projections of inhibitory locations in retinotopic coordinates.

Locations stored in the FHM begin to decay with each subsequent fixation, whether that fixation is a covertly allocated central fixation or an overt movement of the camera. The duration of this memory can be customized, but it defaults to a linear decay which completely ceases after five subsequent fixations (see Section 1.2.1.5 for a discussion of the temporal persistence of IOR). Since the memory of prior fixation locations are stored relative to the current fixation, they must be updated with each camera movement. This is performed for computational simplicity by directly translating the stored values in the FHM based on the magnitude of motor commands, but Zaharescu suggests a method by which this translation could be accomplished in a more biologically plausible manner through neuronal interactions, [96].

Neurophysiological evidence has been found for a similar update mechanism in the posterior parietal cortex of both humans and monkeys, [52, 20]. Interestingly, Duhamel *et al.* additionally found evidence in the parietal cortex for neurons to fire in response to a saccade bringing the location of a previously flashed stimulus into their receptive fields, even though that stimulus was no longer active [20]. This was interpreted as parietal cortex cells anticipating the appearance of the visual field in response to an eye movement. It could alternatively be evidence that the flashed stimulus elicited an attentional allocation (as sudden onset stimuli often will, see the end of Section 1.2.1.2 for a discussion of attentional capture) which is now being remembered as part of an IOR mechanism. It is worthwhile noting that the neurophysiological evidence presented here implicates regions of the parietal cortex as part of an IOR effect, whereas evidence reviewed in Section 1.2.1.5 suggested instead that IOR is controlled by the frontal eye fields (FEF). It is possible that both areas have a role in the implementation of IOR, as some research has posited that the regions are functionally linked as part of an interconnected oculomotor network, [24].

The original interpretation of Duhamel *et al.*'s findings raises an interesting question about the nature of short term visual memory. In this thesis, the only form of visual persistence between fixations is an IOR mechanism; however, it is possible that other forms of low-level information might be beneficial to hang onto as well. For example, if two peripheral elements on opposite sides of the visual field have high saliency values which are nearly identical, it could be worthwhile for the system to remember the location of the slightly lower one as it fixates on the winner. This value could then be compared against the saliency values of the now current peripheral visual field, allowing the system the option of saccading all the way back to this other highly salient element. A saliency modulated set of probability cells described in Section 1.2.3.3 would effectively operate in this manner, as information from a current fixation would have the ability to both inhibit or enhance the probability of a future fixation to a region.

2.3.2 Spatial Extent

Evidence for short-term memory regarding an attended location outside of the visual field was demonstrated by Tark and Curtis using an auditory-visual task which presented auditorally cued points of interest both behind and in front of subjects' heads [73]. While human subjects were able to keep track of regions of interest located fully behind their heads, for the purposes of this thesis it was deemed unlikely that a solely visual system would need to support an FHM extending beyond twice the size of the visual field. This is due to the fact that an auditory cue might cause a person to jump from one fixation to another directly behind, whereas a solely visual system is only able to turn itself around by following a chain of fixations which can be no farther apart than half the width or height of a single visual field. Thus the inhibition of a point which is farther than two visual fields away from a current target is likely to have worn off by the time fixation once again returns to its general vicinity unless a specific task implementation requires an unusually long inhibitory persistence.

In addition to deciding the extent of the FHM, the resolution is also important. The lower the resolution of the map, the fewer computational resources it will require. Nevertheless, reducing the resolution too greatly will adversely affect performance, as the inhibition of one target could potentially inhibit the ability of the system to attend to another nearby yet distinct target. Since memory optimization was not a primary concern in this thesis, the FHM was left at the same resolution as the visual field itself. Future implementations which are more resource limited, however, may need to explore more space efficient sizes of FHM resolution.

2.4 History Biased Priority Map

The History Biased Priority Map (HPBM) integrates the information coming from the central attentional field with that of the FHM and PPM in order to determine an overall attentional decision. Although the aim of this thesis is to utilize the Selective Tuning (ST) model to guide attention in the central attentional field, the full implementation of ST is beyond the scope of this work, and therefore an implementation of ST from previous work will be used [65]. There is nothing in the nature of the HBPM which requires attentional direction from the central field to be provided by the central ST mechanism (and one should in principle be able to use the fixation controller developed here with another attentional model). However, since the development of this model was driven by the view of attention as a set of interacting mechanisms rather than a single overarching algorithm, a view which is integral to the nature of ST, it is natural to concentrate on integrating the system within the overall ST architecture.

2.4.1 Functional Overview

The primary challenge of the HBPM design is in balancing the priority of the next central fixation (NCF) signal from the computations of the central attentional field with input from the PPM. The NCF and the saliency-based signals found within the PPM are derived from quite different calculations, and it is thus difficult to directly compare the two. In previous work which has looked at a separate overt and covert attentional system, the central and peripheral attentional signals were still driven by the same algorithm, and could thus be compared directly in an “Eye Movement Bias” term as the difference between the average neuronal activation in the centre field minus the average neuronal activation in the periphery [96]. In this thesis, selection between the NCF and

PPM signal is instead a winner-take-all comparison of the two signals. The HBPM is intentionally structured to be flexible in the manner with which it integrates the two elements depending on how the NCF signal is generated.

2.4.2 Spatial Extent

The HBPM covers the entire visual field, although it will only activate the saccade controller if a peripheral signal is selected to have the highest priority. Future implementations, however, may benefit by extending the HBPM beyond the size of the visual field if the system seeks to integrate fixational cues from memory (see Section 2.3.1 for a brief discussion), world knowledge, or other sensory modalities.

2.5 Summary

This chapter introduced an overview of the conceptual aspects for the main algorithmic components of this thesis: the Peripheral Priority Map (PPM), the Fixational History Map (FHM), and the History-Biased Priority Map (HBPM). Additionally, an implementation of the AIM algorithm in TarzaNN is discussed in the context of serving as the bottom-up processing stream for the PPM, and a novel method for calculating saliency based on self-information for a heterogeneous filter distribution is proposed. The three algorithmic components described provide a useful computational framework which can be used to test top-down control strategies which have previously been largely ignored in saliency-based approaches to attention. Specific implementation details for these components are discussed in Chapter 3.

Chapter 3

Implementation

3.1 Overview

The majority of work for this thesis was implemented as an extension to the TarzaNN neural network simulator framework, which will be described in Section 3.3. A few minor components were performed in MATLAB or utilized proprietary software libraries; these will be noted when appropriate. The physical hardware used will be reviewed in Section 3.2. Sections 3.4-3.6 describe the implementation details of the corresponding sections from Chapter 2, while Section 3.7 overviews the saccade controller. Section 3.8 overviews an additional structure introduced to the TarzaNN framework in the course of this work which provides a unified environment controller designed to provide a virtual testing environment for the system as well as allow easier and more reliable extension of the elements of this thesis to future physical hardware setups.

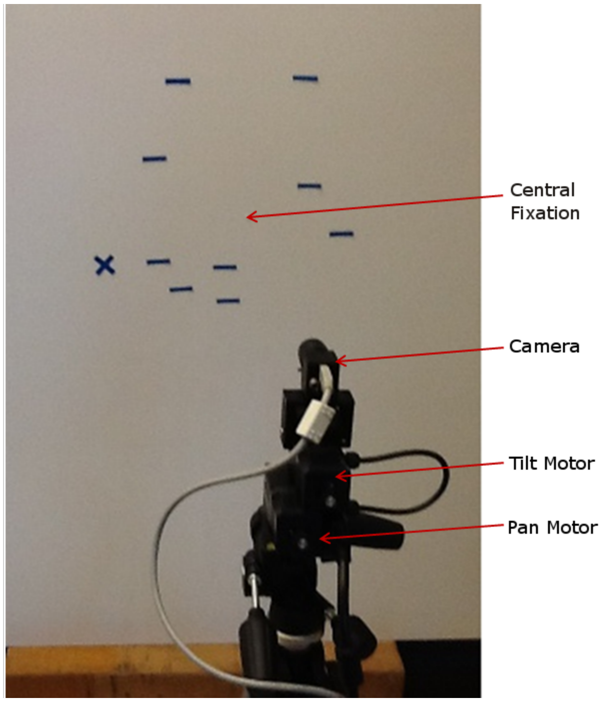


Figure 3.1 – Physical setup of camera and PTU.

3.2 Physical Hardware

3.2.1 Pan-Tilt Unit

The pan-tilt unit used is a Model PTU-D46 from DirectedPerception, [2, 19]. It provides reliable and accurate positioning controllable through ASCII or binary formats. The PTU is connected via a serial port; the serial port address and BAUD rate must be provided to the network through the environment control structure (see Section 3.8).

3.2.2 Camera

The camera used is a Point Grey Flea, a miniature IEEE-1394 (FireWire) camera produced by Point Grey up to January 2010, at which point it was no longer marketed to new customers due to the availability of Flea2 and Flea3 cameras [1]. The camera has numerous customizable operating parameters, and was operated in Mode 1 (downsampling acquired images at a 2:1 ratio) Mono16 for the purposes of this thesis. The lens is a 6mm CCTV lens f1.2.

3.3 TarzaNN Neural Network Simulator

TarzaNN is a general purpose neural network simulator implemented in the C++ programming language. Its primary design was completed by Albert Rothenstein and Andrei Zaharescu; descriptions and details of the design can be found in [66, 65, 96], and an online resource is currently maintained for the project, [67]. The discussion here will therefore simply attempt to provide a descriptive overview of the TarzaNN framework, as further details may be found in the original reference materials. Extensions to TarzaNN implemented as part of this work are described in the appropriate sections following this section's overview of the overall framework.

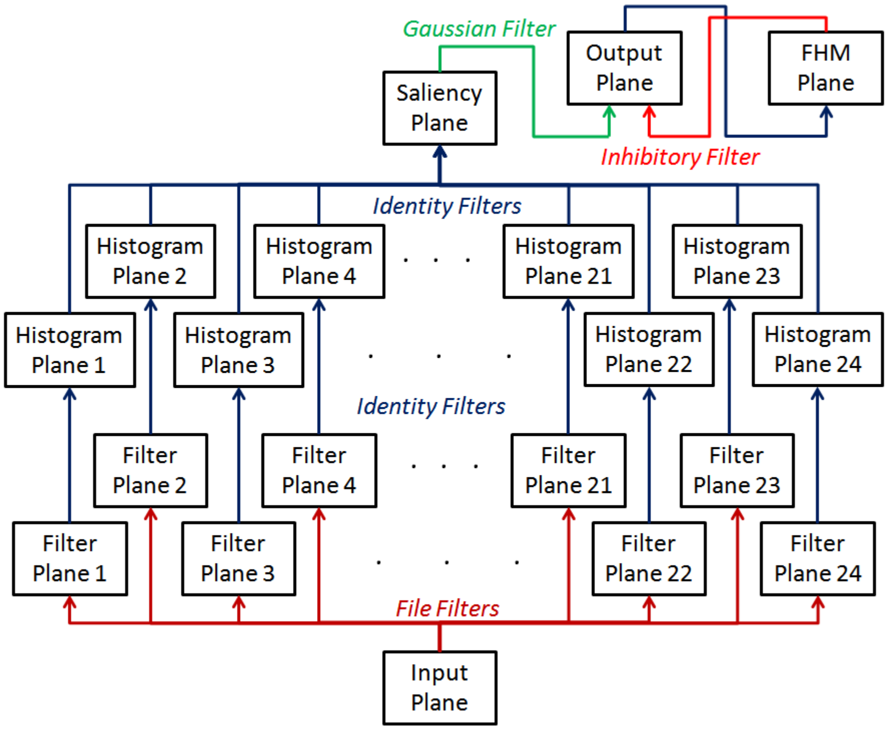


Figure 3.2 – Diagram representation of the network used in this thesis representing the TarzaNN objects. Different filter types are colour coded. Not shown is the environment control object, as this is a global object which is not directly linked to any one feature plane.

TarzaNN is designed to provide a wide set of customizable tools which are combined to create a specific processing network. The building blocks of these networks come in two main forms: Feature Planes and Filters. A diagram of the network developed for this thesis representing the TarzaNN objects is shown in Figure 3.2.

3.3.1 Feature Planes

Feature planes are two-dimensional clusters of neurons, displayed in Figure 3.2 as the labeled black boxes. A feature plane is defined by the receptive field and response properties of its constituent neurons. The receptive field properties of each neuron are determined by the filters linking a given feature plane to the rest of the network, while the response properties of the neurons are set as an intrinsic property of the feature plane. A number of commonly desired response properties are available, including a linear unconstrained neuron which can take on any numerical value determined by the resulting convolutions of its receptive field and a sigmoidal neuron which maps incoming responses onto a sigmoid curve between a minimal and maximal response value. A special case of the feature plane is the input feature plane, which has no receptive fields to modify its neuronal responses but instead provides input to the network through a defined data source such as a camera or saved image.

Feature planes can themselves be organized into layers, allowing an additional degree of control over the network structure. A network can be set to execute layer by layer, pausing before moving on to the next; this is useful for debugging and explanatory purposes.

3.3.2 Filters

Filters are the links between feature planes and define the receptive fields of the constituent neuron. Filters are displayed in Figure 3.2 as the coloured arrows linking feature planes. Each filter consists of a convolution kernel applied to the output values of one feature plane which are then passed as input to a target plane. TarzaNN itself has a number of predefined methods for padding an image edge or interpolating around boundary effects introduced by a kernel convolution, but these were turned off for the purposes of this thesis. A feature plane can receive input from multiple planes via a set of identical (such as the Saliency Plane in Figure 3.2) or heterogeneous filters (such as the Output Plane in Figure 3.2); the resulting input from each individual plane is then additively combined according to the response properties of the feature plane neurons.

A number of common filter types are defined, such as the simple identity filter which simply passes input unchanged between feature planes and a Gaussian kernel. Additionally, a file filter is provided which was originally developed to read in a customized kernel in the form of an image. This was extended for this thesis to allow input either from image formats or text files, thereby allowing completely customizable filter kernels (including negative values which were precluded from image-based file formats) to be developed outside of TarzaNN and easily passed in as part of a network definition.

3.4 Peripheral Priority Map

The peripheral priority map, as mentioned in Section 2.2, provides an implementation of the AIM saliency algorithm. This required several specialized extensions of the existing TarzaNN object classes. Although these extensions were performed specifically with AIM in mind, they may nevertheless be potentially used in any future arbitrary network.

3.4.1 Filters

As described in Section 2.2 and displayed in Figure 2.1, the peripheral priority map applies a bank of filters to the input image from which the image statistics necessary for the AIM algorithm may be derived. The PPM calculation corresponds to the feature plane set from the Input Plane to the Saliency Plane in Figure 3.2. The filters chosen for this implementation and used in both experimental setups described in Chapter 4 were log-Gabor filters generated in MATLAB, [44]. As mentioned in Section 2.2, log-Gabor filters were chosen for a number of positive performance reasons as well as for their similarity to early cortical processing. They additionally have the advantage of not being tuned to any particular problem domain (unlike ICA filters which are derived by sampling image patches from within a given data-set), and thus could be effectively used in both experimental setups explored in this thesis despite the dissimilarity of the two experimental domains. Likewise, the parametric nature of log-Gabor filters should allow for a more straightforward approach in future work into top-down tuning of the priority signal.

The filters were specifically generated using Kovesi's `gaborconvolve` MATLAB function, [44], on a sample image taken from the physical experiment described in Section 4.1.1. A set of 24 filters was generated with eight orientations at three spatial scales with a minimum wavelength of 3 and a scaling factor between spatial scales of 2. One challenge to utilizing log-Gabor filters is that most applications (including Kovesi's code) perform the convolution calculation in the frequency domain; implementing a frequency domain calculation in TarzaNN would require a great deal of task-specific programming which would have minimal biological plausibility. Therefore, the log-Gabor filters were necessarily subsampled kernels which provide a spatial approximation to the frequency domain calculation (see Figure 3.3 for an example). A numerical check was performed on each kernel to ensure that the spatial truncation of the filter kernel introduced a minimal DC-component to the calculation and maintained the majority of the filter information.

3.4.2 Histogram Feature Plane

The histogram feature plane is a child class of the feature plane, specifically modified to produce a probability estimate of the neuronal value calculated rather output that value itself. This probability estimate is currently produced by creating a single histogram of neuron values over the entire feature plane, but in the future will be extended to allow localized histogram estimates as well.

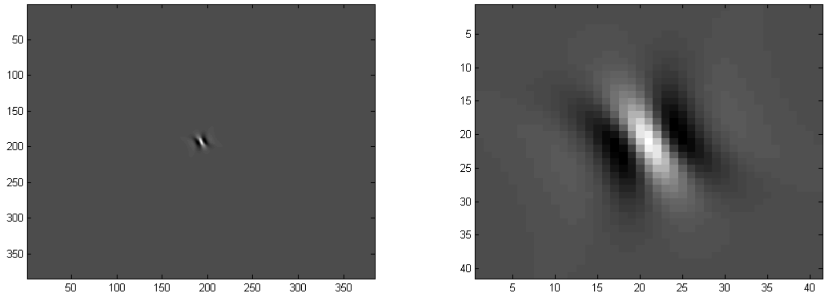


Figure 3.3 – The spatial truncation of a log-Gabor filter. The image on the left shows the full spatial representation of the filter produced by the `gaborconvolve` MATLAB function, while the image on the right shows the truncated kernel used to approximate the filter in TarzaNN. The example shown here is for the largest spatial resolution used in this thesis.

Additionally, the implementation of a heterogeneous filter relation as described in Section 2.2.3 will require additional modifications to define the region in which an entropy-based estimation was required. It is likewise possible that future work may want to estimate probabilities with more complex estimates than a simple counted histogram; this should be readily possible through the addition of tunable parameters in the feature plane constructor.

It should be noted that the network shown in Figure 3.2 includes a layer of Filter Planes for the log-Gabor filter responses followed by a subsequent layer of Histogram Planes which pass probability estimates for each pixel on to the Saliency Plane. Implementing this as two separate layers is functionally unnecessary as each feature plane representing a log-Gabor filter response is linked via an identity filter to its corresponding histogram feature plane, and a computationally faster and more space-efficient network could be implemented which combines the two layers into a single layer of Histogram Planes. Since this thesis was not focused on computational efficiency, however, it was decided to keep the two layers separate in order to allow visual inspection of the filter responses as well as the probability estimates for those responses (see Figure 3.4

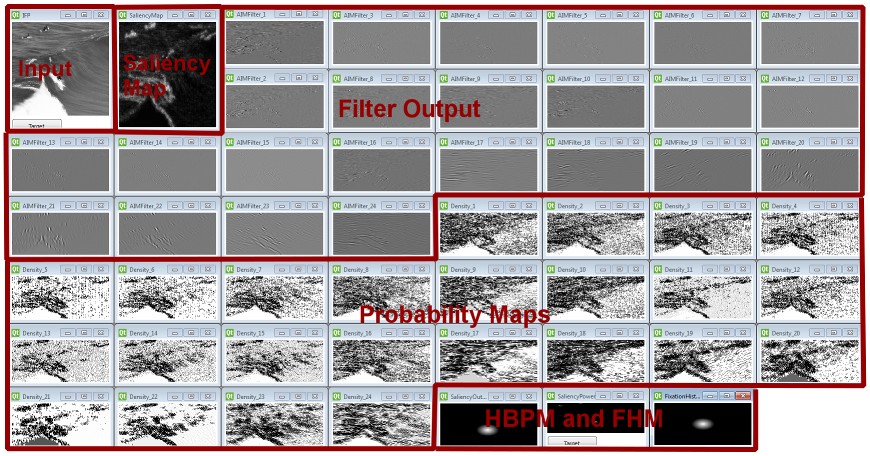


Figure 3.4 – Example of TarzaNN display after a completed pass through the network. The input image and the saliency map are displayed in the top left. The raw output of the log-Gabor filter bank is shown in the upper middle, and the resultant probability maps based on that filter output is displayed below. In the bottom right the target selection and translated fixational history map are shown, along with the resultant output of the HBPM which has been scaled to accentuate the winning saliency region.

for an example of the active output produced during network execution).

3.4.3 Saliency Feature Plane

Like the histogram feature plane, the saliency feature plane is another special-purpose child class of the generic TarzaNN feature plane. Whereas the histogram feature plane introduced a hidden histogram structure in order to replace the standard neuronal output values with an estimate of their probability, the saliency feature plane functions as a normal feature plane which overloads the standard summation of incoming activations with the sum of the negative log of those values. It is assumed that input values should be on the interval $(0, 1]$, and a check is included to ensure that no values explode to infinity. After calculating the value of the sum over the negative log of all input values for all constituent neurons, they are linearly scaled by the maximum response value,

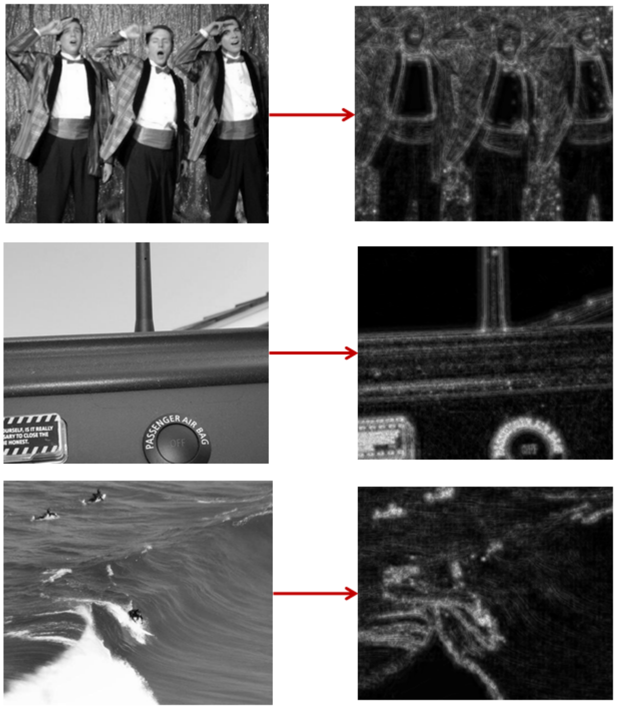


Figure 3.5 – Several examples of saliency maps generated for a variety of image types. The top image is a challenging image for bottom-up saliency due to the many bright and varied textures, as well as the lack of contextual knowledge about human structure and faces. The centre image displays how text tends to produce strong salience signals due to its distinct and strong edges. The bottom image displays a natural scene to which AIM is well-suited; the surfers stand out from the background without requiring any contextual knowledge.

thereby casting all output to the range $[0, 1]$. Several example images and their corresponding saliency maps are shown in Figure 3.5.

