FIXATION SELECTION FOR CATEGORICAL TARGET SEARCHES IN REAL WORLD SCENES

By

NICHOLAS KLEENE

A dissertation submitted to the

School of Graduate Studies

Rutgers, the State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate program in Psychology

Written under the direction of

Melchi Michel

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2019

ABSTRACT OF THE DISSERTATION

Fixation Selection for Categorical Target Searches in Real World Scenes

by **NICHOLAS KLEENE**

Dissertation Director:

Melchi Michel

Computational models have seen widespread success in predicting fixation locations for visual search tasks that use artificial stimuli, such as Gabors in 1/f noise (Najemnik & Geisler, 2005), but comparatively little in predicting fixation locations for visual search tasks with natural stimuli. Critically, previous approaches have not accounted for the effects of our foveated visual system nor implemented decision rules to actually select a sequence of fixations (Torralba, Oliva, Castelhano, & Henderson, 2006; Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009). Here we present a Bayesian model of fixation selection in visual search tasks using natural images. The model used two known sources of information to select fixations: scene context and features that look similar to the target (termed target-relevant features here). Scene context functioned as a prior over possible target locations, while target-relevant features acted as the likelihood function (as in Ehinger et al., 2009). Scene context was measured using GIST (Torralba et al., 2006) and target-relevant features were measured

using Histograms of Oriented Gradients (Dalal & Triggs, 2005). We represented scene context with a mixture of Gaussians and target-relevant features with a multivariate Gaussian distribution. The model selected new fixations using either a maximum a posteriori (MAP) or entropy limit minimization (ELM) rule. To compare the fixations selected by the models we tested human observers on a pedestrian search task in natural images. Prior to the search task, a visibility map was measured using data from human observers in detection task. The visibility map was then used to degrade the target-relevant feature information in our model simulations, representing the effects of foveation. We found evidence that human observers do use scene context and target-relevant features as sources of information to guide their fixations in natural scenes. Additionally, fixations selected by human observers were more consistent with the ELM decision rule than the MAP decision rule. We close by noting the limitations of the models and discuss potential extensions.

Acknowledgements

I would like to thank my adviser, Melchi Michel, and the members of my committee: Eileen Kowler, Manish Singh and Ahmed Elgammal. I would also like to thank Yelda Semizer and the other members of the computational vision lab at Rutgers University. Finally, thank you to all who participated in this study.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Visual search, the act of making eye-movements (saccades) to find a target in a visual environment is a ubiquitous and essential part of everyday life. This is in large part because visual search subserves action, and is therefore a critical subtask in achieving larger behavioral goals (Land, 2009; Hayhoe & Ballard, 2005). Everyday activities that involve visual search include reading, looking for your car in a parking lot, deciding where to place your foot when walking on uneven terrain, and guiding your arm/hand when reaching for your mug to take a sip of coffee. To demonstrate the frequency with which we conduct visual search, consider the simple task of making a peanut butter and jelly sandwich (Land & Hayhoe, 2001). Doing this requires the following search-action pairs:

1. Locate the bread and guide a slice to the plate

2. Locate the peanut butter and collect some on a knife
3. Re-localize the bread to smear the peanut butter onto it
4. Locate the jelly and collect some on a knife
5. Re-localize the bread to smear the jelly onto it
6. Find and place the second slice of bread on top to complete the sandwich

Given that a relatively simple task like making a peanut butter and jelly sandwich required six visual search subtasks, the number of searches required for more complex behaviors such as driving a car will be extremely large. Visual search tasks can be either overt (with eye-movements) or covert (without eye-movements). Interestingly, covert search tends to be part of overt search tasks because humans use information from non-fixated locations to guide future fixations. On the other hand, tasks that exclusively require covert search are relatively rare in natural settings. Therefore, this study is restricted to overt visual search tasks since they are more general than exclusively covert search tasks.

Due to the prevalence of visual search in everyday life, completing searches accurately and quickly should be of critical importance. In the context of visual search, accuracy is often measured through detection ('the target is present/absent') or localization ('the target is at a specific location'). Two measures of search speed are the number of fixations required for a search and the amount of time spent on a search (often termed 'response time'). Given

similar task conditions, these two measures of speed will often be correlated. The dual goals of maximizing accuracy and minimizing search time can be in conflict, resulting in an antagonistic relationship referred to as the speed-accuracy trade-off. If the visual system truly is trying to optimize for both these quantities, then there should be features of human fixation patterns that aid the optimization. Much of what we know about the features of fixations patterns that might help maximize accuracy and/or minimize search time comes from laboratory experiments using artificial search elements, such as an array of letters or colored shapes presented against a grey background. One particular feature of human scan paths is the tendency to fixate between search elements, rather than foveating (fixating directly on) them (Coeffe & O'Regan, 1987; Kaufman & Richards, 1969; He & Kowler, 1989; Findlay, 1997). Fixating between objects allows the observer to gain information about multiple potential targets (Verghese, 2012), rather than just one, which can reduce the number of eye-movements required to find the target (Najemnik & Geisler, 2009) by allowing the observer to rule out multiple search elements at once. Another persistent feature of human fixation patterns is the tendency to avoid fixating the same locations twice (Klein, 1988, 2000; Tipper, Weaver, Jerreat, & Burak, 1994). The benefit here should be obvious: if a search element has already been classified as a distractor, then fixating it again would be completely uninformative. Therefore, it seems that human fixation patterns exhibit features that could be used to maximize accuracy and search speed.

Although the immediate goal of many visual search studies is to characterize search in artificial displays, as a field we would ultimately like to be able to

draw conclusions regarding search in the real world. Ideally, we would like to be able to predict how humans will search given a particular target and scene. Computational models provide a promising avenue of research in this respect since they produce quantifiable, and therefore testable, predictions about how humans search. Additionally, computational models can be used to evaluate what sources of information humans use in visual search. If the behavior of a model that uses a particular information source does not match human behavior then it is unlikely that humans use that source of information. A related concern is how humans actually make decisions once they have acquired the relevant information. In computational models this takes the form of a decision rule, a theoretical computation that synthesizes the relevant information to produce a behavior. Known constraints on visual processing can also be incorporated into computational models. Comparing the performance of a constrained model with an unconstrained model can tell us how the constraint impacts search behavior. Unfortunately, as we move from artificial stimuli to more naturalistic stimuli it becomes more difficult to measure or estimate the various parameters of computational models. This makes it more difficult to apply computational models to natural images than artificial stimuli. Deriving approximations that will work on natural images remains a challenge for developing computational models of visual search.

Computational models can either be predictive or normative. Predictive models of visual search seek to predict how observers *will* search when given a particular target and scene. Importantly, it is not necessary for a predictive model to perform the same computations as humans, all these models

need to do is match human behavior. On the contrary, normative models seek to predict how observers *should* search when given a particular target or scene. The distinction here is that the primary goal for a normative model of is to replicate the same computations that humans perform as closely as possible. The primary question of this study is how humans should search, therefore I will define a 'good' model as one that has a normative base and incorporates as many known biological constraints and sources of information as possible while still producing a close match to human behavior. This means that computational models don't only need to match human behavior, they also need to do so with the same constraints that limit humans and the same sources of information that humans use. Models that can match human performance but don't include important constraints and sources of information will not be considered 'good' models of visual search because these models are necessarily performing different computations than human observers, and are therefore unable to make predictions about how humans *should* search. For our purposes a 'good' computational model is one that provides normative predictions about how human observers search, given a defined task and stimulus. Predictive models that only seek to predict human behavior without incorporating known biological constraints or sources of information would not be considered 'good' computational models because they are simply matching observer behavior, not modeling the actual computations the observer is performing. While normative computational models may not perform the same exact computations as human observers they are far more likely to perform similar computations than predictive models.