3.5 Fixation History Map

The fixation history map (FHM), as described in Section 2.3, provides a record of previously fixated locations in order to provide an IOR element to the network which prevents continuous oscillation between a small number of highly salient regions. This is implemented via the introduction of an inhibitory filter and an

FHM feature plane, and is shown as the FHM Plane in Figure 3.2.

3.5.1 Inhibitory Filter

The inhibitory filter provides a number of inhibitory kernels based on the mode and size specified. For all experiments in this thesis inhibition was provided in the form of a negative identity filter (a 1×1 kernel with a value of -1), but the option for a negative Gaussian and a negative filter either uniformly averaged or uniformly applied over a kernel of a specified size were also implemented. All outgoing filter connections of the FHM feature plane are inhibitory filters.

3.5.2 FHM Feature Plane

The FHM feature plane is a child class of the generic TarzaNN feature plane. The FHM feature plane includes an additional memory array as a private data structure hidden from the rest of the network. This memory array must be at least as large as the feature plane itself, but may be arbitrarily larger. The memory array is cocentric with the feature plane, and stores a history of all remembered fixations. Each time the feature plane steps, elements of the memory array decay according to the equation:

$$\mathbf{M}_{i,j}^{t+1} = \max(\mathbf{M}_{i,j}^{t} - \frac{1}{\alpha}, 0)$$

where $\mathbf{M}_{i,j}^{t}$ represents the memory value at time step t and location (i, j), and α is a defined decay constant. Linear decay appears to give acceptable results for now, but future work will investigate other methods of inhibitory decay (such as exponential decay). Once all memory traces have been decayed, they are translated in order to centre the new locus of attention provided by the HBPM, at which point the new locus of attention is added to memory. The output of the FHM feature plane is then taken from the central the portion of the memory matrix which is of the same size as the feature plane.

It is important to note that the FHM feature plane is only defined to receive input from a single feature plane. This contrasts with the functionality of other TarzaNN feature planes which are capable of supporting an unspecified number of input or output connections. This limitation was imposed on the FHM feature plane since it receives as input a representation of the locus of attention, something for which multiple inputs does not make sense. In the future when top-down task influences are incorporated, this will either have to be done through intermediate feature planes or a slight modification to the manner in which the FHM feature plane is programmed to function.

3.6 History Biased Priority Map

The history biased priority map (HBPM), as described in Section 2.4, integrates the saliency map with inhibition from the FHM to determine the peripheral target with the highest priority. It is represented by the Output Plane in Figure 3.2. It performs its function with a special-purpose child class of the TarzaNN feature plane called the saliency output feature plane. The input from the saliency map is generally smoothed via a Gaussian filter and combined with the output of the FHM feature plane through an inhibitory filter. The smoothing of the saliency map is performed to increase the signal-to-noise ratio, since the winning target is based on the largest single pixel response (see Section 4.2.1.2 for further discussion of salience smoothing).

Currently the HBPM will output its peripheral priority target directly to the PTU controller (see Section 3.7.2) when operating with a physical environment. When operating in a virtual environment, the peripheral priority target is instead sent to the environment controller (see Section 3.8). In the future this will be modified such that the HBPM will always send its peripheral target information to the environment controller, which will then determine the appropriate response based on the physical or virtual properties in which the network is operating.

3.7 Saccade Controller

The saccade controller consists of two major components: image acquisition and PTU control. Currently these two components are controlled separately and directly by the specialized feature planes developed for this thesis; in the future it is planned to move these components to be a component environment in the environment control class (see Section 3.8).

3.7.1 Image Acquisition

Image acquisition is controlled through a specialized input feature plane called an input camera feature plane. This feature plane receives input from the camera described in Section 3.2.2, utilizing Point Grey FlyCapture development tools. Because of this reliance on proprietary software, the network only works on the Windows operating system, in contrast with the rest of the TarzaNN project. Attempts were made to utilize open-source cross-platform imaging tools from OpenCV, but it was found that the OpenCV tools did not support the given camera model in a Windows 7 environment. In the future it is hoped to remove this reliance on proprietary software in order to ensure full cross-platform functionality of the TarzaNN framework.

3.7.2 PTU Control

The active visual component of this thesis utilizes the pan-tilt unit (PTU) described in Section 3.2.1. Control of the PTU is executed by the saliency output feature plane described in Section 3.6, and makes use of a programming API provided with the PTU. Unlike the proprietary software involved in the camera control, this programming API is cross-platform, and so there is currently no requirement to replace it with an alternative method of cross-platform control.

3.8 Environment Control

The environment control is a singleton class introduced to the TarzaNN framework designed to encapsulate all necessary environmental parameters and data. The class was introduced late in the design process of this thesis, and thus it has been only partially integrated into the network when operating in a physical environment (the elements described in Section 3.7 will be controlled through environment control in the future rather than directly from the network feature planes). The environment control class does have a fully implemented virtual environment, however, which allows the system to simulate an active visual system over a pre-loaded image.

3.8.1 Intent of Design

Creating a dedicated environment control class in the TarzaNN framework is intended to simplify the process involved in designing and implementing a TarzaNN network for a particular task setup. As a singleton class the system will be sure that all environment calls interact with the same class, which provides a straightforward method for parameters such as pixel-to-degree conversion ratios to be defined at run-time (when the environment is set up) but still globally constant (in the past such parameters have been included as defined macros in TarzaNN, but this forces users to modify elements of the code when changing between environments).

More importantly, feature planes which must interact with the environment, whether it is input feature planes acquiring information or feature planes outputting control signals to physical or virtual hardware, can be programmed with a generic input or output function call to the environment control object. If a new environmental setup is desired (for example, due to the acquisition of a new camera), the only code which would have to be extended to interact with the new hardware would be the environment class, which should greatly increase development speed.

3.8.1.1 Virtual Environment

The environment control class grew out of the original need for a virtual environment which could run simulations of the network. Operating in a virtual environment allows for faster and more repeatable experimentation, as well as the use of common data sets which have been used by other systems (see Section 4.1.2). When the virtual environment is created, a specified image file is loaded; this image constitutes the "environment" in which the network will now operate. When an image is requested by the input camera feature plane, a virtual camera with specified height and width properties will sample a portion of the environment image based on the current fixation coordinates (these are set by default to the centre of the environment image at the start of execution). When a command is given to saccade to a new target in the image, the fixation coordinates are updated and a new input image is sampled from the environment image. If any portion of the desired camera image is outside the environment image bounds, it is padded with the average value of all the environment edge pixels.

3.8.1.2 Data Logging

Another important element of the environment control class is the introduction of data logging tools. Current data logging options include a history of fixation command sequences as well as the central response network's decision regarding the central visual field (if applicable). When TarzaNN finishes execution of a network, a call is automatically made for the environment control class to save its recorded data in a new file if data logging is turned on.

3.8.2 Future Work

As previously mentioned, the environment control class was introduced relatively late in the thesis design process, and thus has not been completely integrated into the TarzaNN framework. In addition to not yet completely encapsulating control over the supported physical hardware, the use of the environment control class is currently restricted to predefined networks, as support for the environment setup has not yet been added to the Network Parser which handles networks defined in XML. Likewise, the ability to modify the environment control class through the network editor in TarzaNN has not yet been implemented.

3.9 Summary

This chapter overviewed TarzaNN, the computational framework with which this thesis was implemented, in Section 3.3. Additionally, the hardware details of the physical camera and PTU which the system has been implemented to control are described in Sections 3.2 and 3.7. The software contributions of this thesis have predominantly consisted of programming a number of extensions to the generic filter and feature plane object classes in TarzaNN, as described in Sections 3.4-3.6. An additional software component to extend the TarzaNN framework, the environment control class, is discussed in Section 3.8. The environment control class is designed to provide a generic control module for any active visual system constructed within TarzaNN, thereby greatly increasing the speed and ease with which new active control networks or hardware configurations may be implemented.

Chapter 4

Empirical Evaluation

4.1 Experimental Overview

Two experimental methodologies were used to evaluate the system developed for this thesis: a visual search task using physical hardware, and an eye-tracking task in a virtual environment utilizing a database of natural images. The physical experiment setup is described in Section 4.1.1, and the results are presented in Section 4.2.1. The virtual experiment setup is described in Section 4.1.2, and the results are presented in Section 4.2.2.

The physical task was designed to be a proof-of-principle experiment which would demonstrate a completed ST network functioning on a physical platform. The system should be capable of efficiently locating a visual target among distractors, with particular interest in the case when it is faced with an experimental field greater in size than the visual field of the camera. There were no top-down strategies or considerations beyond the central field target recognition to guide the search, which means that there are certain configurations of targets and distractors in which the system will fail; nevertheless performance should be significantly better than the brute force approach in the average case.

The virtual experiment was selected to explore how well peripherally guided fixations would evaluate against psychophysical data for human fixations in comparison to AIM. AIM was chosen as a benchmark both since it is provides the basis on which the peripheral priority map of this thesis was developed, but also because it was shown by Borji *et al.* to be among the top four saliency detection models of the thirty four tested for predicting human fixations, [6]. Therefore, a favourable performance compared to AIM would place our system among the top performers of this algorithm class for fixation prediction.

4.1.1 Physical Experiment Setup

The first test of this thesis work is to demonstrate a fully integrated ST network operating in a physical environment with equipment described in Section 3.2. This experiment is meant primarily as a proof of principle for the system behaviour, and thus uses a well-established visual search task of searching for an "+" or "x" amongst "-"s (see Figure 4.1). For the purposes of this thesis, the term *experiment field* refers to the region in which targets and distractors could be placed, while the *visual field* is the portion of the acquired image over which either peripheral or central processing is defined.

The target orientation was randomly determined to be either a "+" or an "x" on each trial. Placement of all elements in the visual field were randomly determined using a pseudorandom number generator in MATLAB. Any element which was placed with its centre point within 3.5 cm of another element was shifted to a new random location (this was roughly the length of the average distractor). The image acquired by the camera covered a 32×32 cm^2 region of the experiment field, which was a visual field of approximately 32.75°. To facilitate processing speed, the camera was set to capture at half resolution,

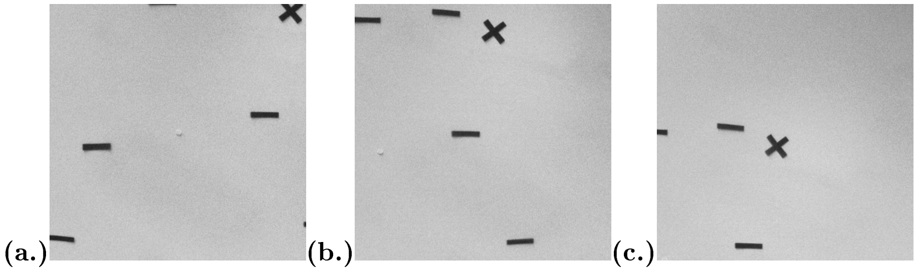


Figure 4.1 – Example sequence of images acquired during a visual search trial. This specific image sequence was from Trial 1 of Experiment 2 (see Section 4.1.1.3) **(a.)** Initial fixation, the target x is visible in the upper right corner but outside of the peripheral processing boundary. **(b.)** First fixation on a distractor, the target x is now within the peripheral salience field. **(c.)** Final fixation with the target x centered.

producing images of size 384 × 384. Convolutions in the peripheral processing network resulted in some loss of defined output in the extreme periphery of the visual field, which ranged in size from $20 - 47$ pixels ($1.7° - 4°$) from the image edge depending on the degree of smoothing (see Section 4.2.1.2 for a discussion of salience smoothing). The central region was defined as a square $10° \times 10°$ centered on the central fixation point. The number of trials conducted was rather low (20 overall, 10 at each experiment field size), largely due to prohibitive time required to accurately set up each trial.

4.1.1.1 Target Recognition in the Central Field

A recognition network for identifying the presence of a target "x" or "+" was produced by a two-stage network. The first layer consists of "W-filter" edge detectors at four orientations: 0°, 45°, 90°, and 135°, [90]. The outputs of these edge detectors are then passed to a recognition layer which smooths the output of the edge detectors via a Gaussian filter and combines the results of the vertical and angled edges in a spatially downsampled feature plane. A threshold is applied to this output and the target is reported as found if any region is found with a response above threshold.

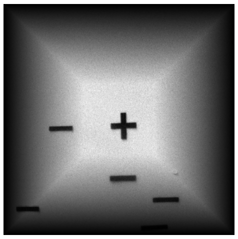


Figure 4.2 – Example of the masked image provided to the central field recognition network.

This is obviously a fairly simple network to operate in a problem domain specifically chosen for its straightforwardness as a proof of principle functional demonstration. Therefore, in order to attempt to still capture the essence of the Boundary Problem, a linear mask was applied to the input image which gradually decreased the pixel activation outside the central field (see Figure 4.2 for an example). It was felt that such a mask would effectively capture a sense of gradually reduced recognition capacity as one moved further from the unmodified central field. The mask was generally effective at reducing target detection outside of the central field, although targets just outside the central field were still occasionally detected (see Section 4.2.1 for a full discussion).

4.1.1.2 Experiment 1: Equal Dimensions of Visual Field and Experiment Field

The first physical experiment was performed with an experiment field of the same size as the visual field over which peripheral results were defined. This was done with minimal smoothing of salience values (the salience map was run through a 16×16 Gaussian filter with a standard deviation of 3 pixels) in order to try and minimize the boundary effect in the system (see Section 4.2.1.2 for a discussion of smoothing). Taking into account the minimal boundary effect introduced by the log-Gabor filters and smoothing of salience values, the experiment field was 29×29 cm^2. Five trials were performed with six distractors, and five trials were performed with nine distractors. If the system functions optimally, each trial should end either on the initial fixation (if the target is placed within the initial central field) or after a single saccade (which should be targeted to bring the target into the central field). There should be no major performance difference based on the number of distractors, since the visual search task is an efficient one.

4.1.1.3 Experiment 2: Visual Field Dimensions Smaller than Experiment Field Dimensions

The second physical experiment was performed with an experimental field twice the size of the defined peripheral field in each dimension. Following the results of the first experiment, however, the size of the Gaussian smoothing kernel applied to the PPM output was increased to a 40×40 filter (see Section 4.2.1.2 for a discussion of why). Due to this increase in the degree of smoothing, the boundary effect of the system was correspondingly increased and the defined peripheral field was therefore reduced in size to 26×26 cm^2. Therefore, the total experimental field was 52×52 cm^2. Ten trials were performed with nine distractors. The number of distractors was kept constant through all trials due to the fact that the number of distractors visible in each frame would naturally vary based on the random object placement, and nine targets provided a reasonable probability of ensuring at least one distractor was in each quadrant of the experimental field without increasing the time required to set up each trial. Unlike the first experiment, the optimal number of fixations required to find the target is not easily quantified, but it should on average be below that required for a brute-force search with the central field.

4.1.2 Virtual Experiment Setup

A second type of experiment was run based on the public data-set provided by Judd *et al.* which provides a set of 1003 images at a decently high resolution (the longest dimension of each image was 1024 pixels, and the other dimension ranged from $405 - 1024$ pixels) with eye tracking fixation data for fifteen human subjects, [38, 37]. A randomly selected subset of 50 images was taken from this data-set on which to run our network (the full data-set will be run in the future, but batch testing has not yet been fully implemented in TarzaNN). Each image was loaded into the environment controller of the network (see Section 3.8), at which point it served as the entire experimental field in which the system would operate. Selecting the size of the visual field was somewhat challenging, since a specific conversion between pixels and degrees of visual field was not provided by Judd *et al.* Based on the experimental description provided in [38], it was estimated that the images were presented to human viewers at a resolution of $0.0274°$ per pixel. This meant that even the largest image dimension of 1024 pixels represented only approximately $28°$ of visual space, and thus the maximum eccentricity of image elements was never more than $14°$ from the initial fixation. In the future it therefore would be worthwhile to gather a data set of human eye-tracking data in which the peripheral field was represented to a much larger extent, but as a proof of concept it was decided to have the network subsample a 480×400 subregion of the image (representing $13° \times 11°$ of visual field). The central $2°$ of that image would be deemed the foveal fixation region, and thus all new fixations would represent saccades at least $1°$ degree from the previous fixation.

The fixations produced by the peripheral saliency map model of this thesis were compared against a set of fixations produced by the AIM algorithm run over the entire image. To produce the AIM fixations, AIM was run using the same filter set as the peripheral saliency map model and the most salient points which were at least 1° apart were recorded as the fixation points. The peripheral saliency map model should produce similar results to AIM, but the implicit central bias introduced by operating only over a fixated subregion of the image should yield fixation sequences which more closely resemble those of humans.

4.2 Experimental Results

4.2.1 Physical Experiment Results

4.2.1.1 Experiment 1 Results

By restricting target placement to be within the initial visual field, the system should ideally find the target within at most two fixations (it should either be found immediately on the initial central fixation or saccade directly to the target). However, as mentioned in the experiment description, this experiment was performed with only a small degree of smoothing, which resulted in a fairly high signal-to-noise ratio between the most salient points on the target and the most salient points on the distractors (see Section 4.2.1.2 for an in-depth discussion of smoothing). As a result, there were several trials in which the initial fixation was made to a distractor rather than the target; these are recorded as fixation errors. There were no recognition errors in these trials; each time the majority of the target was located within the central field, the recognition network found it.

Figures 4.3-4.7 display abstracted saccade paths for trials with six distractors, and Figures 4.8-4.12 display saccade paths for trials with nine distractors. Each image represents the experimental field, with the coordinates of the plane given in centimeters. Distractors are represented by blue rectangles while the target is represented by a red dotted circle displaying the extent of the target

Trial Number	1	2	3	4	5	**Avg.**	**Std. Dev.**
Number of Fixations	2	3	3	2	1	2.2	0.8
Fixation Errors	0	1	1	0	0	0.4	0.5

Table 4.1 – Chart of the results for six distractors. The number of fixations for each trial and the number of fixation errors are reported.

Trial Number	1	2	3	4	5	**Avg.**	**Std. Dev.**
Number of Fixations	2	3	2	4	1	2.4	1.1
Fixation Errors	0	1	0	1	0	0.4	0.5

Table 4.2 – Chart of the results for nine distractors. The number of fixations for each trial and the number of fixation errors are reported.

(the character in the centre of the circle is an "×" or a "+" to match the orientation of the target on that particular trial). The extent of the peripheral field is represented by the dashed outer square, while the inner dashed green square represents the central recognition region. Fixations are represented by black circles, with arrows representing the transition between each fixation target.

Table 4.1 displays the results for trials with six distractors, while Table 4.2 displays the results for trials with nine distractors. As can be seen, the target was fixated first on half of the trials in which it did not start in the central fixation, and on all other trials it was only missed by a single fixation error. Thus, the target was being assigned the highest peripheral priority with a likelihood significantly greater than chance. Nevertheless, the occurrence of fixation errors at all is a concern. Although the majority of fixation errors were immediately corrected and therefore resulted only in a single extra fixation, the fourth trial with nine distractors required an extra fixation to bring the target back within the peripheral field (see Figure 4.11). It was actually a lucky occurrence that the third fixation brought the target back within the peripheral field, as once the target has moved beyond the bounds of the visual field there are no other control mechanisms in the system to either recognize the complete absence of the target nor to provide an appropriate strategy for bringing the target back within visual range. Thus, any fixation error has the potential to result in a far greater number of unnecessary fixations. The detrimental effect of fixation errors on system performance would likely only be exacerbated with a larger visual field, and thus steps were taken to attempt to minimize these errors by modifying the degree of salience smoothing.