In this paper I will first present biological constraints and sources of information that influence search performance. Next, I will critically examine several computational models of visual search that incorporate some of the biological constraints and sources of information discussed. I will then present a Bayesian model of fixation selection that combines some of the strengths of these computational models. Model predictions will be compared to eye-movement data collected in a behavioral experiment where observers searched for a natural target in real world images. Finally, I will discuss the implications of these results.

1.1 Biological Constraints on Visual Processing

When attempting to model the behavior of an organism, an important first step is identifying the inherent limitations of the organism. If the organism's perceptual or behavioral systems are constrained in some way so that certain tasks prove impossible, or exceedingly difficult, then a normative model of that organism should not be able to complete those tasks with ease. For the task of visual search the most important biological constraints to consider are the degraded quality of information in the periphery due to our foveated visual system and limitations on cognitive resources such as memory and attention.

1.1.1 Degraded Peripheral Sensitivity

It is well known that the quality of visual information in the periphery is greatly reduced when compared with the fovea. Although photoreceptors tile the entire retina, the number of ganglion cells they project to changes dramatically as a function of eccentricity. Photoreceptors in the fovea have an approximately 1 : 1 connection with foveal ganglion cells, but at greater eccentricities larger numbers of photoreceptors project onto the same peripheral ganglion cells (Fischer, 1973). As a result, the receptive field size of ganglion cells in the periphery is much larger than in the fovea. Smaller receptive field sizes are associated with greater acuity and spatial resolution, so the quality of information transmitted by foveal neurons is greater than neurons in the periphery. For visual search this means that the representation of search elements displayed in the periphery will necessarily have more uncertainty associated with them, which can reduce accuracy and increase search time (Geisler & Chou, 1995; Scialfa & Joffe, 1998; Carrasco, Evert, Chang, & Katz, 1995). The eccentricity effect holds true regardless of whether search time is quantified in terms of the number of fixations or response time.

Foveation also gives rise to the phenomenon known as crowding, where recognition of search elements in the periphery is impaired by other nearby search elements. Crowding has also been shown to have detrimental effects on search accuracy and can increase search times (Vlaskamp, Over, & Hooge, 2005; Vlaskamp & Hooge, 2006; Rosenholtz, Li, & Nakano, 2007). Since human performance is impaired as search elements become more and more

peripheral, so should the performance of any computational model for visual search. One method for controlling for the fall-off in peripheral acuity is by constraining models with visibility maps that quantify this fall-off (e.g., Najemnik & Geisler, 2005; Zelinsky, 2008; Zelinsky, Adeli, Peng, & Samaras, 2013). For well controlled stimuli (e.g. a Gabor in noise), visibility maps can be measured for individual participants using psychophysics and signal detection theory (Najemnik & Geisler, 2005). This helps control for any differences in target visibility across observers. Unfortunately, it is difficult to quantify the visibility maps for natural scenes (e.g. a teddy bear in a bedroom). In this case a common visibility map is often used across observers (Zelinsky, 2008; Zelinsky et al., 2013). Common visibility maps have the benefit of requiring much less data collection, but come with the drawback of having worse predictive power.

1.1.2 Limited Memory

There is little debate that visual memory (Miller, 1956; Brady, Konkle, & Alvarez, 2011) is a limited resource, but it's impact on visual search is still an open question. Visual memory could theoretically be used for search in a number of ways, such as storing target-relevant information (e.g. features associated with the target) or tracking previously fixated locations. Most studies on this issue have concluded that memory is either not involved in search (Horowitz & Wolfe, 1998) or that it is not a limiting factor on search performance (Najemnik & Geisler, 2005). Another study found that memory

was only minimally required for storing target-relevant information (Woodman & Chun, 2006). However, memory may also be required to store previously fixated locations to avoid re-fixation. The fact that humans tend to avoid fixating the same locations hints at the idea that memory may play a more important role in visual search than previously thought. Additionally, memory has been shown to be an important component of some saccadic targeting models (e.g., Epelboim & Suppes, 2001; Aivar, Hayhoe, Chizk, & Mruczek, 2005), showing that memory can be used to guide saccades. Unfortunately, due to the popular belief that memory is not a constraining factor in visual search (e.g., Najemnik & Geisler, 2005; Horowitz & Wolfe, 1998) and lack of agreement over *how* it should impact search there have been few computational models explicitly developed for visual search including the influence of limited visual memory.