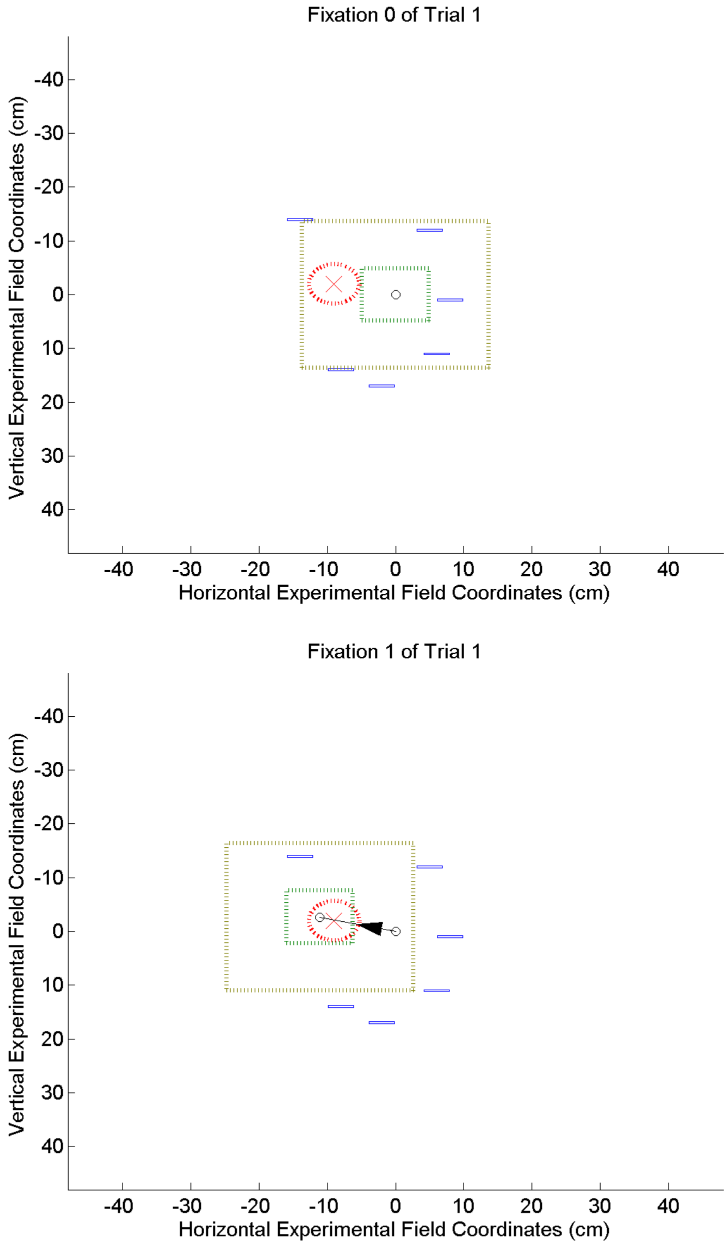
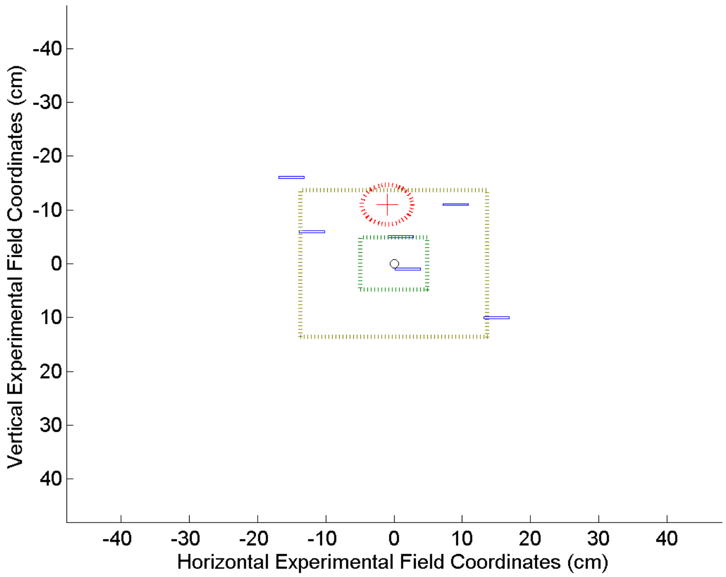
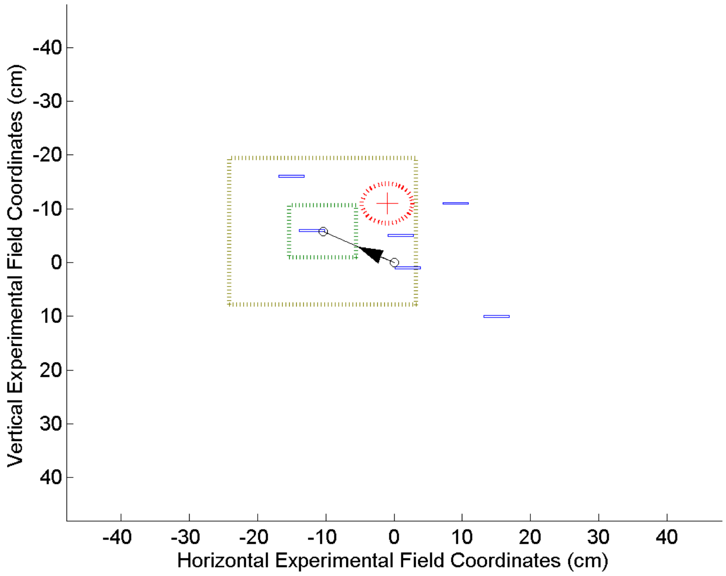


Figure 4.3 – Trial 1 with 6 distractors. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.

Fixation 0 of Trial 2



Fixation 1 of Trial 2



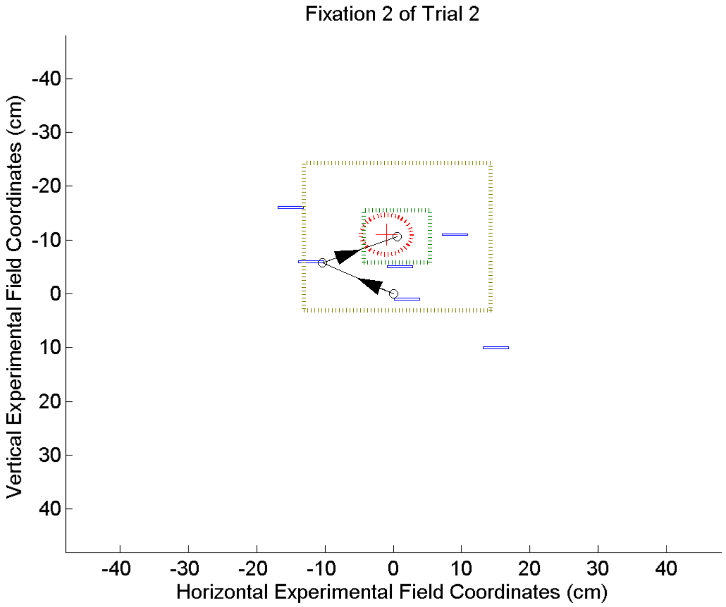
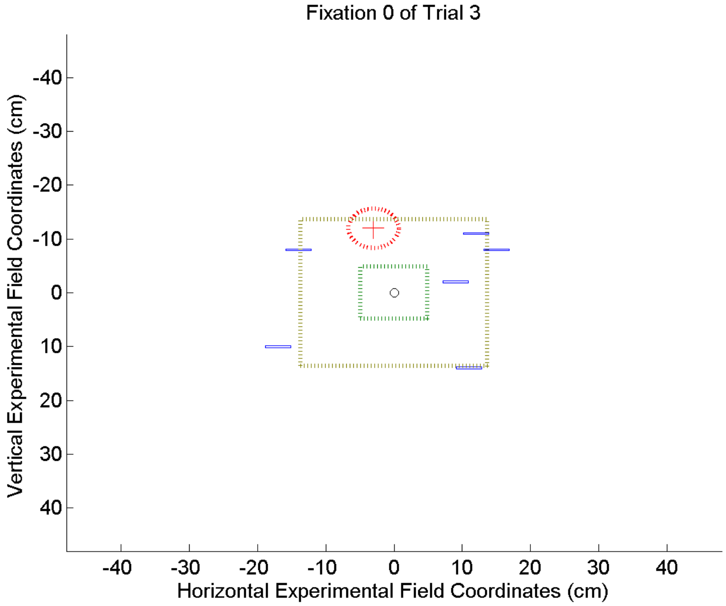
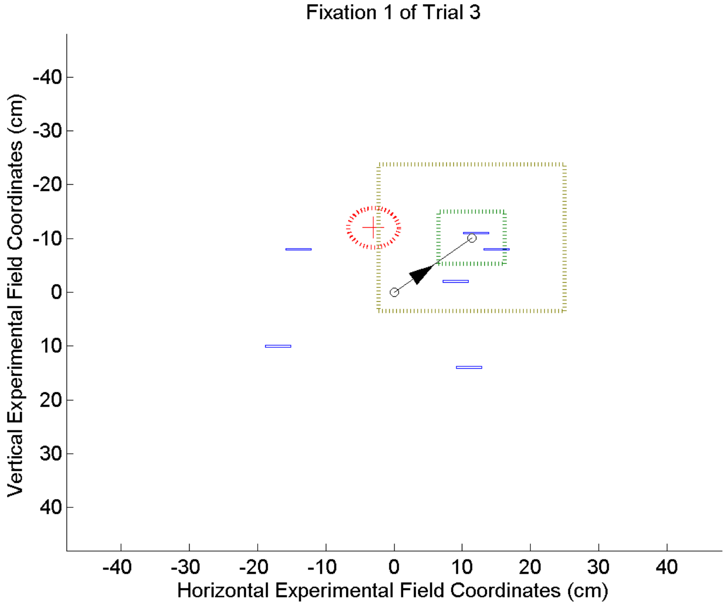


Figure 4.4 – Trial 2 with 6 distractors. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.


Fixation 0 of Trial 3


Fixation 1 of Trial 3

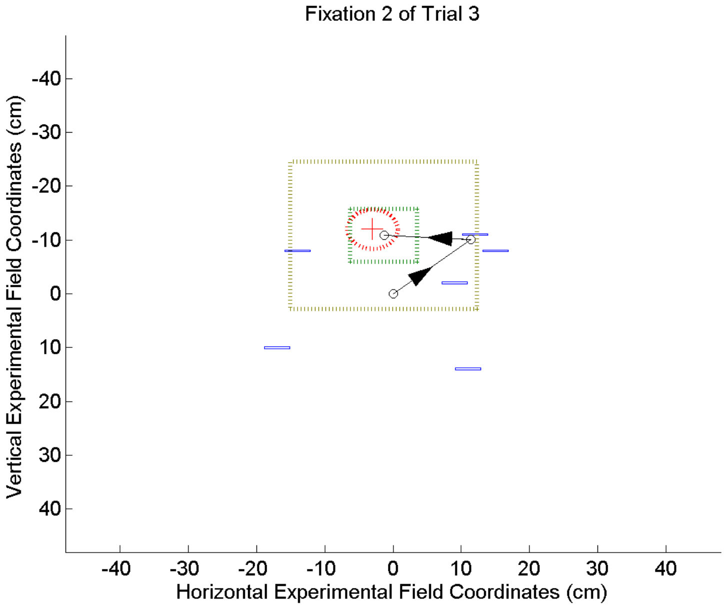
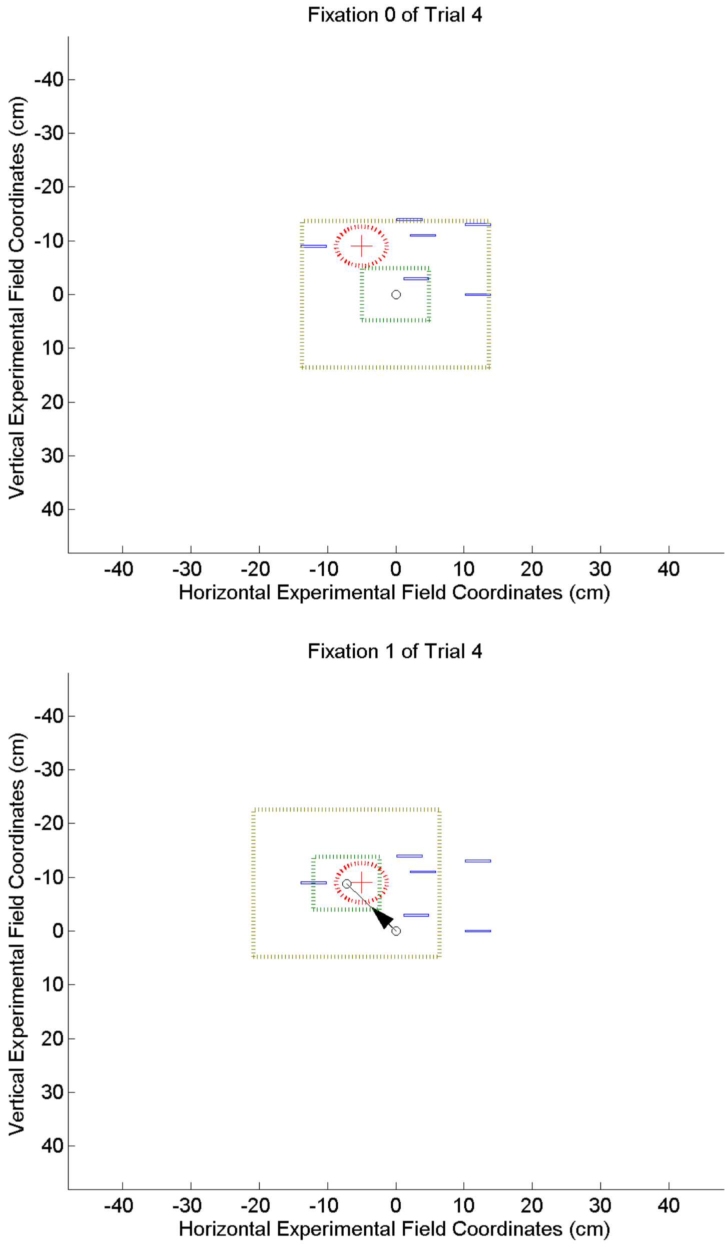


Figure 4.5 – Trial 3 with 6 distractors. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.



Fixation 0 of Trial 4

Fixation 1 of Trial 4

Figure 4.6 – Trial 4 with 6 distractors. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.

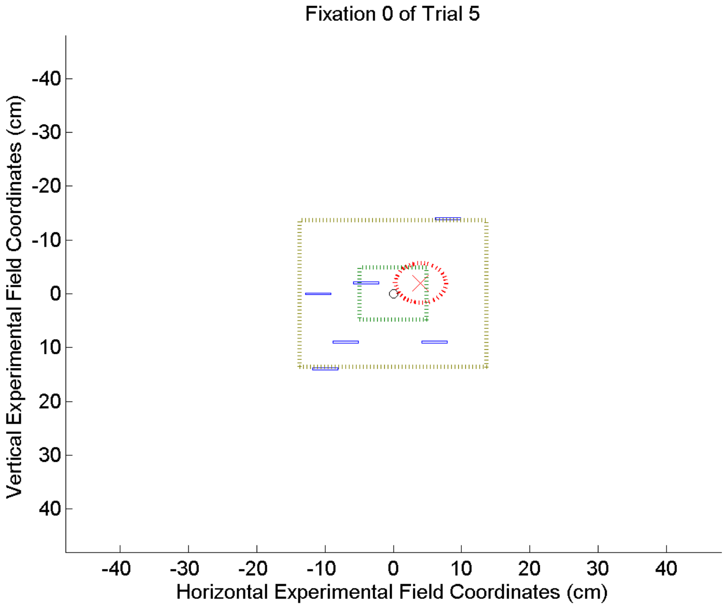


Figure 4.7 – Trial 5 with 6 distractors. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.

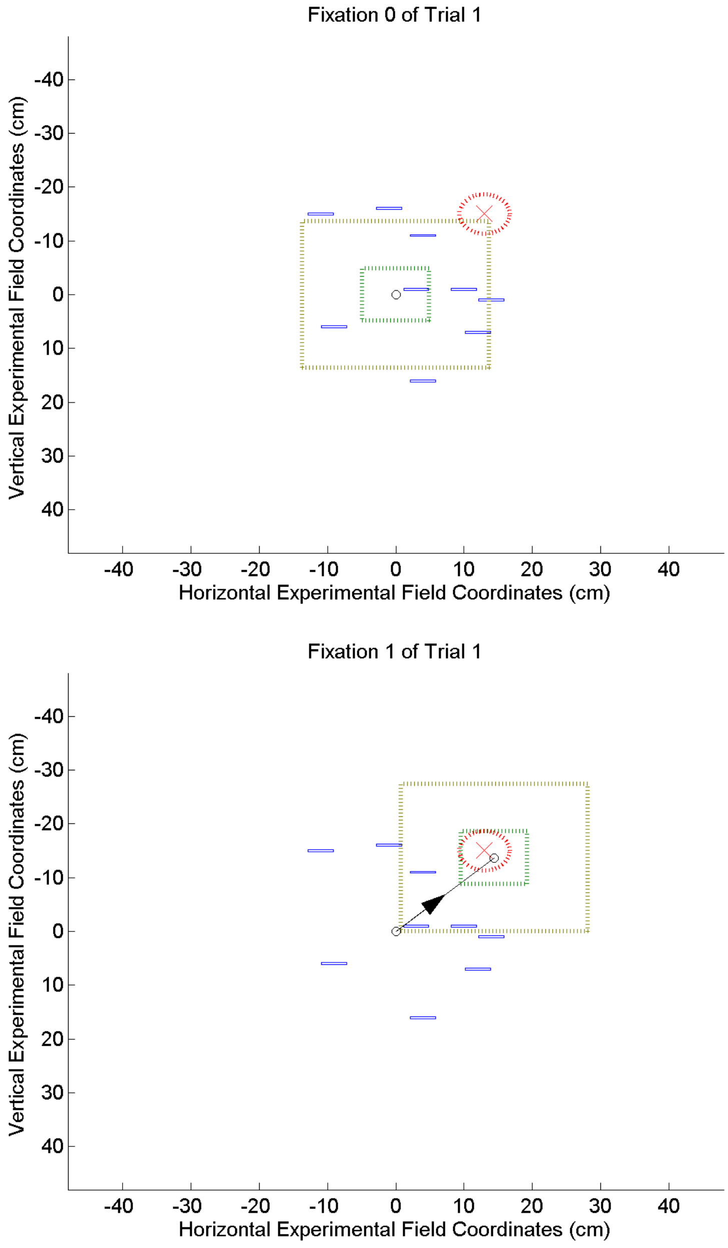
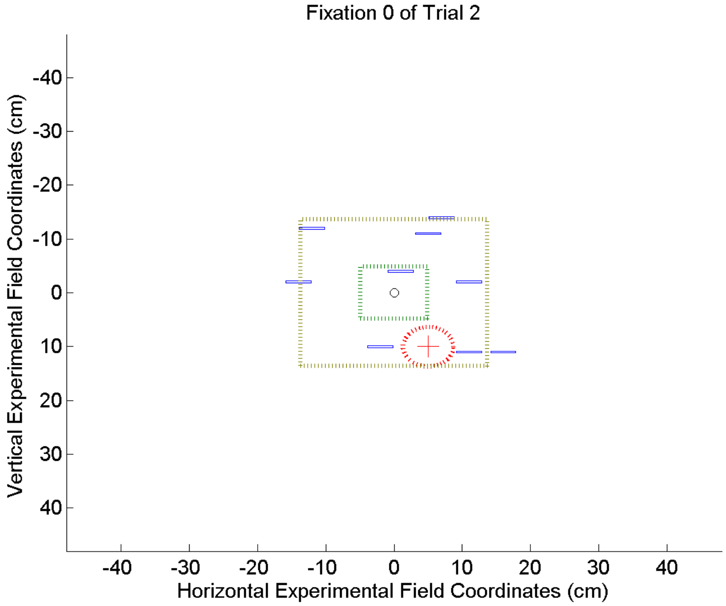
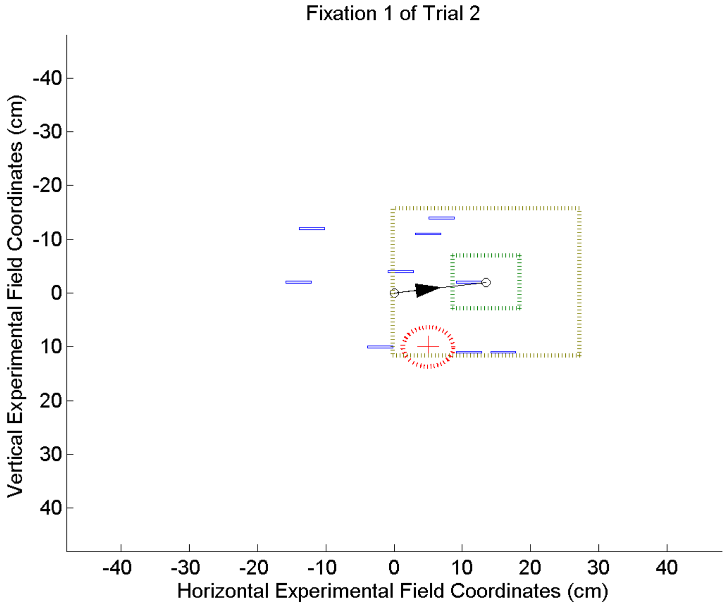


Figure 4.8 – Trial 1 with 9 distractors. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.


Fixation 0 of Trial 2
-40
-30
-20
-10
0
10
20
30
40
Vertical Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40
Horizontal Experimental Field Coordinates (cm)


Fixation 1 of Trial 2
-40
-30
-20
-10
0
10
20
30
40
Vertical Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40
Horizontal Experimental Field Coordinates (cm)

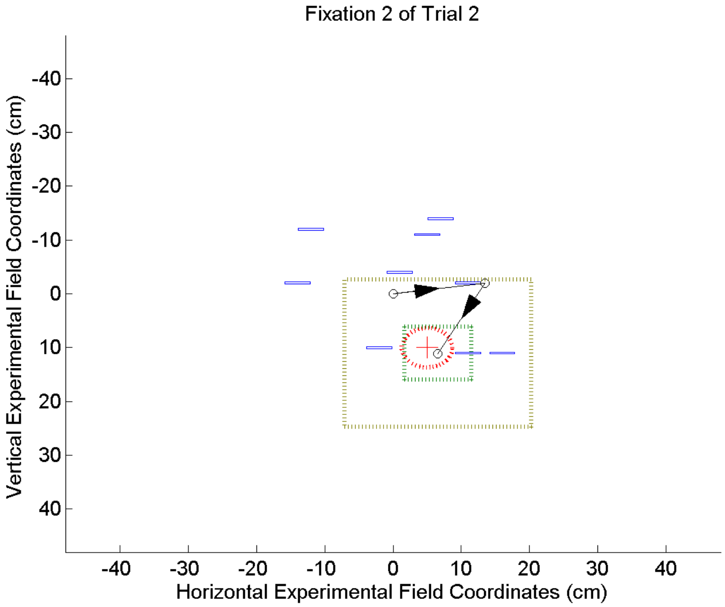


Figure 4.9 – Trial 2 with 9 distractors. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.

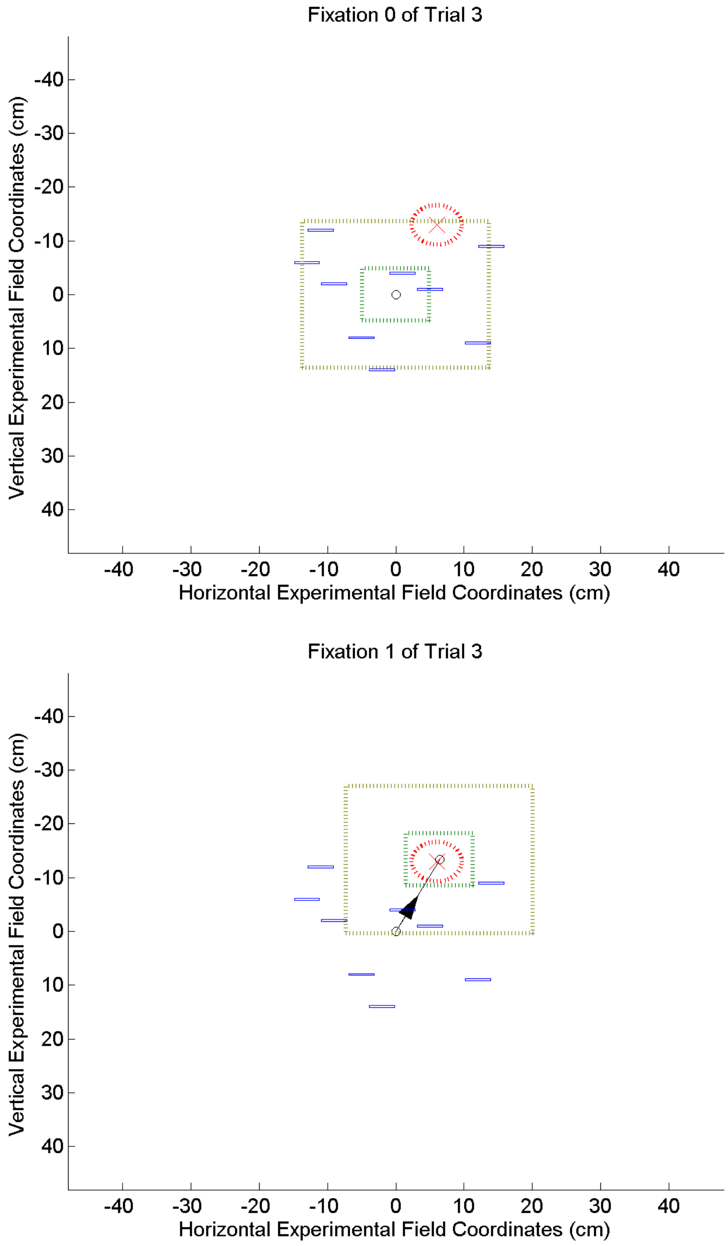
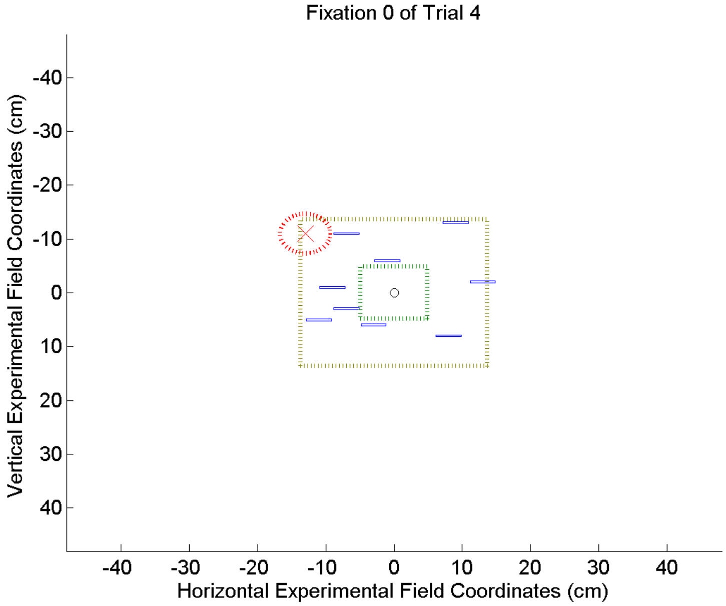
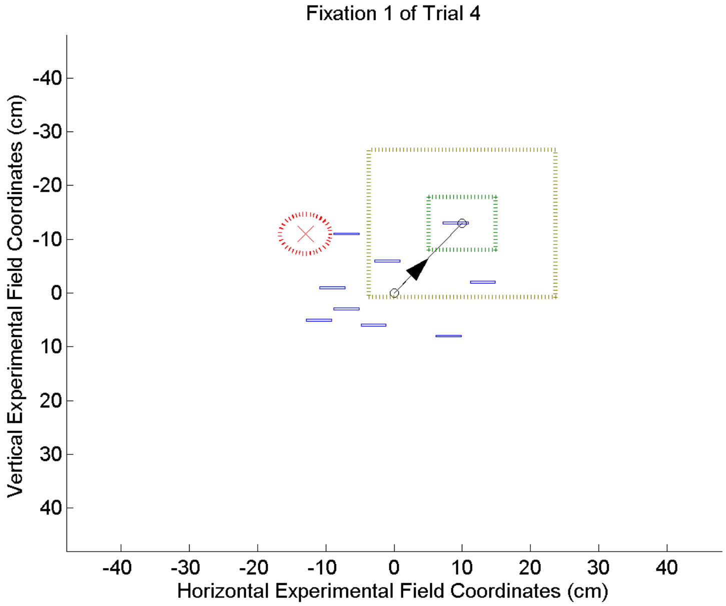
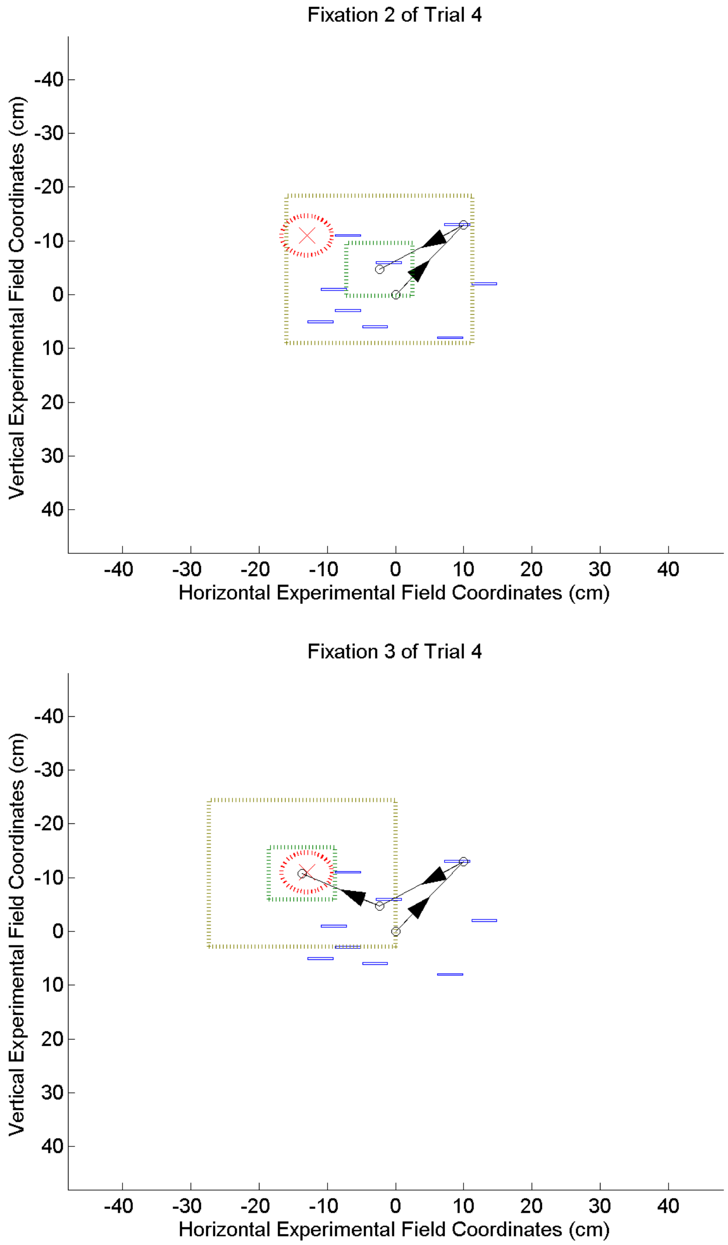


Figure 4.10 – Trial 3 with 9 distractors. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.


Fixation 0 of Trial 4
Vertical Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40
Horizontal Experimental Field Coordinates (cm)


Fixation 1 of Trial 4
Vertical Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40
Horizontal Experimental Field Coordinates (cm)


Fixation 2 of Trial 4

Fixation 3 of Trial 4

Figure 4.11 – Trial 4 with 9 distractors. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.

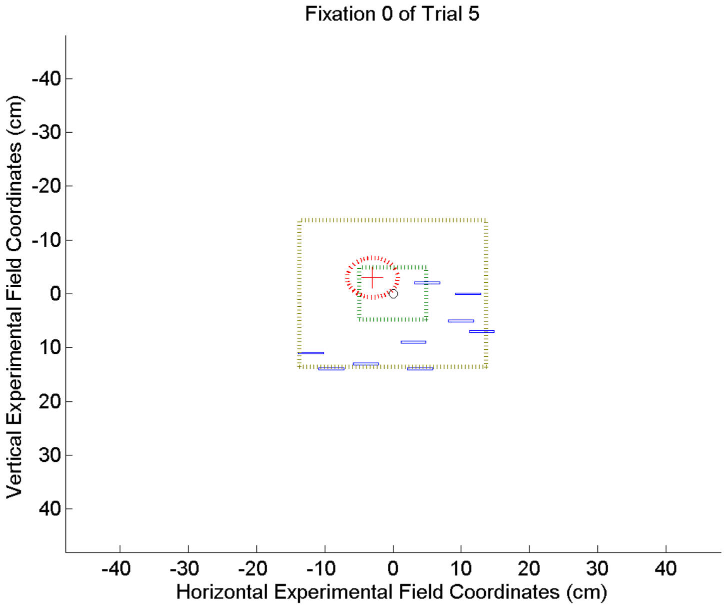


Figure 4.12 – Trial 5 with 9 distractors. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.

4.2.1.2 Smoothing of Salience Values

The number of peripheral errors in Experiment 1 was higher than would be desirable, despite the fact that none of the errors were responsible for a drastic number of extra fixations. Fixation errors were caused by erroneously selecting a distractor rather than the target for a fixation; it therefore became necessary to examine the saliency map being produced by the AIM algorithm.

As can be seen in Figure 4.13, the most salient components of the image are the corners of all the visual elements. This makes intuitive sense from the perspective of the AIM algorithm since corners will elicit an unusual response across a broad range of the log-Gabor filters. However, it creates a conceptual problem in that we would ideally like a saliency measure which is based on objects, rather than point-based (as was mentioned in Section 1.2.1.5, there is evidence that IOR can operate in an object-based fashion, which suggests the possibility that the human equivalent of salience may function in a similar manner). Likewise, it introduces a practical problem if the noisiness of real images means that the corners of distractor elements begin to approach the salience of the desired target. Despite individual distractor pixels displaying strongly salient elements, it is clear that the target possesses a greater concentration of these saliency “hot spots”. As can be seen in Figure 4.14, increasing the size of the smoothing convolution not only emphasizes this greater concentration, it draws the maximal salience in toward the centre of the object rather than selecting the object edge.

While a larger smoothing kernel provides more reliable fixation results in this specific search task, it does further increase the size of the boundary over which peripheral results are undefined. Additionally, too large a kernel would risk conflating the saliency response of nearby objects, possibly emphasizing two closely placed distractors over the desired target. In a more unconstrained

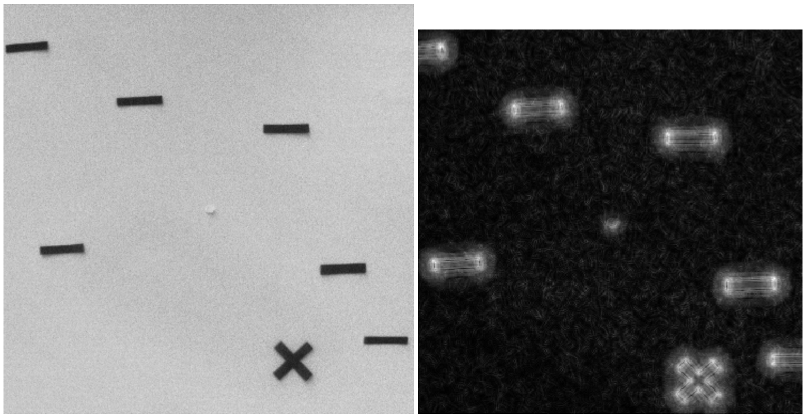


Figure 4.13 – Initial image fixation from Experiment 2 Trial 3 and its corresponding saliency map without smoothing.

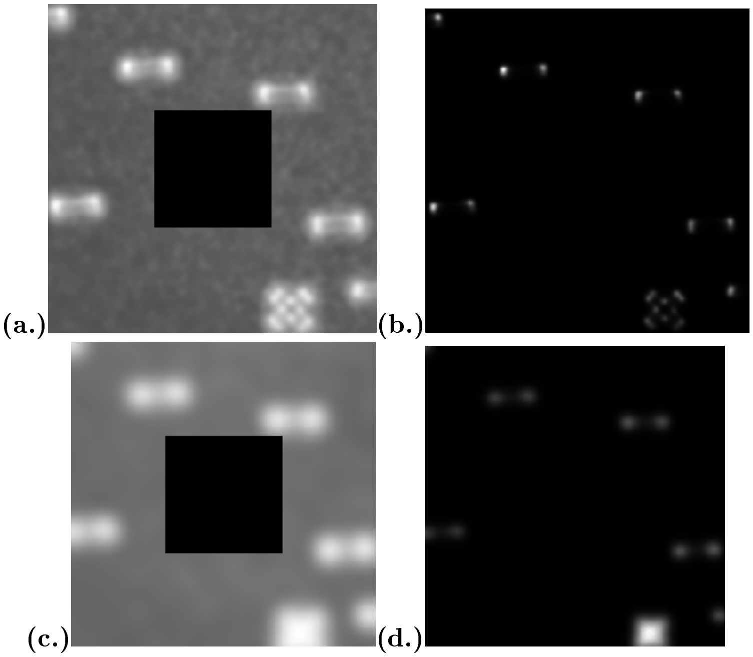


Figure 4.14 – Saliency maps from Figure 4.13 with Gaussian smoothing applied (the central visual field has been set to zero). **(a.)** Saliency output smoothed via a 16 × 16 Gaussian kernel with a 3 pixel standard deviation. This was the smoothing kernel used in Experiment 1. **(b.)** The same saliency map from (a.) raised to the tenth power in order to visually accentuate the most salient elements. **(c.)** Saliency output smoothed via a 40 × 40 Gaussian kernel with a 10 pixel standard deviation. **(d.)** The same saliency map from (c.) raised to the tenth power in order to visually accentuate the most salient elements.

environment, the size of the smoothing kernel biases the peripheral selection to objects closest in size to the spatial scale of the kernel (in this case, both the target and distractors were of the same spatial scale). It therefore would be worthwhile investigating an appropriate control strategy either to find an optimal smoothing parameter for natural scene viewing or, more interestingly, seek a method to tune priority on an object-basis rather than a point-basis.

4.2.1.3 Experiment 2 Results

Figures 4.16-4.25 display abstracted saccade paths for all Experiment 2 trials. As with the saccade paths displayed for Experiment 1, each image represents the experimental field with the coordinates of the plane given in centimeters. Distractors are represented by blue rectangles while the target is represented by a red dotted circle displaying the extent of the target (the character in the centre of the circle is an "×" or a "+" to match the orientation of the target on that particular trial). The extent of the peripheral field is represented by the dashed outer square, while the inner dashed green square represents the central recognition region. Fixations are represented by black circles, with arrows representing the transition between each fixation target.

By increasing the size of the smoothing kernel (see Section 4.2.1.2), there were no fixation errors in the second experiment. However, because the target was no longer guaranteed to be within the initial fixation's visual field, a number of trials required several fixations to distractors to first bring the target into the peripheral field, at which point a direct fixation was made and the trial completed. The number of fixations required to find the target are reported in Table 4.4. It should be noted that a recognition error occurred in Trial 7; the target was actually fixated first on the third fixation, but the recognition score was slightly below the necessary threshold (0.46 instead of the required 0.5) and the network therefore continued looking for the target. After saccading away, it

Trial Number	1	2	3	4	5	6	7	8	9	10	**Avg.**	**Std. Dev.**
Number of Fixations	6	3	2	1	2	2	5*	6	2	3	3.2	1.8

Table 4.4 – Results for physical experiment 2. Trial 7's results (denoted with a *) are the number of fixations to recognize the target; the target was fixated first on the third fixation, but a recognition error prevented it from being located. Attention returned to the target again on the fifth fixation, at which point it was properly recognized. There were no fixation selection errors.

luckily returned to fixate a nearby distractor which was sufficiently close to the target for it to be recognized on the second try.

As expected, the results are generally better than a systematic grid search would be at a resolution of the central field but are not guaranteed to find the target in all possible experimental setups. A brute force search would, assuming some overlap between each fixation to avoid missing a target on a frame edge, require approximately six fixations in each direction, or a total of thirty six fixations to cover the entire experimental field. Given the uniform probability of placing the target anywhere in the experiment field, such a naive search would therefore be expected to take on average eighteen fixations to find the target. Although the current system performs much better than the naive approach, its reliance on what is essentially a random search between distractors until the target enters the visual field means that it is possible for the system to fail completely to find the target. An example of a setup in which this is the case is shown in Figure 4.15(b.), which is a customized experimental layout based on swapping the position of the target with an isolated distractor in the upper right corner of the experiment field from the experimental setup of Trial 9 (shown in Figure 4.15(a.)). The new target position placed it outside the peripheral field of the initial fixation, as well as outside the periphery of any distractor-centered fixations. Therefore, in order to avoid failing on unusual (but entirely possible) experimental setups such as this one, the system would

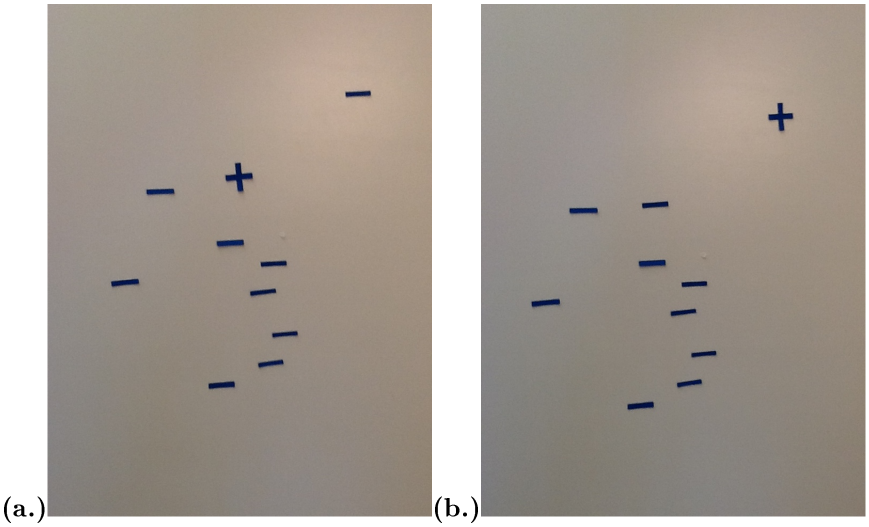
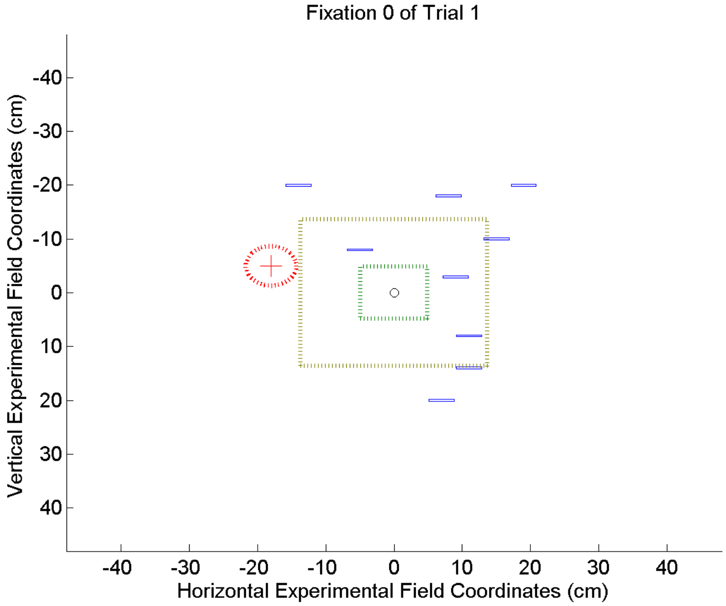
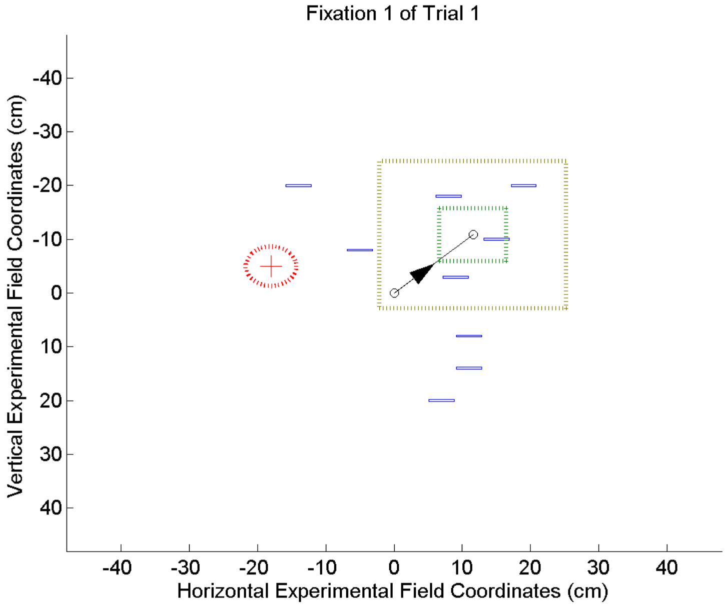
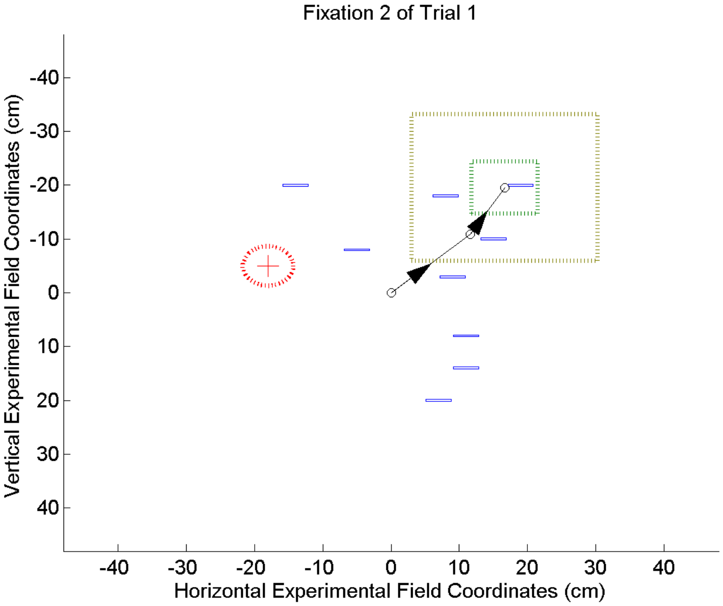