1.1.3 Limited Attention

Limited attention has frequently been hypothesized to be a constraining factor for visual search. Attention can either be overt (attention that is directed using eye-movements) or covert (attention that is directed without using eye-movements). Overt attention is necessarily limited since directing eye-movements to different scene locations is a sequential process. Directing overt attention to a location brings that location into the fovea, which improves the quality of information we receive there. This means that overt attention is a sequential process because of our foveated visual system. Therefore, the effects

of limited covert attention are more relevant to the current study than those of overt attention.

There are two particular types of covert attention that are relevant for visual search. The first is covert spatial attention, which consists of directing attention towards a particular spatial location without moving one's eyes. In visual search tasks spatial attention is used for selecting spatial locations to direct overt attention (eye-movements) towards. Laboratory tasks using artificial stimuli have shown that attention can help ameliorate the detrimental effects of our foveated visual system. For example, applying covert attention to peripheral search elements actually increases the spatial resolution of their representation, thereby attenuating the eccentricity effect (Carrasco & Yeshurun, 1998). The second relevant type of attention is feature-based attention. In feature-based attention, attention is covertly directed towards specific features (e.g. color, orientation), regardless of their physical location. The role of feature-based attention in visual search is relatively straightforward. If the observer knows the target-relevant features then they can selectively direct attention towards search elements that have similar features. For example, if the observer is searching for a red baseball cap then feature-based attention would be directed towards red objects. The utility of feature-based attention has borne out in visual search as previous research has found that feature-based attention can guide eye-movements towards and prioritize processing of search elements with target-relevant features (Moore & Egeth, 1998; Shih & Sperling, 1996). Therefore, restricting attention to search elements with target-relevant features can lead to reduced search times and improved accuracy.

Although attention has been shown to play a role in visual search, including limited attention as a constraint for computational models has so far been difficult. A major hurdle is that there is no agreement on how attention should influence search performance. In particular, two hypotheses about the mechanism through which attention acts are information selection and information enhancement (for a more thorough review of the different attentional mechanism hypotheses, see Carrasco, 2011). These hypotheses are not actually mutually exclusive, meaning that attention may act through both of these mechanisms. This lack of agreement regarding the attentional mechanism makes it difficult to generate quantifiable predictions for how limited attention constrains search behavior. Additionally, there is some debate regarding whether effects on search previously attributed to limited attention are actually caused by limited attention. Instead, set-size (the number of search elements in a display) effects attributed to limitations on attention may actually result from the stochastic nature of visual processing (Palmer, Verghese, & Pavel, 2000). Due to our inability as a field to agree on a mechanism through which attention acts and the difficulty disentangling the effects of limited attention and variability in neural firing rates, developing computational models that include limited attention has not been especially fruitful.

1.2 Sources of Information

There are several sources of information that humans seem to use to reduce uncertainty about target locations and guide eye-movements in search. These include target-relevant features, saliency, scene context, and object co-occurrence. Of these information sources, the most commonly used components of computational models for visual search have been saliency (Itti & Koch, 2000) and target-relevant features. In the following section I will briefly discuss each of these four sources of information and what makes them a good or bad candidate for use in a computational model of search. I will close this section by describing two candidate decision rules for fixation selection.