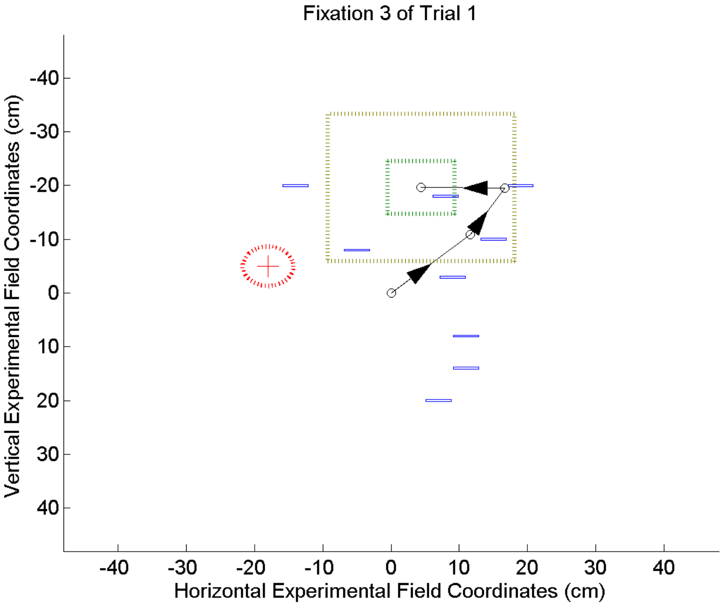
Figure 4.15 – Example images of a trial in which the system fails to find the target. **(a.)** The search setup from Trial 9 for the entire experiment field. **(b.)** The same setup as in (a.), except with the target position swapped with the upper right distractor. Once the target is sufficiently isolated from all other elements in the experiment field, it can no longer be found.

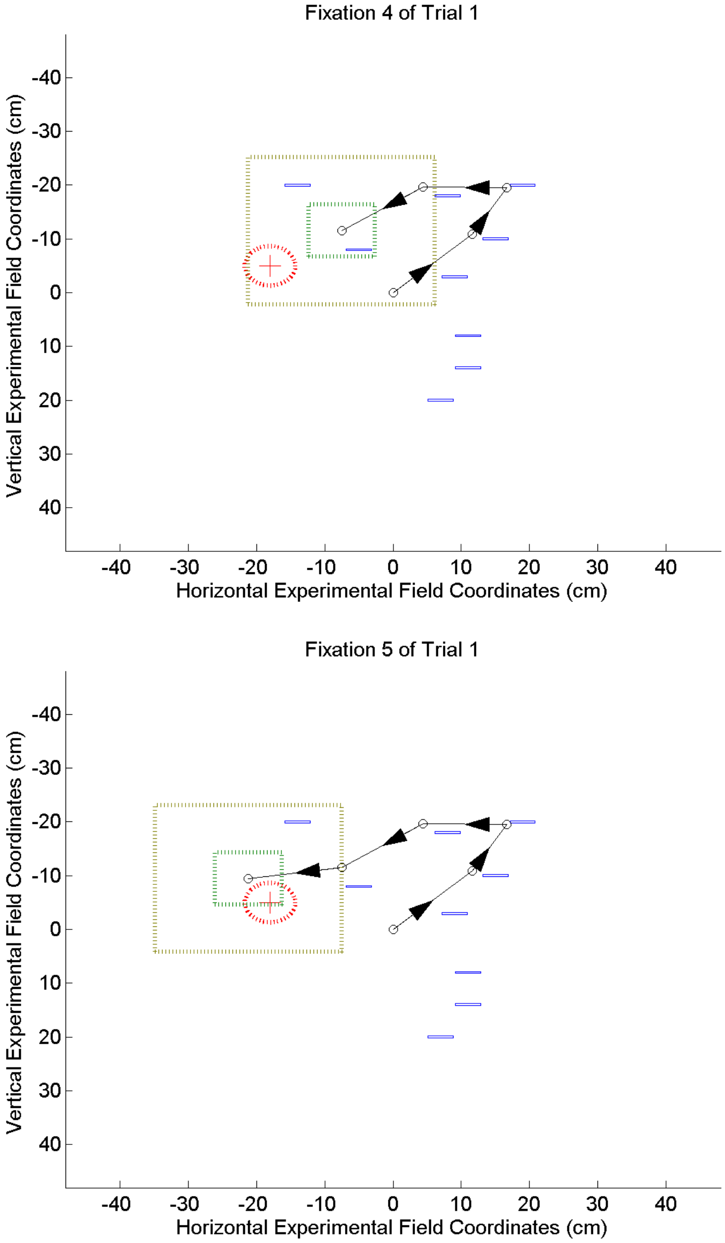
need to incorporate further contextual task information into its control, such as by using a probability field similar to that used in the SYT algorithm discussed in Section 1.2.3.3, [70].


Fixation 0 of Trial 1
Vertical Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40
-40
-30
-20
-10
0
10
20
30
40
Horizontal Experimental Field Coordinates (cm)


Fixation 1 of Trial 1
Vertical Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40
-40
-30
-20
-10
0
10
20
30
40
Horizontal Experimental Field Coordinates (cm)


Fixation 2 of Trial 1
Vertical Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40
Horizontal Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40

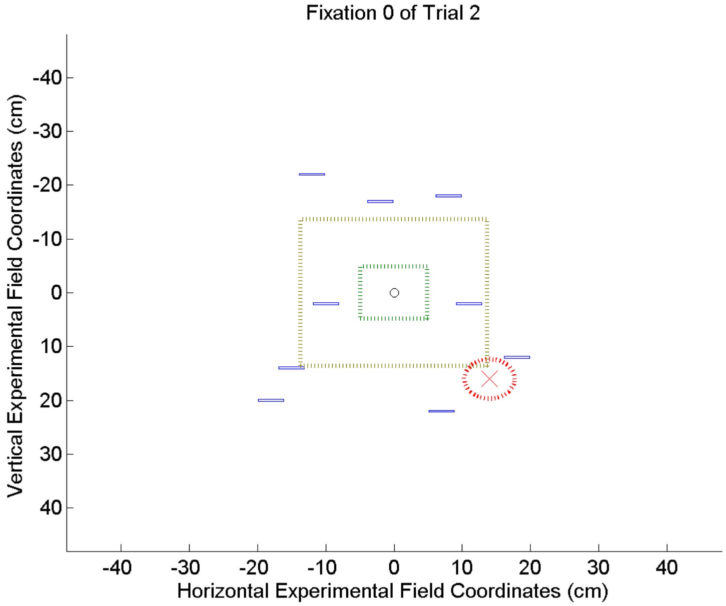

Fixation 3 of Trial 1
Vertical Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40
Horizontal Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40

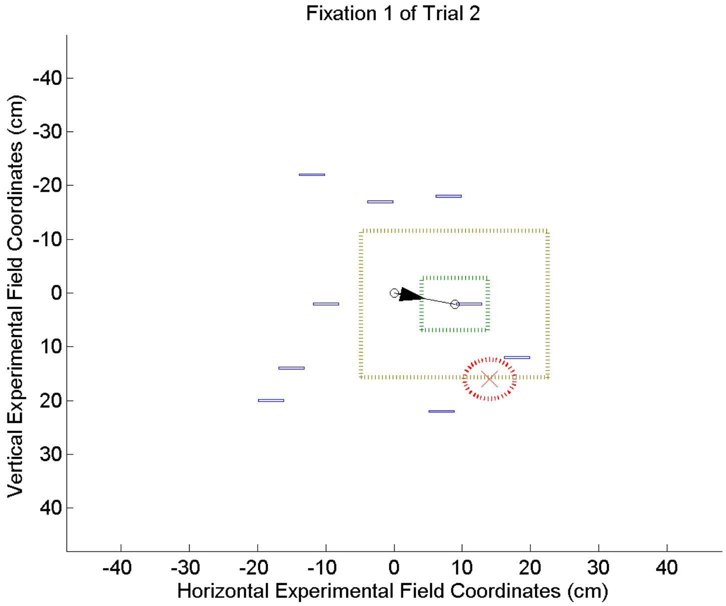


Fixation 4 of Trial 1

Fixation 5 of Trial 1

Figure 4.16 – Trial 1 Experiment 2. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.


Fixation 0 of Trial 2
Vertical Experimental Field Coordinates (cm)
Horizontal Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40


Fixation 1 of Trial 2
Vertical Experimental Field Coordinates (cm)
Horizontal Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40

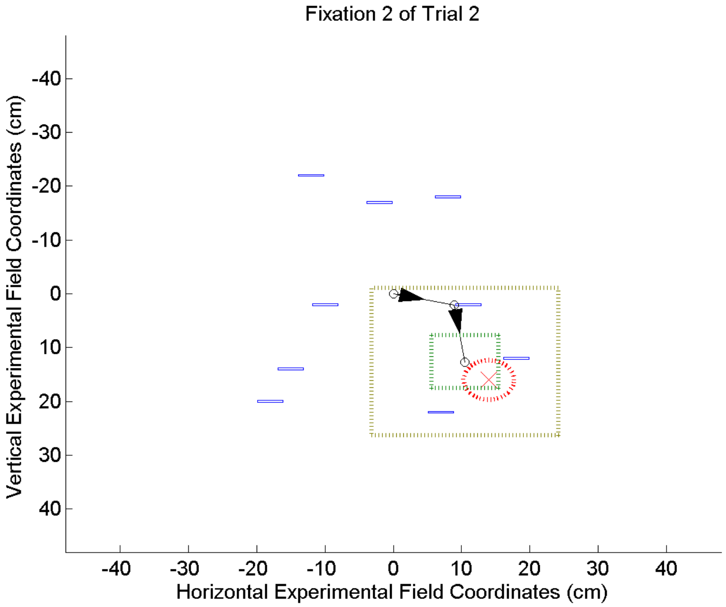


Figure 4.17 – Trial 2 Experiment 2. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.

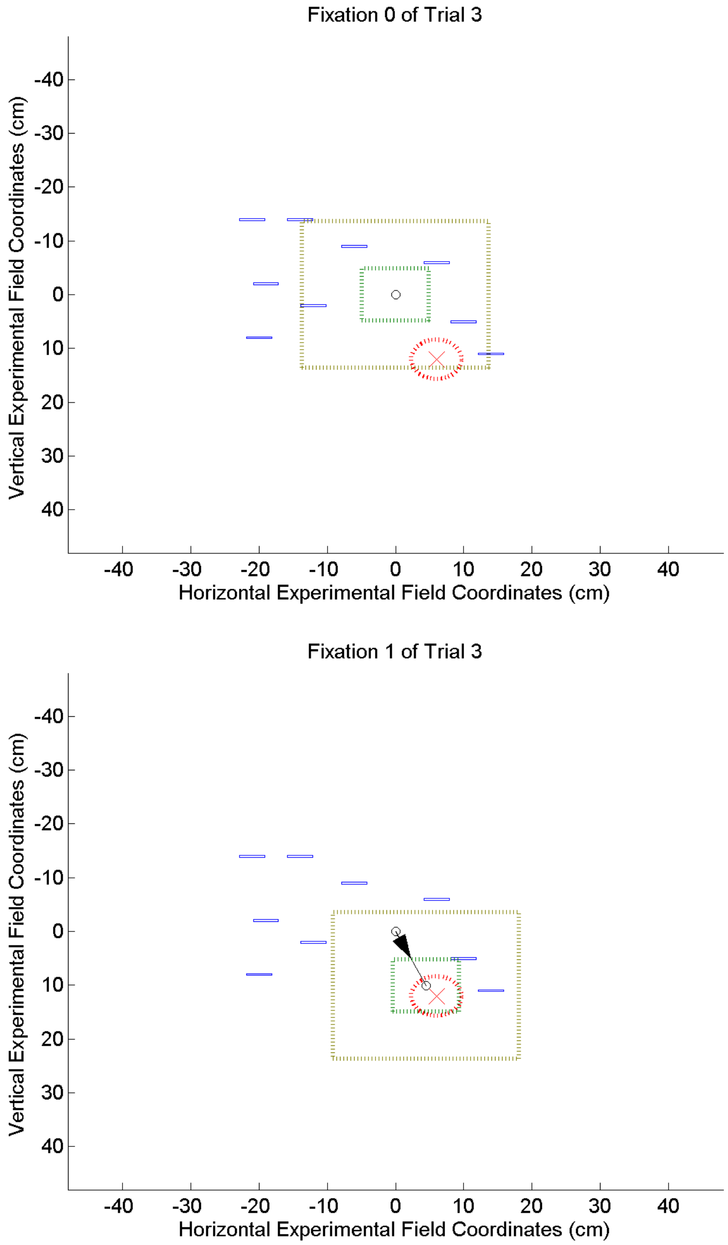


Figure 4.18 – Trial 3 Experiment 2. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.

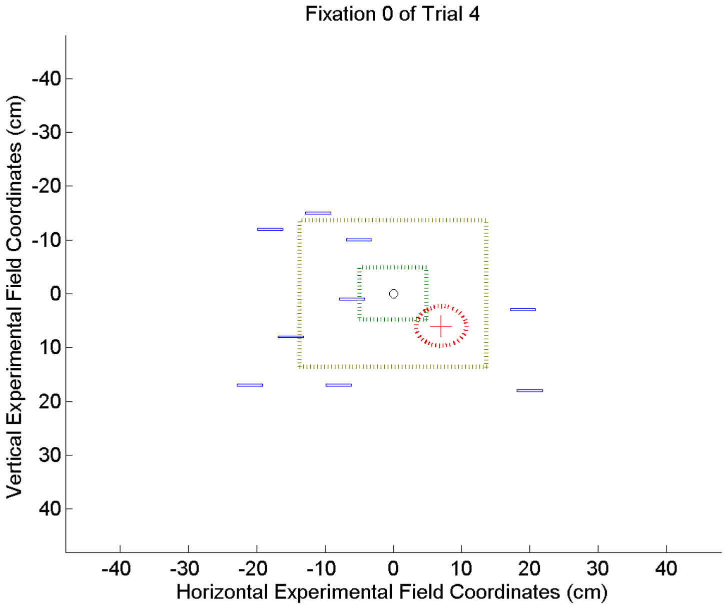


Figure 4.19 – Trial 4 Experiment 2. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.

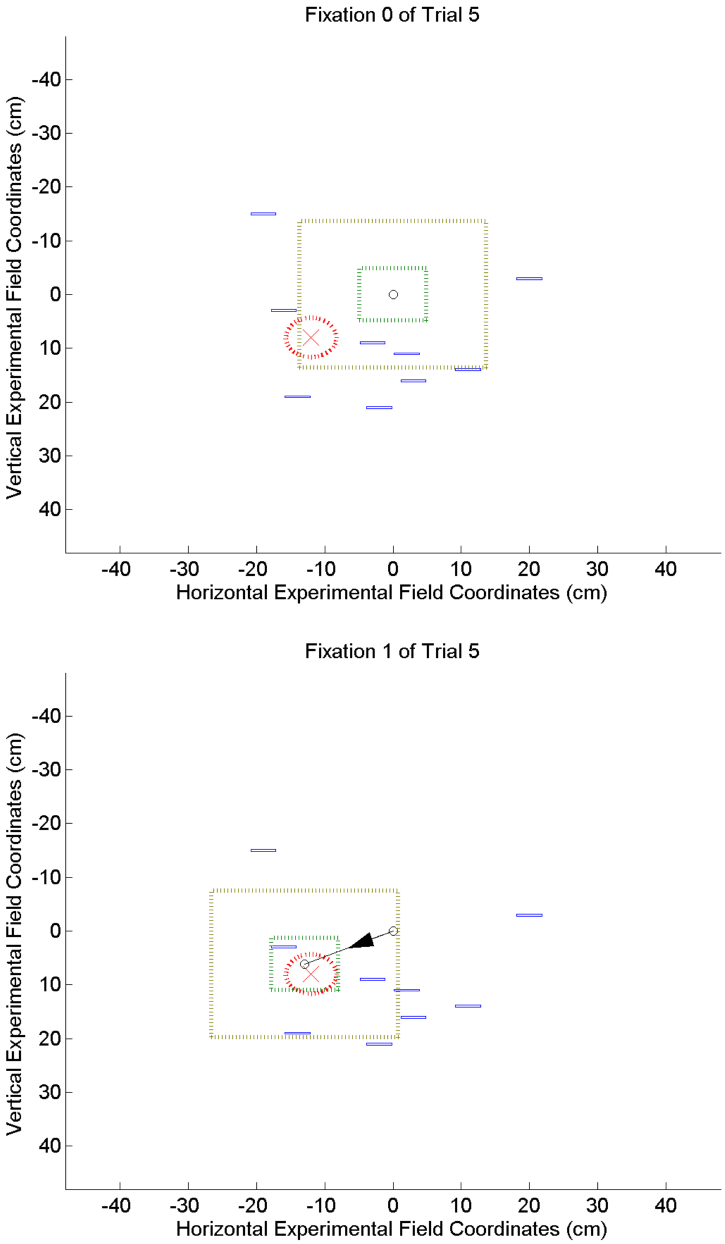
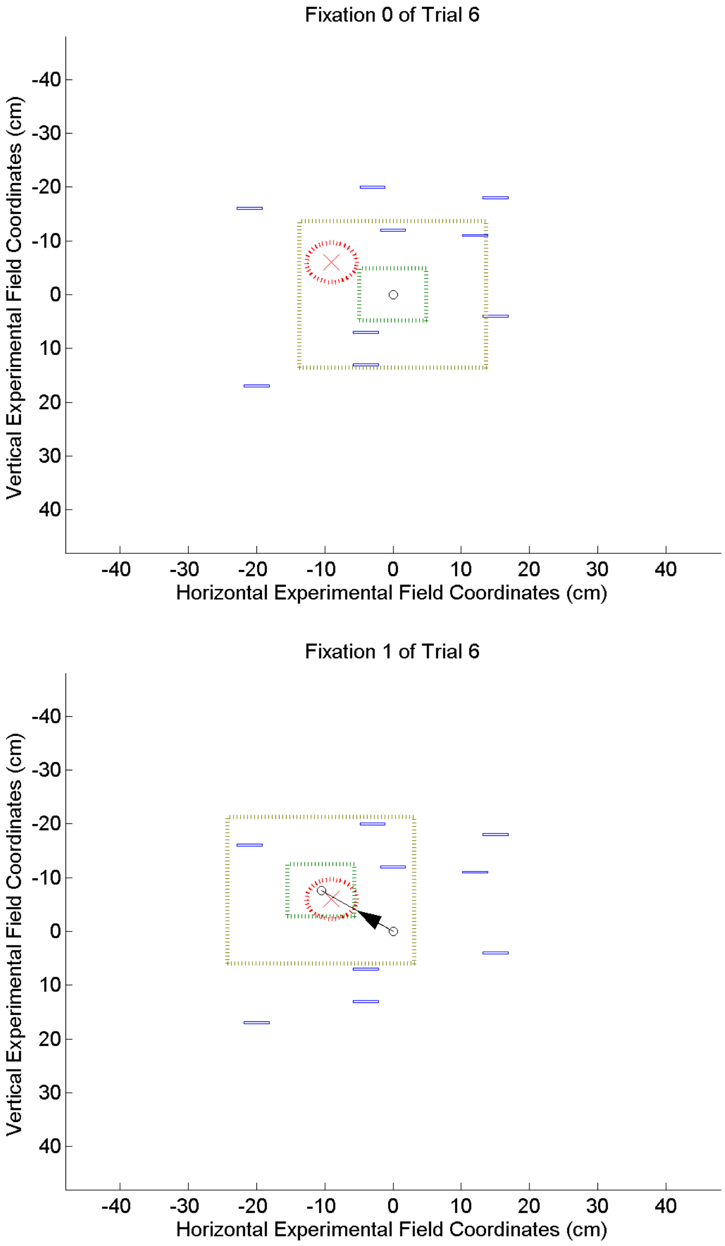


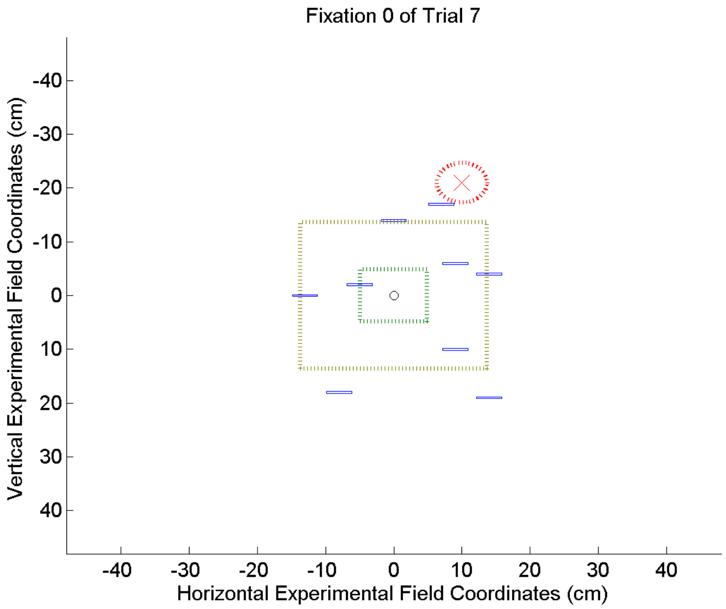
Figure 4.20 – Trial 5 Experiment 2. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.

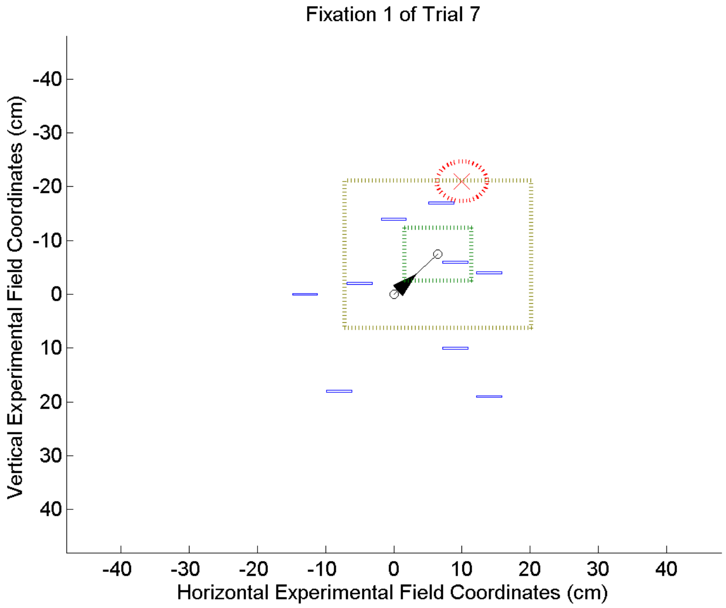


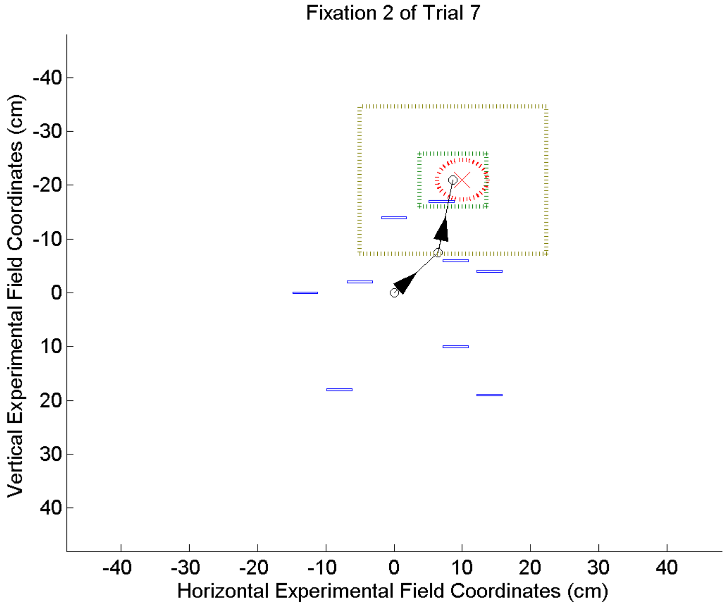
Fixation 0 of Trial 6

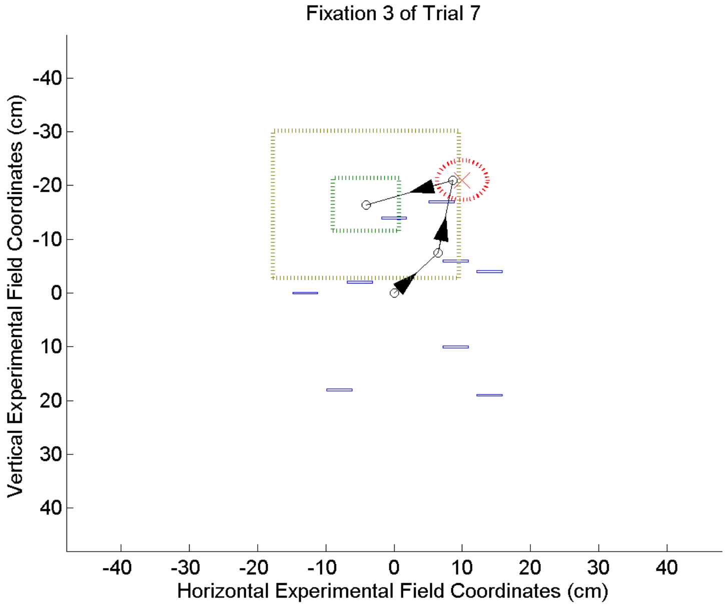
Fixation 1 of Trial 6

Figure 4.21 – Trial 6 Experiment 2. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.


Fixation 0 of Trial 7
Vertical Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40
Horizontal Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40


Fixation 1 of Trial 7
Vertical Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40
Horizontal Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40


Fixation 2 of Trial 7
Vertical Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40
Horizontal Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40


Fixation 3 of Trial 7
Vertical Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40
Horizontal Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40

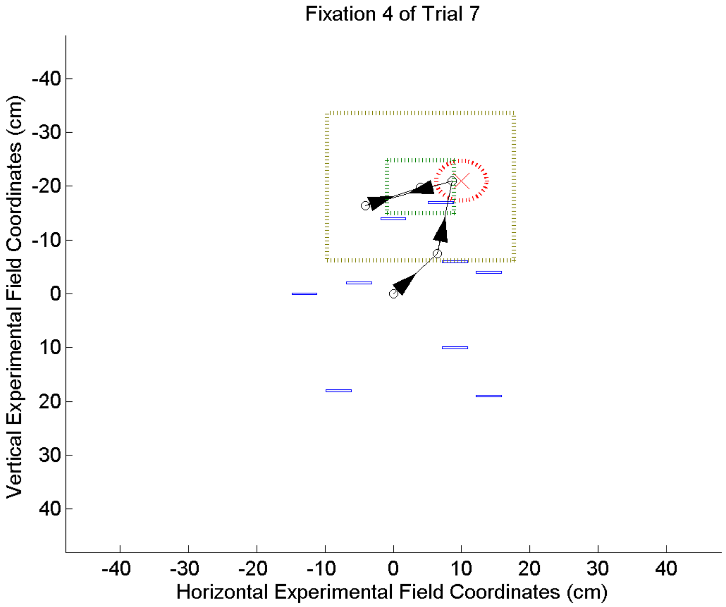
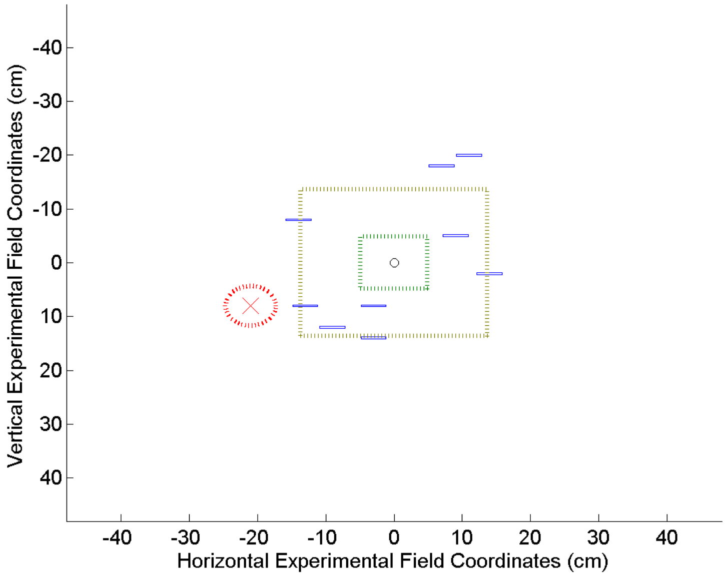
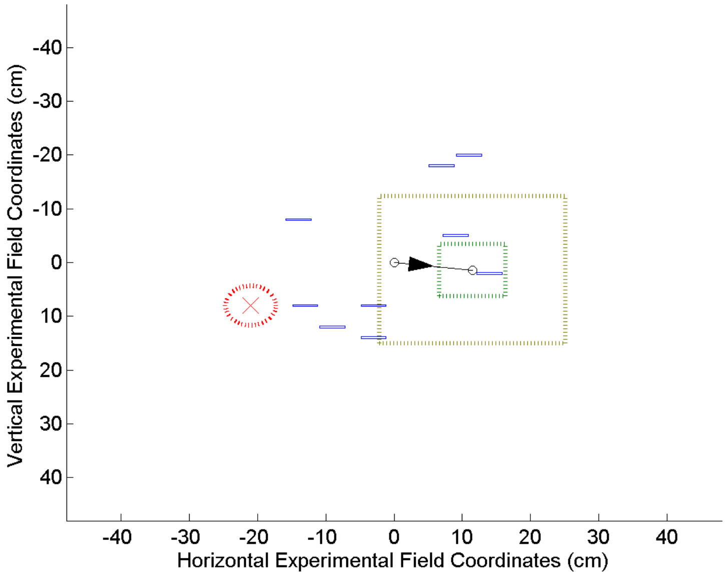


Figure 4.22 – Trial 7 Experiment 2. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.

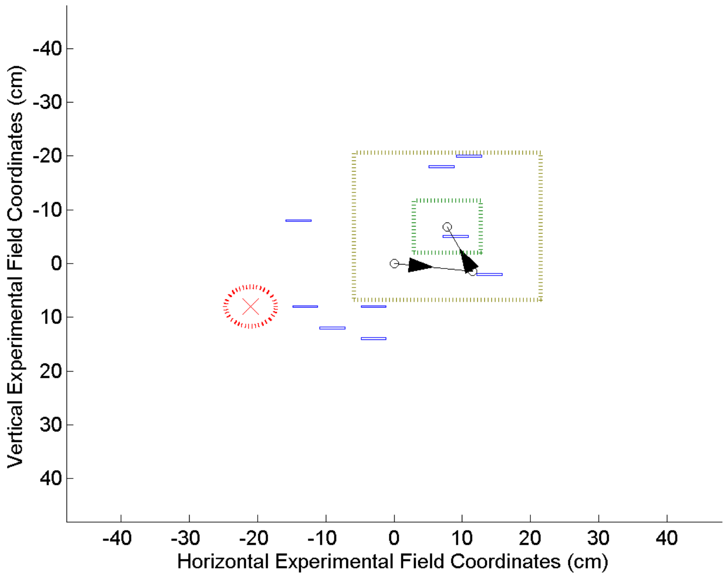
Fixation 0 of Trial 8



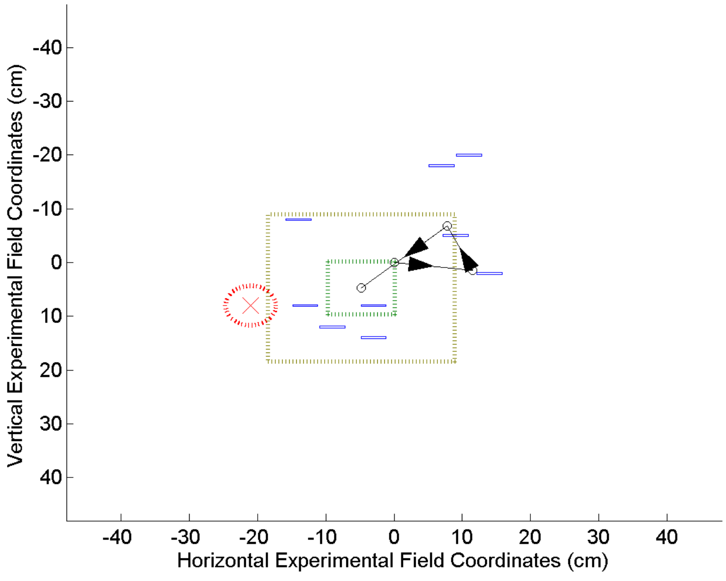
Fixation 1 of Trial 8

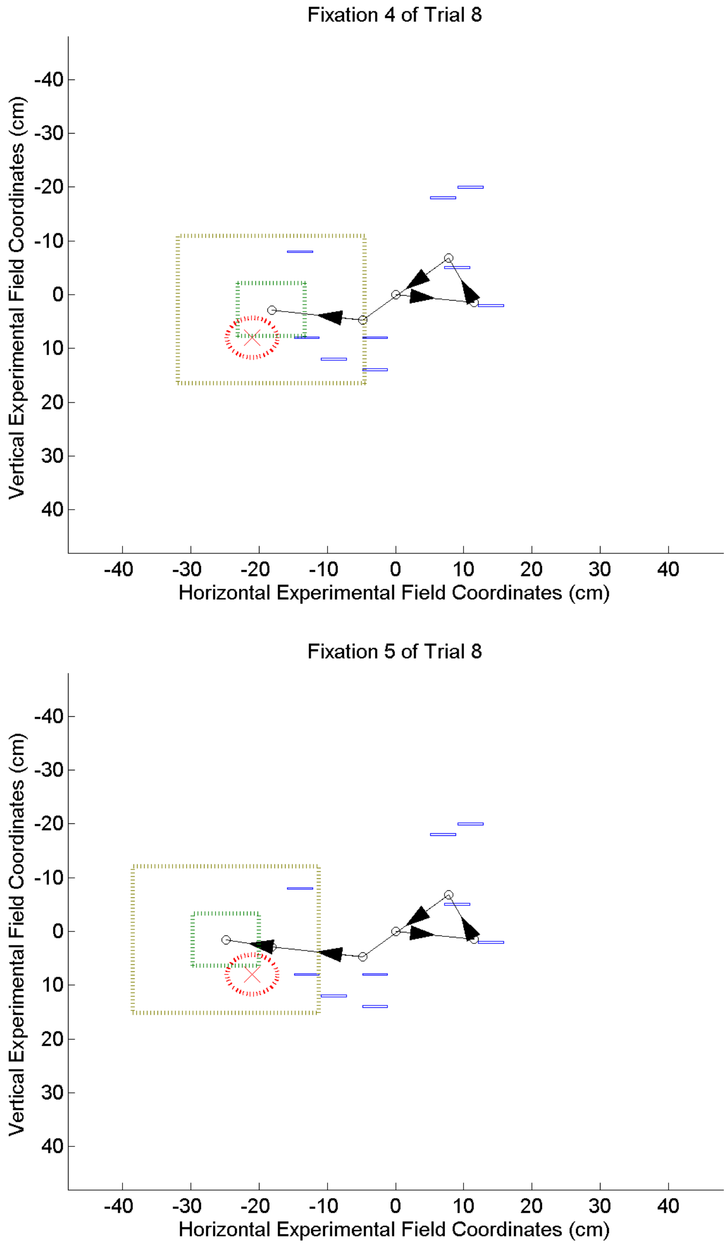


Fixation 2 of Trial 8



Fixation 3 of Trial 8





Fixation 4 of Trial 8

Fixation 5 of Trial 8

Figure 4.23 – Trial 8 Experiment 2. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.

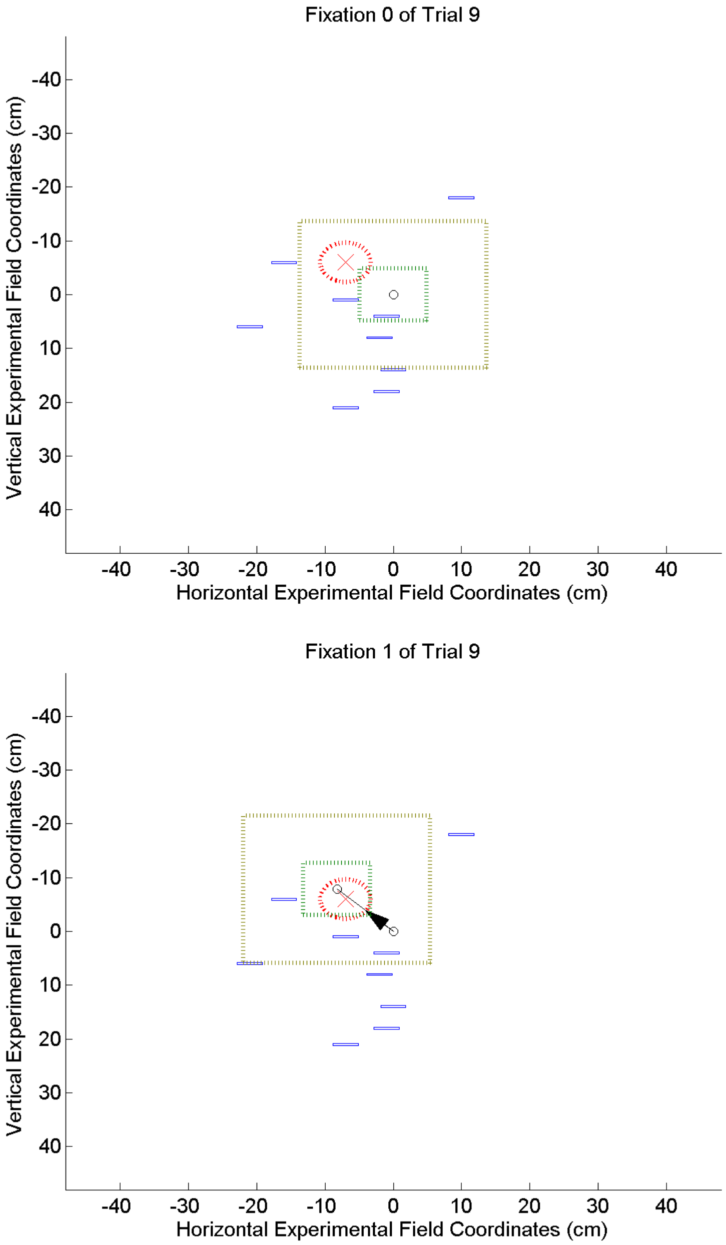
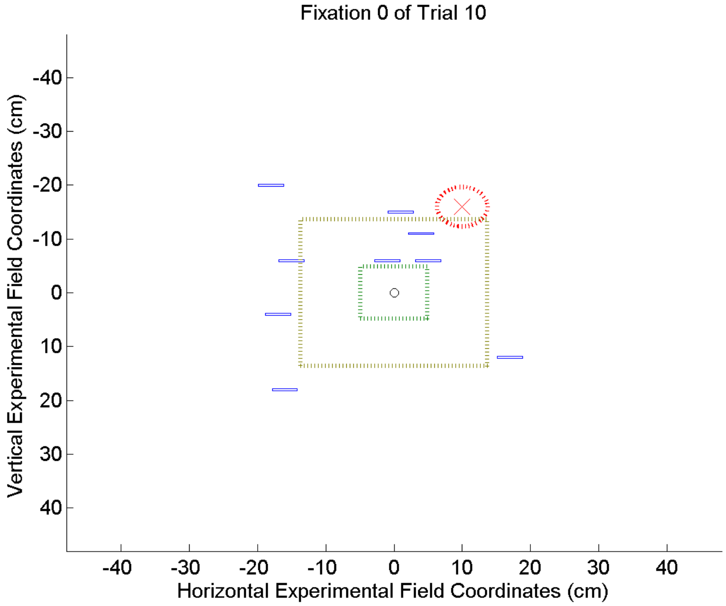
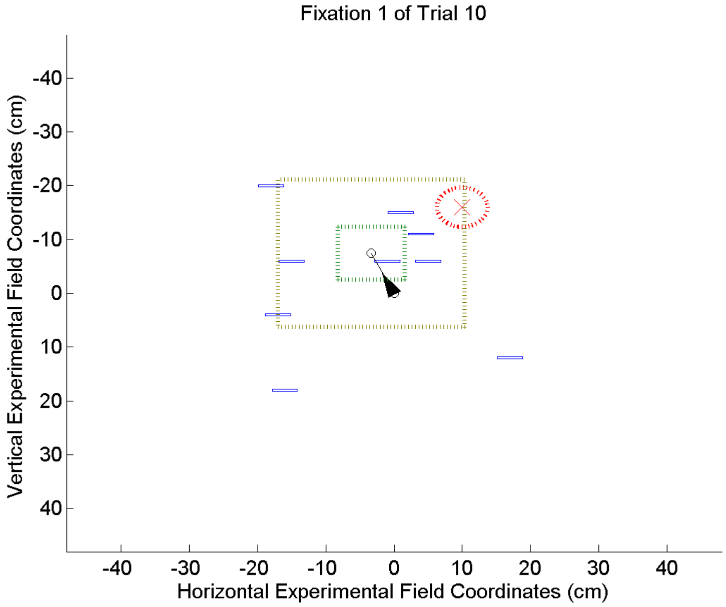


Figure 4.24 – Trial 9 Experiment 2. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.