1.2.1 Target-relevant Features

One reasonable source of information in visual search are features associated with the target. The reasoning for this is intuitive: if you are searching for a red apple you should direct your gaze towards things that look like the target (e.g. red, circular objects). The evidence is quite strong that target-relevant features provide information about the location of the target, and that humans use this information to guide their search (Findlay, 1997; Eckstein, Beutter, Pham, Shimozaki, & Stone, 2007; Ludwig, Eckstein, & Beutter, 2007; Rajashekar, Bovick, & Cormack, 2006; Tavassoli, Linde, & Cormack, 2009). In computational modeling, a popular approach to quantifying the effect of target-relevant features is template matching. Template matching consists

of computing a measure of similarity between a template of the target and each search element in the display. For example, when searching for a Gabor embedded in Gaussian white noise the template would be a copy of the target itself. In practice template matching is carried out by convolving the target template with an entire image, generating a 'similarity map'. Regions of the similarity map that respond most strongly (i.e. have features that are similar to the target) will have the greatest activation following convolution, and will therefore be most likely to draw the observer's gaze.

Unfortunately, template matching has some limitations. First, it is surely affected by the degraded resolution of information in the periphery. As such, models that use target-relevant features are most successful when constrained with visibility maps (Najemnik & Geisler, 2005; Zelinsky, 2008; Zelinsky et al., 2013). Another limitation is that while template matching is easy to carry out with artificial stimuli (e.g. natural objects on a roughly uniform background or Gabors embedded in noise) we don't know what the template looks like for real-world scenes. In part this is due to variations in pose, lighting, and scale that occur naturally and are uncontrollable. Occlusions also present a unique problem by hiding part of the target from view, meaning the template must be flexible enough to compensate for only part of the target being in view. Many real world searches are also categorical in nature, meaning the task is to find an unspecified member of a category (e.g. find any mug). In this case the template will almost never match the target exactly. These variations in appearance and uncertainty about target parameters have been shown to degrade search performance relative to tasks where the search target

matches the template exactly (Eckstein & Abbey, 2001; Bravo & Farid, 2009; Vickery, King, & Jiang, 2005; Wolfe, Horowitz, Kenner, Hyle, & Vasan, 2004; Ludwig et al., 2007). However, humans are able to at least partially compensate for variations in target appearance that reduce the match with the search template (Bravo & Farid, 2009). This means humans can use target-relevant features in a flexible manner to compensate for changes in pose, lighting, and other sources of naturally arising variation. Any computational model of visual search that uses target-relevant features as a source of information should also be able to account for this invariance to some degree. Some computational modeling approaches have taken this into account by computing features that are invariant to these naturally arising fluctuations in target appearance (e.g., Lowe, 1999; Dalal & Triggs, 2005). In these approaches, features are extracted using linear filters that are selective for particular spatial frequencies and orientations. Linear filters are considered a model of early visual neurons, so this method of feature extraction has some biological plausibility. Each linear filter essentially functions as a 'feature detector' for a given spatial frequency and orientation. These approaches fare reasonably well, although their object detection performance is still inferior to that of humans. This implies that these approaches may not be using the same target-relevant features as humans, or that humans are better able to compensate for variations in target appearance. However, these approaches do come close to human performance and therefore may be a good approximation.

1.2.2 Saliency

Saliency in its original conception is essentially a measure of distinctiveness, or how different the statistics in a local image region are given the statistics in the rest of the image (Itti & Koch, 2000). In this sense, saliency matches the intuitive definition of local image regions that stand out from their surroundings. Thinking of anecdotal examples where saliency might be informative about probable target locations in visual search are extremely easy, such as a red target embedded among blue distractors. This is an example of a pop-out search, where targets are found so rapidly that search times do not increase significantly as a function of the number of distractors. Regions where the statistics match the rest of the image have low saliency (i.e. this region is not 'distinctive' given the rest of the image, like the blue distractors) while regions where the statistics are unlikely given the rest of the image are highly salient (i.e. this region is 'distinctive,' like the red target). For the purposes of this review, I will use the original definition of saliency described in Itti & Koch (2000).