Fixation 0 of Trial 10
Vertical Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40
Horizontal Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40


Fixation 1 of Trial 10
Vertical Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40
Horizontal Experimental Field Coordinates (cm)
-40
-30
-20
-10
0
10
20
30
40

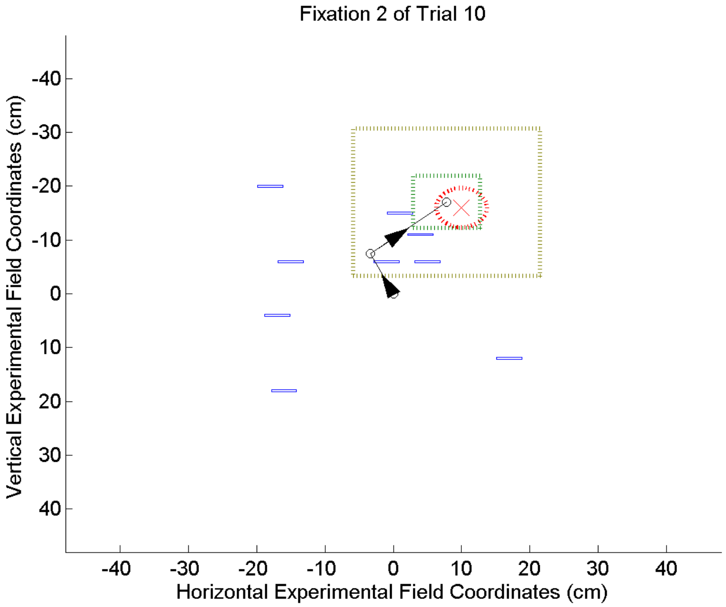


Figure 4.25 – Trial 10 Experiment 2. Distractors are represented by blue bars while the target is shown as a red dotted circle. The green dotted square displays the extent of the central recognition field, while the outer dotted square displays the extent of the peripheral field.

4.2.2 Virtual Experiment Results

Judd *et al.* evaluated the performance of several saliency algorithms on their data-set (including AIM) using an ROC metric to compare salience values over the entire image [38]. Such a metric was not directly applicable for evaluating this system, however, since the important aspect of performance is the actual fixation sequence rather than any particular distribution of salience within the image (plus, in the case of the peripheral saliency map model, the salience of any particular pixel is liable to differ depending on the current fixation). However, Judd *et al.* also came up with a method for evaluating human fixation data against the other human subjects by convolving a Gaussian kernel with the fixation points to produce a fixation-based "saliency map". The fixation-based saliency map was then thresholded by taking the top $n\%$ points to produce a binary mask (see Figure 4.26 for an example). Using the binary mask, the percent of fixation points from other subjects which fell within this mask were calculated. This approach was repeated for our experimental data, both for the explicit fixation sequence generated by peripheral saliency map model of this thesis and for a fixation sequence generated by taking the five top-most salient points from AIM. The results are displayed in Figure 4.27.

Not surprisingly, the results for the peripheral saliency map model are quite close to those of AIM itself; the peripheral priority measure in this experiment was derived from AIM and independent of top-down influences, after all. However, the combination of an implicit central bias combined with a localized measure of saliency based on the current fixation does appear to shift the results to be more consistent with human data. The first fixation of the peripheral saliency map model includes the strongest central bias (since it is impossible for the system to fixate beyond the bounds of the initial fixation's visual field), and so it is possible that the performance benefit of the peripheral model would de-

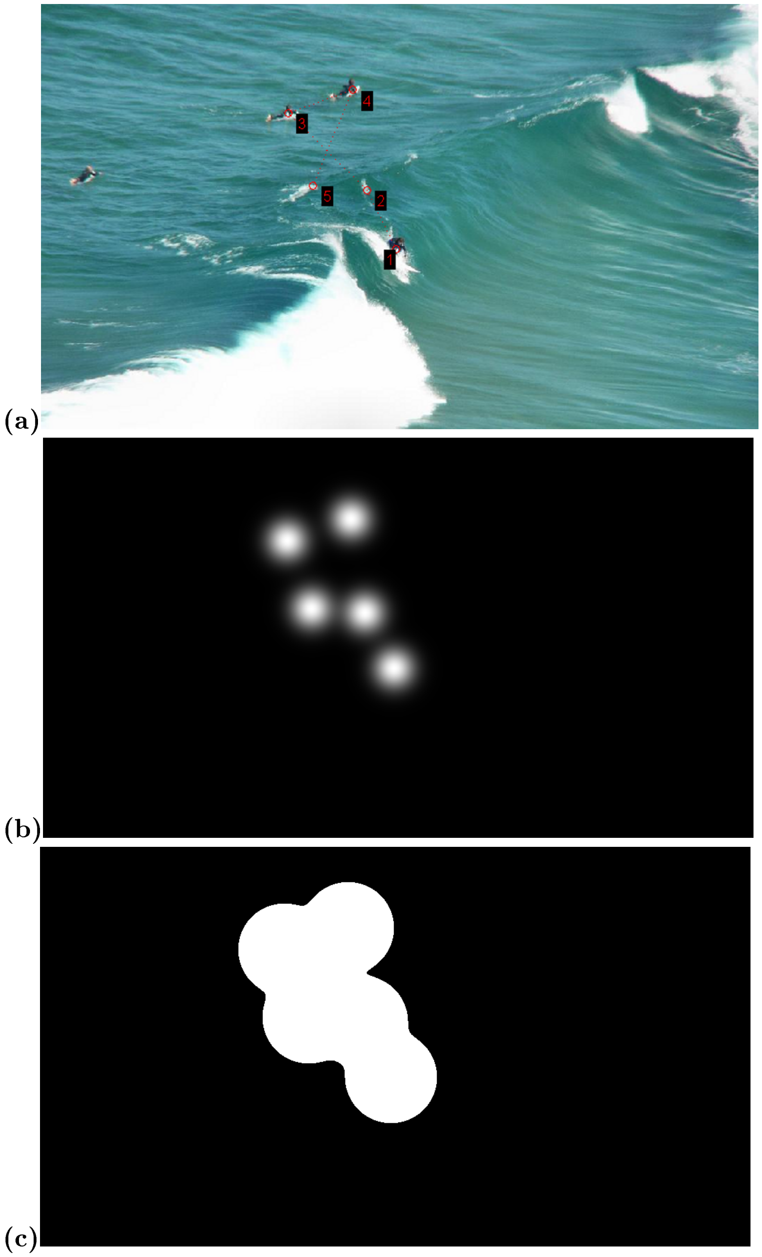


Figure 4.26 – Example of the generation of a fixation based saliency map and binary mask. **(a)** Fixation sequence produced by the peripheral saliency map model. **(b)** The fixation points shown in (a) convolved with a Gaussian filter to create a fixation-based saliency map (image values have been rescaled to aid in viewing). **(c)** A binary mask created by taking the top 10% of salient points from (b).

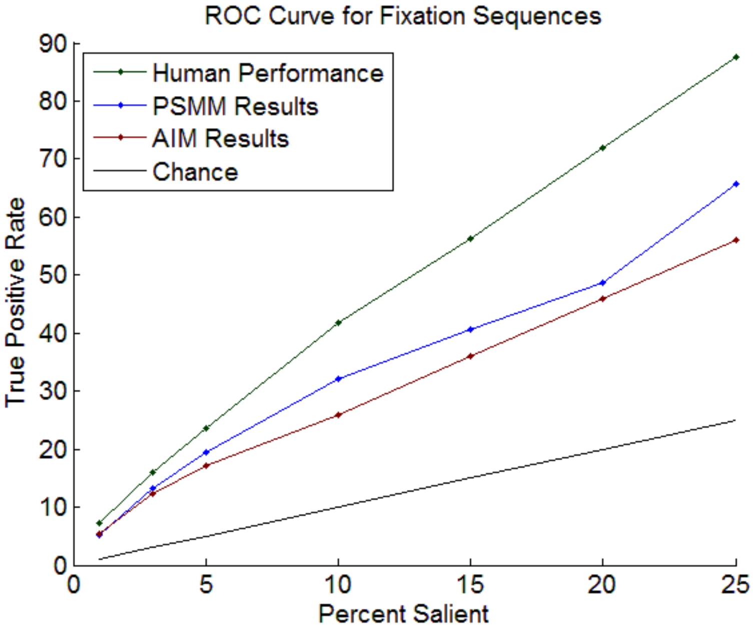


Figure 4.27 – The ROC curve of performances for human data, the Peripheral Saliency Map Model (PSMM) of this thesis, and AIM. Chance is also plotted for comparison.

crease with an increased number of fixations. Qualitative analysis of the fixation sequences suggests that this is not the case, but further study will be needed to be sure (see Figure 4.28 for a specific example in which the peripheral model produces a more human-like fixation sequence. In this image, AIM highlights all surfers in the water for fixation, whereas the peripheral saliency map model concentrates fixations instead only on the more central figures. Human observers show a similar pattern in their initial fixations, overwhelmingly concentrating their fixations on the central figures to the neglect of the leftmost surfer).

There are a number of additional aspects of this experiment which are worth noting. Neither system compared here has any semantic or contextual knowledge of the scene content; therefore scene elements which strongly draw human fixations (such as faces and text) were only fixated if they happened to also elicit strongly dissimilar filter responses from their surroundings. While this actually worked quite well for most images with text (see Figures 4.29 and 4.30) due to the tendency for lettering to produce unusually strong oriented edges, it was much less consistent for faces (see Figures 4.30, 4.31, and 4.32). In addition to the lack of contextual knowledge, all images were processed in greyscale; although colour information was not highly relevant for some of the images, there were others in which it would have aided in overcoming the systems' lack of worldly knowledge (see Figure 4.32).

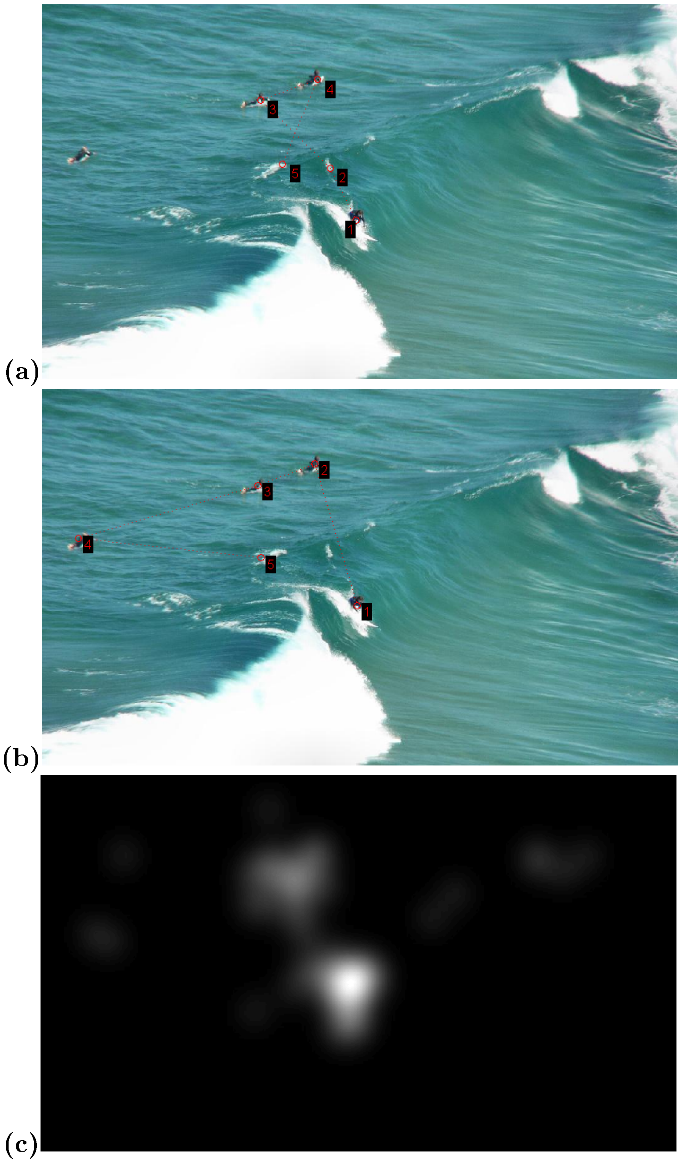


Figure 4.28 – Fixation sequences for **(a)** Peripheral saliency map model and **(b)** AIM. **(c)** Shows the human saliency map produced by Judd *et al.*, [38], which can be seen to neglect the leftward most surfer and therefore match the fixation sequence of (a) more closely.

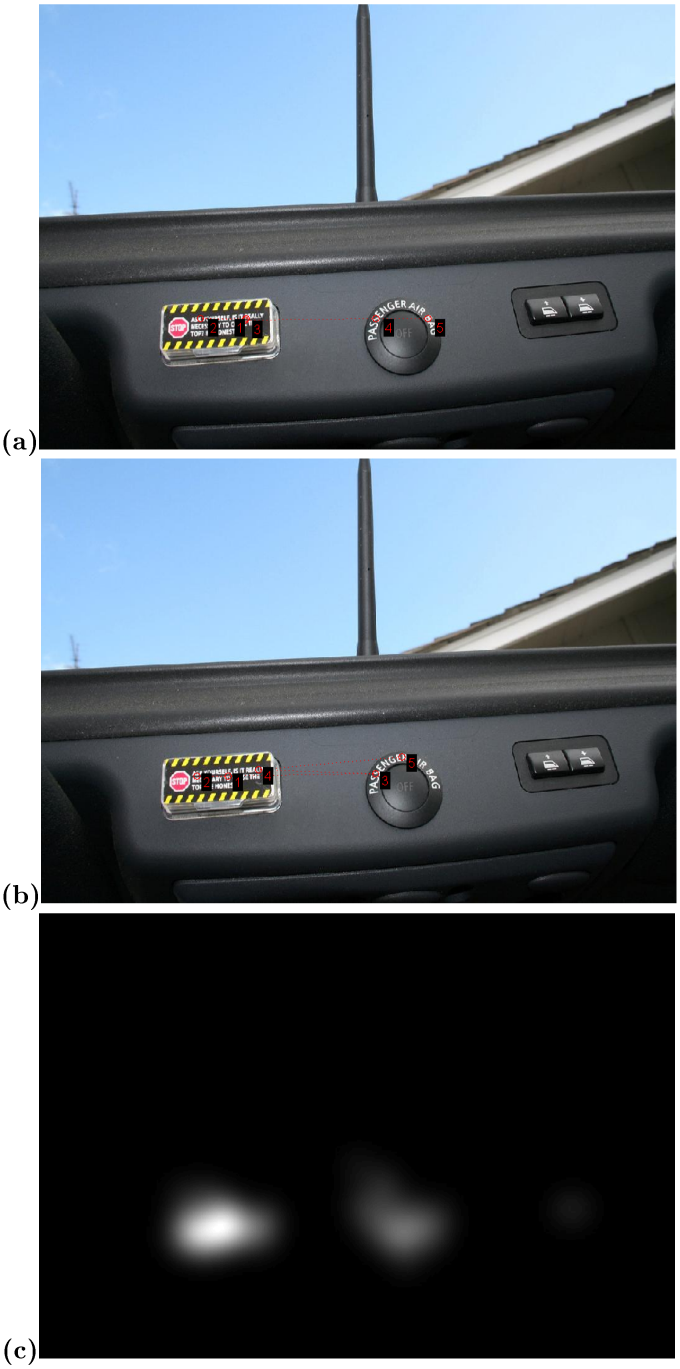


Figure 4.29 – Example of an image displaying text. Both the peripheral saliency map model, **(a)**, and AIM, **(b)**, show very similar fixation patterns which closely mirror the fixations of humans as displayed by the human fixation-based saliency map in **(c)**.

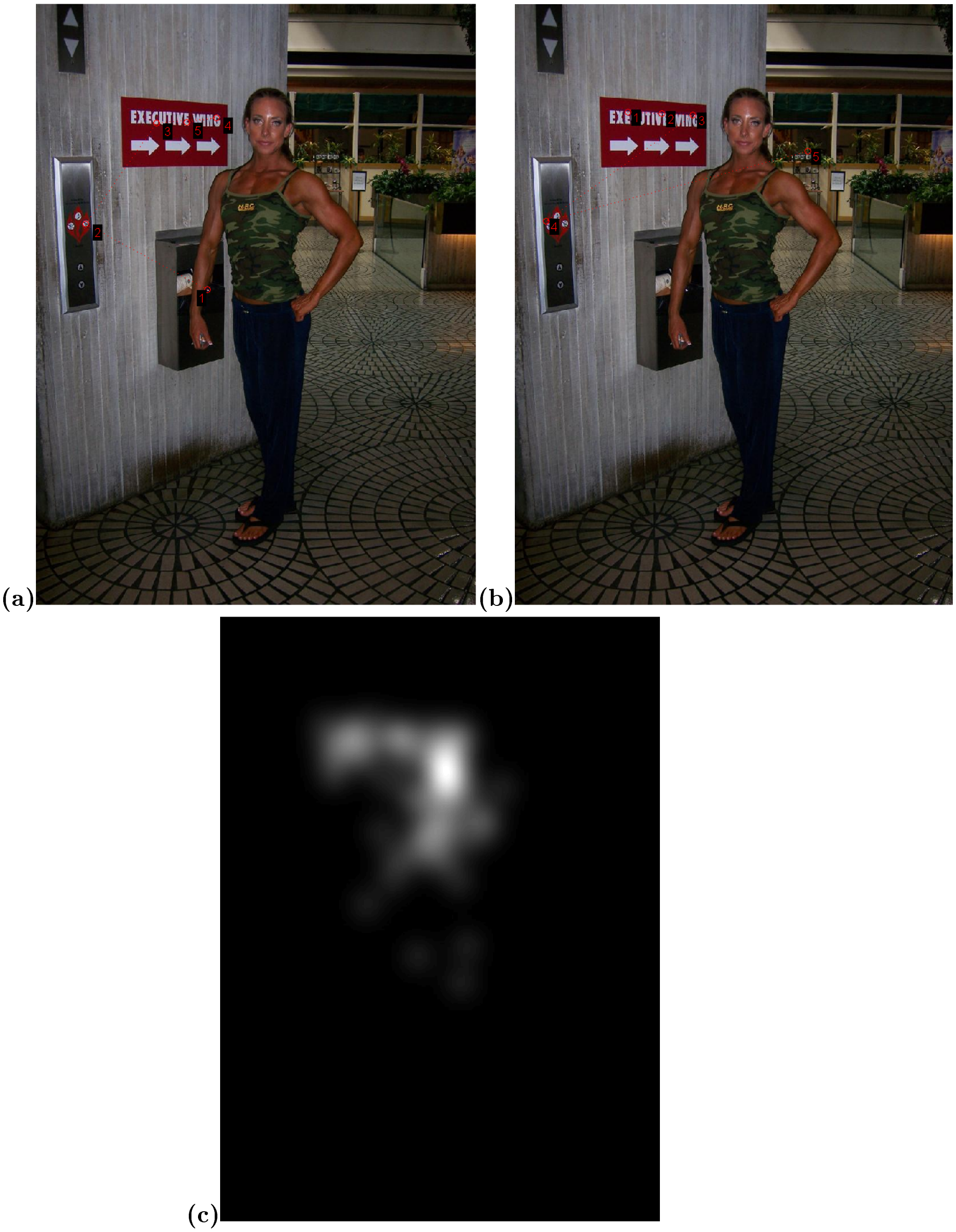


Figure 4.30 – Fixation sequences are shown for **(a)** the peripheral salience map model and **(b)** AIM. Both models fixate the elevator control panel and the sign text, but fail to fixate the woman's face. Image **(c)** displays the fixation-based saliency map for human fixations, which concentrated primarily on the woman's face and the sign text.

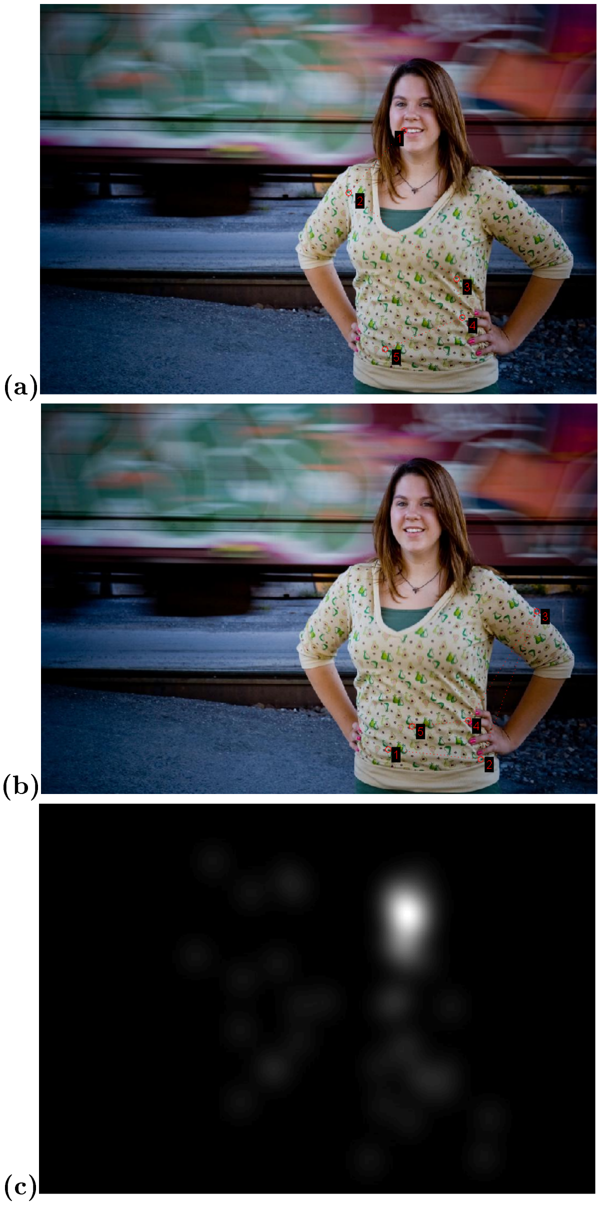


Figure 4.31 – Example in which the peripheral saliency map model, shown in **(a)**, fixated a face but AIM, shown in **(b)**, did not. Despite having a single fixation on the woman's face, however, the peripheral saliency map model, like AIM, still predominantly fixated on the woman's torso in contrast to the vast majority of human fixations represented by the fixation-based saliency map shown in **(c)**.