The ease of its computation has made saliency a popular component of computational models of visual search (e.g., Torralba et al., 2006; Parkhurst, Law, & Niebur, 2002; Nakayama & Martini, 2011). The essential computation can be framed probabilistically as $p\left(L|G\right)^{-1}$ where L corresponds to the features in a local region of the image (obtained by linear filtering) and G is the distribution of features measured over the entire image. The inverse of $p\left(L|G\right)$ is taken since saliency is a measure of how unlikely a set of features

is, meaning local features that are more probable given the rest of the image should have low saliency and vice versa. However, a fundamental problem with saliency is that highly salient regions of an image are not necessarily likely target locations. For example, saliency would not be useful if the task was to find a blue target among blue distractors and one red distractor. In this situation, saliency would guide the observer's gaze towards the red distractor rather than the target. Therefore, saliency is only informative when the target is salient compared to the rest of the image, otherwise it can actually impair search performance by drawing eye-movements to distractors. Critically, this is because the computation of saliency does not include a term associated with the target. While saliency has been shown to be a good predictor of fixations in free viewing tasks (which do not have a target defined), it is a bad predictor for search tasks where a target has been defined (Torralba et al., 2006; Einhauser, Rutishausher, & Kock, 2008; Foulsham & Underwood, 2008; Pomplun, 2006; Tatler, Hayhoe, Land, & Ballard, 2011). Therefore, although Itti & Koch (2000) saliency is easily quantifiable it is a poor source of information for visual search tasks.

1.2.3 Scene Context

Scene context refers to the structure of natural scenes. Scenes with similar contexts (e.g. forest scenes) tend to have similar configurations of features, such as orientations and spatial frequencies. However, instead of operating over localized regions of the image like target-relevant features, scene context

acts over the entire image. For example, forest scenes have many vertically oriented edges (the trees) connected to horizontal edges (the grass), whereas beach scenes have very few vertical edges and a few prominent horizontal edges (denoting the divisions between the beach, ocean, and horizon). Importantly, scene context often naturally reduces the set of probable target locations. For example, when searching for an apple in a forest two highly probable target regions would be tree branches and the forest floor, but in a kitchen scene the highly probable locations would be on counter tops. Since scenes with similar context often have extremely similar structures, the set of highly probable target locations tends to be the same. There is strong experimental evidence that scene context does guide eye-movements in search, as a wide range of studies have consistently found effects of guidance by scene context (e.g., Castelhano & Heaven, 2010; Ehinger et al., 2009; Hidalgo-Sotelo, Oliva, & Torralba, 2005; Neider & Zelinsky, 2006; Torralba et al., 2006). These studies find that scene context can greatly reduce search times by restricting the set possible target locations, thereby reducing the number of image regions that must be fixated.

Scene context also has the nice feature of being quantifiable in a measure known as 'gist' (Oliva & Torralba, 2006). Scene gist is determined by using computer vision methods to pool together the output of groups of linear filters at different scales and orientations. The computation of gist is closely related to that of object detection algorithms in computer science, but has an important distinction. Rather than being computed for a single object at a time, gist representations are computed over an entire image. This means

gist representations essentially consider an entire scene as one object made up of different texture regions. One way to represent a forest scene is to use an object-based representation where each individual object is represented (e.g. trees, grass, animals, etc.) with a collection of filter outputs. Conversely, the gist representation of that forest scene would consist of a few large texture regions, with each texture region defined by a particular configuration of filter outputs. It has been hypothesized that scene gist builds up rapidly over the first few hundred milliseconds of viewing an image since research has shown humans can reliably categorize scenes even when viewed for less than 200 ms (Potter & Levy, 1969). This means it is more likely that humans represent scene context using a representation similar to gist, as opposed to a collection of objects representation. Additionally, since each fixation lasts approximately 200-300 ms, the fact that humans can reliably categorize scenes with less than a full length fixation indicates that the perception of scene context may not be impaired by our foveated visual system. Therefore, it is not necessary for computational models to constrain gist representations with visibility maps.