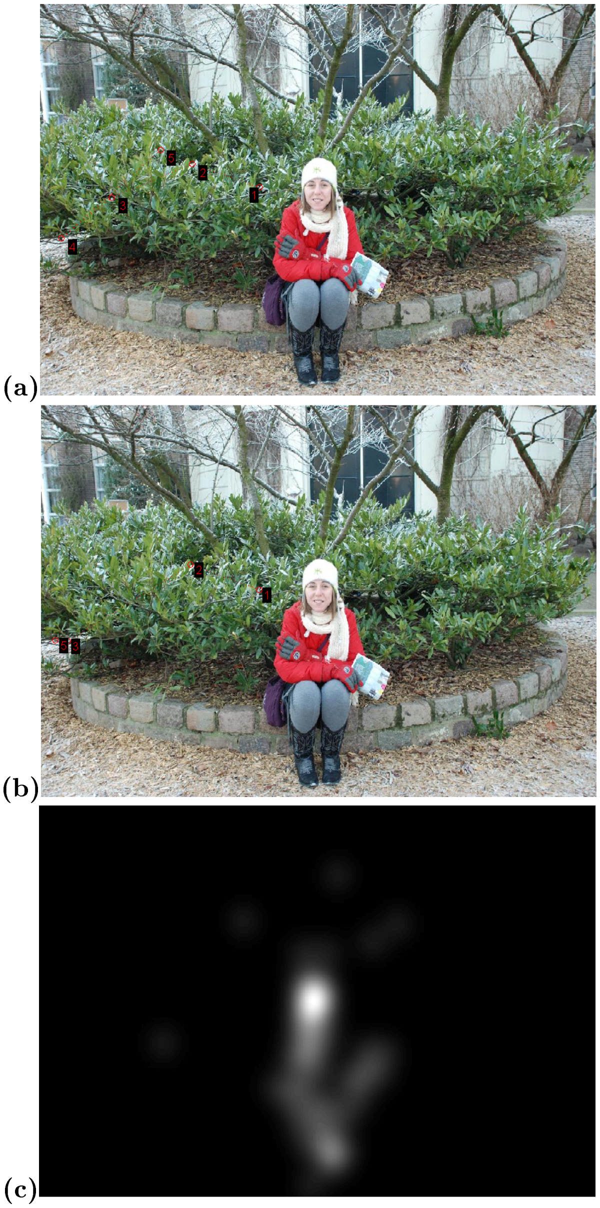


Figure 4.32 – Example in which both the peripheral saliency map model, **(a)**, and AIM, **(b)**, both failed in a highly similar manner to replicate human fixation patterns, represented by the fixation-based saliency map in **(c)**. The addition of colour channels would likely have improved performance in this image, as the woman's face, her clothing, and the paper she is holding would all likely increase in salience in comparison to the background.

4.3 Summary

The system developed for this thesis was evaluated over two experimental methodologies: a visual search task to find a form-singleton utilizing physical hardware, and an evaluation of fixation sequences for natural images utilizing a publicly available psychophysical database.

Evaluative criteria in the physical task were based on the capacity for the system to find the target and the average number of fixations required to do so. Despite an almost entirely data-driven approach, the system was able to find the target in all randomized trials with significantly fewer fixations than would be expected with a brute-force search of the experiment field. The lack of top-down knowledge does mean that experimental layouts could be designed on which the system would fail. Future avenues of research will therefore explore top-down control strategies which may be integrated into the system's operation.

Performance of the system in the second task was evaluated by comparing its fixation sequences to those of humans and to the fixation sequences generated by taking the top salient points from the AIM algorithm. An ROC curve was generated based on the methodology of Judd *et al.*, [38], which indicates an improvement in the performance of our system with respect to AIM in generating human-like fixation sequences. Given the favourable performance of AIM at predicting human fixations found by Borji *et al.*, [6], it appears that the peripheral saliency map model is able to compete with the current state of the art.

Chapter 5

Discussion and Conclusions

5.1 Summary

This thesis presents a framework for visual fixation control based on a peripheral priority signal, which implements a previously proposed but untested component to the overall Selective Tuning architecture proposed by Tsotsos as a solution to the Boundary Problem, [82]. Previous work on a biologically plausible active visual model by Zaharescu, [96], is extended in the TarzaNN neural network simulator architecture through novel implementations of a Peripheral Priority Map, a Fixational History Map which provides inhibition of return functionality, a History Biased Priority Map and saccade controller, and an environment control structure which provides faster and more extensible development for future application of active visual systems. Additionally, the AIM saliency map algorithm has been reimplemented in TarzaNN, providing an additional programming platform outside of MATLAB in which it can be utilized. As part of the development of the AIM implementation, a novel method for combining salience based on self-information from a heterogeneous distribution of filters was proposed using entropy to estimate the self-information. Overall the system implemented provides a computational substrate for the future exploration of many areas of control of active vision.

The system has been tested in both a physical environment on real hardware solving a visual search problem using an integrated Selective Tuning network, and in a virtual environment on a psychophysical data-set. The system was able to find a visual search target with significantly fewer fixations than a brute force method in all randomized trials, although in the current naive formulation there remain some specific problem formulations which it is unable to solve. Future avenues of research may extend the current system with top-down control strategies to rectify this problem.

When executed over the virtual data-set, the system was able to improve on AIM's performance in reproducing human-like fixation patterns. By finding an explicit fixation train over a sub-portion of the image, the peripheral saliency map model introduced an implicit bias against large translations between fixation points, thereby providing a more organic central bias term than is typically applied when attempting to predict eye fixation data. Given that AIM has been ranked among the top performing saliency detection models for predicting human fixations, [6], this indicates that the system ranks among the current state of the art.

5.2 Future Work

Much of the focus of this thesis has been in producing an active vision framework which can most easily be extended in the future to explore a variety of aspects of active vision. It is entirely possible that some of these research avenues have yet to even be envisioned; nevertheless there are a number of applications which I believe will be worthwhile exploring. The system developed in this thesis provides an excellent computational platform which may be used to test aspects of attention. This includes investigating more complex IOR mechanisms such as the possibility of multiple time courses depending on the level of scrutiny applied to a given region of the image, and short-term priming of pop-out mechanisms.

A well-developed peripheral priority mechanism for guiding active visual control may also provide novel applications in active search platforms. One specific example which seems promising would be to use the priority signal to inform the probability map of the SYT algorithm outlined in Section 1.2.3.3. In such a system a wide-field fish-eye lens could be used which would provide extensive coverage of the visual field while still limiting the higher level recognition task to the central field where distortion is minimal.

One additional avenue of research which would open up many more applications would be to extend the system to function in a dynamic rather than a static environment. Existing work exists on incorporating temporal components into the salience measure of AIM, [69], and much of the work discussed in Section 1.2.1.2 suggests that temporal onset cues are the strongest mechanism for peripheral attentional capture.

Bibliography

[1] Legacy product cameras, August 2012.

[2] Pan-tilt unit D-46 models, August 2012.

[3] R. A. Abrams and J. Jonides. Programming saccadic eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 14:428–443, 1988.

[4] John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1:333–356, 1988.

[5] Ruzena Bajcsy. Active perception vs. passive perception. In Linda Shapiro and Avi Kak, editors, *IEEE Workshop on Computer Vision: Representation and Control*, 1985.

[6] A. Borji, D. N. Sihite, and L. Itti. Salient object detection: A benchmark. In *Proc. European Conference on Computer Vision (ECCV), Florence, Italy*, Oct 2012.

[7] Kevin A. Briand, Abigail L. Larrison, and Anne B. Sereno. Inhibition of return in manual and saccadic response systems. *Perception and Psychophysics*, 62:1512–1524, 2000.

[8] Neil D. B. Bruce, Xun Shi, Evgueni Simine, and John K. Tsotsos. Visual representation in the determination of saliency. In *Eighth Canadian Conference on Computer and Robot Vision (CRV 2011)*, 2011.

[9] Neil D. B. Bruce and John K. Tsotsos. Saliency based on information maximization. *Advances in Neural Information Processing*, 18:155–162, 2006.

[10] Neil D. B. Bruce and John K. Tsotsos. An information theoretic model of saliency and visual search. In E. Rome L. Paletta, editor, *WAPCV*, pages 171–183, 2007.

[11] Neil D. B. Bruce and John K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9:1–24, 2009.

[12] James J. Clark and Nicola J. Ferrier. Modal control of an attentive vision system. In R. Bajcsy and S. Ullman, editors, *Proc. IEEE 2nd International Conference on Computer Vision*, pages 514–523, 1988.

[13] C. L. Colby, R. Gattass, C. R. Olson, and C. G. Gross. Topographical organization of cortical afferents to extrastriate visual area PO in the macaque: A dual tracer study. *Journal of Comparative Neurology*, 269:392–413, 1988.

[14] Laila Craighero, Arturo Carta, and Luciano Fadiga. Peripheral oculomotor palsy affects orienting of visuospatial attention. *Cognitive Neuroscience and Neuropsychology*, 12:3283–3286, 2001.

[15] Laila Craighero, Mauro Nascimben, and Luciano Fadiga. Eye position affects orienting of visuospatial attention. *Current Biology*, 14:331–333, 2004.

[16] T. J. Crawford and H. J. Muller. Spatial and temporal effects of spatial attention on human saccadic eye movements. *Vision Research*, 32:293–304, 1992.

[17] Christine A. Curcio and Kimberly A. Allen. Topography of ganglion cells in human retina. *The Journal of Comparative Neurology*, 300(1):5–25, 1990.

[18] P. M. Daniel and D. Witteridge. The representation of the visual field on the cerebral cortex in monkeys. *Journal of Physiology*, 159:203–221, 1961.

[19] DirectedPerception. *Computer Controlled Pan-Tilt Unit Model PTU-D46 User's Manual Version*, 2.14.0 edition.

[20] Jean-Rene Duhamel, Carol L. Colby, and Michael E. Goldberg. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, 255:90–92, 1992.

[21] F.L. Engel. Visual conspicuity, directed attention and retinal locus. *Vision Research*, 11:563–576, 1971.

[22] Patrizia Fattori, Sabrina Pitzalis, and Claudio Galletti. The cortical visual area v6 in macaque and human brains. *Journal of Physiology - Paris*, 103:88–97, 2009.

[23] Jillian H. Fecteau. Priming of pop-out depends upon the current goals of observers. *Journal of Vision*, 7(6):1:1–11, 2007.

[24] Jillian H. Fecteau and Douglas P. Munoz. Salience, relevance, and firing: a priority map for target selection. *Trends in Cognitive Sciences*, 10:382–390, 2006.

[25] Daniel J. Felleman and David C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991.

[26] Pedro Felzenswalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[27] David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America*, 4:2379–2394, 1987.

[28] Charles L. Folk and Roger Remington. Selectivity in distraction by irrelevant featural singletons: Evidence for two forms of attentional capture. *Journal of Experimental Psychology: Human Perception and Performance*, 24:847–858, 1998.

[29] Ziad M. Hafed and James J. Clark. Microsaccades as an overt measure of covert attention shifts. *Vision Research*, 42:2533–2545, 2002.

[30] Anne P. Hillstrom and Steven Yantis. Visual motion and attentional capture. *Perception and Psychophysics*, 55:399–411, 1994.

[31] George Houghton and Steven P. Tipper. Inhibitory mechanisms of neural and cognitive control: Applications to selective attention and sequential action. *Brain and Cognition*, 30:20–43, 1996.

[32] A. R. Hunt and A. Kingstone. Covert and overt voluntary attention: linked or independent? *Cognitive Brain Research*, 18:102–105, 2003.

[33] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000.

[34] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1254–1259, 1998.

[35] John Jonides. *Attention and performance IX.* Lawrence Erlbaum Associates Inc., 1981.

[36] John Jonides and Steven Yantis. Uniqueness of abrupt visual onset in capturing attention. *Perception and Psychophysics*, 43:346–354, 1988.

[37] Tilke Judd, Krista Ehinger, Fredo Durand, and Antonio Torralba. Learning to predict where humans look.

[38] Tilke Judd, Krista Ehinger, Fredo Durand, and Antonio Torralba. Learning to predict where humans look. In *International Conference on Computer Vision*, 2009.

[39] Raymond M. Klein. *Does oculomotor readiness mediate cognitive control of visual attention?* Springer-Verlag, 1980.

[40] Raymond M. Klein. Inhibitory tagging system facilitates visual search. *Nature*, 334:430–431, 1988.

[41] Raymond M. Klein. Inhibition of return. *Trends in Cognitive Sciences*, 4:138–147, 2000.

[42] Raymond M. Klein and Amanda Pontefract. *Does Oculomotor Readiness Mediate Cognitive Control of Visual Attention? Revisited!* The International Association for the Study of Attention and Performance, 1994.

[43] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[44] P.D. Kovesi. Matlab and octave functions for computer vision and image processing.

[45] E. L. Lawler and D. E. Wood. Branch-and-bound methods: a survey. *Operations Research*, 14:699–719, 1966.

[46] Juan Lupiáñez, Raymond M. Klein, and Paolo Bartolomeo. Inhibition of return: Twenty years after. *Cognitive Neuropsychology*, 23:1003–1014, 2006.

[47] Juan Lupiáñez and Bruce Milliken. Inhibition of return and the attentional set for integrating versus differentiating information. *The Journal of General Psychology*, 126:392–418, 1999.

[48] Juan Lupiáñez, Emilio G. Milán, Francisco J. Tornay, Eduardo Madrid, and Pío Tudela. Does ior occur in discrimination tasks? yes, it does, but later. *Attention, Perception, and Psychophysics*, 59:1241–1254, 1997.

[49] Vera Maljkovic and Ken Nakayama. Priming of pop-out: I. role of features. *Memory and Cognition*, 22:657–672, 1994.

[50] Vera Maljkovic and Ken Nakayama. Priming of pop-out: Ii. the role of position. *Perception and Psychophysics*, 58:977–991, 1996.

[51] Vera Maljkovic and Ken Nakayama. Priming of popout: Iii. a short-term implicit memory system beneficial for rapid target selection. *Visual Cognition*, 7:571–595, 2000.

[52] W. Pieter Medendorp, Herbert C. Goltz, Tutis Vilis, and J. Douglas Crawford. Gaze-centered updating of visual space in human parietal cortex. *The Journal of Neuroscience*, 23:6214–6209, 2003.

[53] B. C. Motter and E. J. Belky. The guidance of eye movements during active visual search. *Vision Research*, 38:1805–1815, 1998.

[54] Vidhya Navalpakkam and Laurent Itti. Modeling the influence of task on attention. *Vision Research*, 45:205–231, 2005.

[55] Vidhya Navalpakkam and Laurent Itti. Search goal tunes visual features optimally. *Neuron*, 53:605–617, 2007.

[56] Hirokazu Ogawa, Yuji Takeda, and Akihiro Yagi. Inhibitory tagging on randomly moving objects. *Psychological Science*, 13:125–129, 2002.

[57] Evan M. Palmer, Todd S. Horowitz, Antonio Torralba, and Jeremy M. Wolfe. What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, 37:58–71, 2011.

[58] Stephen E. Palmer. *Vision Science: Photons to Phenomenology*. The MIT Press, 1999.

[59] A. J. Parker and M. J. Hawken. Two-dimensional spatial structure of receptive fields in monkey striate cortex. *Journal of the Optical Society of America*, 5:598–605, 1988.

[60] M. I. Posner and Y. Cohen. *Components of Visual Orienting. In Attention and Performance*, volume Vol. X. Erlbaum, 1984.

[61] Michael I. Posner, Robert D. Rafal, Lisa S. Choate, and Jonathan Vaughan. Inhibition of return: Neural basis and function. *Cognitive Neuropsychology*, 2:211–228, 1985.

[62] RD Rafal, PA Calabresi, CW Brennan, and TK Sciolto. Saccade preparation inhibits reorienting to recently attended locations. *Journal of Experimental Psychology: Human Perception and Performance*, 15:673–85, 1989.

[63] Patricia A. Reuter-Lorenz and Robert Fendrich. Oculomotor readiness and covert orienting: Differences between central and peripheral precues. *Perception and Psychophysics*, 52:336–344, 1992.

[64] Giacomo Rizzolatti, Lucia Riggio, Isabella Dascola, and Carlo Umilta. Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25:31–40, 1987.

[65] Albert L. Rothenstein. *Beyond the Limits of Feed-Forward Processing: Visual Feature Binding and Object Recognition*. PhD thesis, York University, 2011.

[66] Albert L. Rothenstein, Andrei Zaharescu, and John K. Tsotsos. Tarzann: A general purpose neural network simulator for visual attention modeling. In *WAPCV*, 2004.

[67] Albert L. Rothenstein, Andrei Zaharescu, and John K. Tsotsos. Tarzann: General purpose neural network simulator, 2012.

[68] B. M. Sheliga, L. Riggio, and G. Rizzolatti. Spatial attention and eye movements. *Experimental Brain Research*, 105:261–275, 1995.

[69] Xun Shi, Neil Bruce, and John Tsotsos. Biologically motivated local contextual modulation improves low-level visual feature representations. In Aurélio Campilho and Mohamed Kamel, editors, *Image Analysis and Recognition*, volume 7324 of *Lecture Notes in Computer Science*, pages 79–88. Springer Berlin / Heidelberg, 2012.

[70] Ksenia Shubina and John K. Tsotsos. Visual search for an object in a 3D environment using a mobile robot. Technical report, York University, 2008.

[71] Hans Strasburger and Ingo Rentschler. Contrast-dependent dissociation of visual recognition and detection field. *European Journal of Neuroscience*, 8:1787–1791, 1996.

[72] Hans Strasburger, Ingo Rentschler, and Martin Jüttner. Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11:1–82, 2011.

[73] Kyeong-Jin Tark and Clayton E Curtis. Persistent neural activity in the human frontal cortex when maintaining space that is off the map. *Nature Neuroscience*, 12:1463–1468, 2009.

[74] Benjamin W. Tatler, Mary M. Hayhoe, Michael F. Land, and Dana H. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11(5):1–23, 2011.

[75] Tracy L. Taylor and Raymond M. Klein. Visual and motor effects in inhibition of return. *Journal of Experimental Psychology: Human Perception and Performance*, 26:1639–56, 2000.

[76] Kathleen M. Terry, Leslie A. Valdes, and W. Trammell Neill. Does "inhibition of return" occur in discrimination tasks? *Attention, Perception, and Psychophysics*, 55:279–286, 1994.

[77] Jan Theeuwes. Perceptual selectivity for color and form. *Perception and Psychophysics*, 51:599–606, 1992.

[78] Jan Theeuwes and Frank L. Kooi. Parallel search for a conjunction of contrast polarity and shape. *Vision Research*, 34:3013–3016, 1994.

[79] Thomas L. Thornton and David L. Gilden. Parallel and serial processes in visual search. *Psychological Review*, 114:71–103, 2007.

[80] Anne Treisman and Garry Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[81] John K. Tsotsos. Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13:423–469, 1990.

[82] John K. Tsotsos. *A Computational Perspective on Visual Attention.* MIT Press, 2011.

[83] John K. Tsotsos, Sean M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507 – 545, 1995. Special Volume on Computer Vision.

[84] Stefan van der Stigchel and Jan Theeuwes. The relationship between covert and overt attention in endogenous cuing. *Perception and Psychophysics*, 69:719–731, 2007.

[85] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.

[86] Hermann von Helmholtz and translated by James P.C. Southall. *Helmholtz's treatise on physiological optics.* Dover, translated from the 3rd [g]erman edition, 1962.

[87] Gooitzen S. Wal and Peter J. Burt. A vlsi pyramid chip for multiresolution image analysis. *International Journal of Computer Vision*, 8:177–189, 1992. 10.1007/BF00055150.

[88] Brian A. Wandell. *Useful quantities in vision science (from introductory pages of Foundations of Vision).* Sinauer Associates, 1995.

[89] Zhiguo Wang and Raymond M. Klein. Searching for inhibition of return in visual search: A review. *Vision Research*, 50:220–228, 2010.

[90] Hugh R. Wilson. Non-fourier cortical processes in texture, form, and motion perception. In Philip S. Ulinski, Edward G. Jones, Alan Peters, Edward G. Jones, and Alan Peters, editors, *Cerebral Cortex*, volume 13 of *Cerebral Cortex*, pages 445–477. Springer US, 1999.

[91] Jeremy M. Wolfe. *Visual Search*. Psychology Press, 1998.

[92] Jeremy M. Wolfe, Evan M. Palmer, and Todd S. Horowitz. Reaction time distributions constrain models of visual search. *Vision Research*, 50:1304–1311, 2010.

[93] Steven Yantis. *Visual Search*. Psychology Press, 1998.

[94] Steven Yantis and John Jonides. Abrupt visual onsets and selective attention: Evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 10:601–621, 1984.

[95] Yiming Ye and John K. Tsotsos. Sensor planning for 3D object search. *Computer Vision and Image Understanding*, 73:145–168, 1999.

[96] Andrei Zaharescu. Towards a biologically plausible active search model. Master's thesis, York University, 2004.

[97] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7:32):1–20, 2008.