1.2.4 Object Co-occurrence

Object co-occurrence refers to the fact that certain objects tend to consistently be located near each other in natural images. The presence of a particular object can then act as a cue that a search target is nearby. For example, pencils may tend to be located near notebooks. If you are given the task of finding a pencil then locations adjacent to notebooks are highly likely to contain the tar-

get. Indeed, research on object co-occurence has shown that humans use it to guide their searches towards expected target locations (Castelhano & Heaven, 2011; Eckstein, Drescher, & Shimozaki, 2006; Mack & Eckstein, 2011). The differences between object co-occurrence and scene context are subtle. Scene context describes information obtained from representations of the scene as a whole, whereas object co-occurrence describes a spatial relationship between objects in space. Object co-occurrence is more similar to the 'collection of objects' representation of a scene described in the Scene Context section. This means that information from object co-occurrence will operate at a finer scale than information from scene context.

Although object co-occurrence provides information about highly probable target locations, it has so far not been integrated into a computational model of visual search. This is in large part because object co-occurrence is difficult to quantify explicitly. Quantifying the effects of object co-occurrence requires measuring the conditional distribution of target locations given associated objects (e.g. $p\,(Pencil|Notebook)$). However, measuring the conditional distribution for even one target requires an immense amount of training data, both in the number of images needed and the number of associated objects that must be conditioned on. Additionally, object co-occurrence is only discussed on the level of object-to-object relationships, but it could actually be that features associated with co-occurring objects are the true guiding force here. This would mean that rather than measuring the conditional distributions between associated objects and the search target, we would need to measure the conditional distributions between features of associated objects and the search

target. Finally, it is also possible that gist measurements of scene context include the effects of object co-occurrence as well. Since scene context is a larger scale source of information it may already capture the effects of object co-occurrences. Given particular configurations of features in an image, a gist representation can tell us which locations are more likely to contain the search target. Since particular contexts tend to have a set of objects that frequently appear there, it may be that part of the gist representation is actually feature configurations arising from frequently co-occurring objects. So although measuring the information present in object co-occurrence is difficult, it may be unnecessary.

1.2.5 Decision Rules

Decision rules for fixation selection in visual search tasks tend to fall into two classes: maximum a posteriori (MAP) decision rules and information gain decision rules. MAP decision rules are based around the idea that the best strategy is to fixate the location that has the highest probability of containing the target (e.g., Beutter, Eckstein, & Stone, 2003; Zelinsky, 2008; Zelinsky et al., 2013). If the target is not there, an observer using a MAP decision rule would fixate the next most likely location, and so forth. Some models make use of an averaged-MAP decision rule, where the posterior is thresholded and the average of the remaining locations is used instead of the true MAP (Zelinsky et al., 2013), but the main idea is the same.

Information gain decision rules, also known as uncertainy reduction decision rules, take the long-view and instead try to maximize the information gained on each fixation (e.g., Najemnik & Geisler, 2005, 2008; Vincent, 2011). There are many situations where the two decision rules select similar locations for fixation, in large part because ruling out highly probable locations often substantially reduces uncertainty. One major difference is that uncertainty reduction decision rules also produce 'exclusion' fixations, which are fixations to locations with high uncertainty. The largest deviations between the two decision rules are seen in situations where the signal-to-noise ratio in the periphery falls off gradually compared with the fovea (Zhang & Eckstein, 2010).

Research on which decision rule humans use has been mixed, with some studies finding support for uncertainty reduction decision rules (Najemnik & Geisler, 2005, 2008, 2009; Ghahghaei & Verghese, 2015) and others finding support for MAP decision rules (Verghese, 2012; Ghahghaei & Verghese, 2015). Additionally, this topic has only been investigated using artificial stimuli and carefully constructed trade-offs between the rules. This means that using natural stimuli may actually be the best way to resolve this debate.

1.3 Previous Approaches

In the subsequent section I will examine some of the more successful approaches to modeling visual search in natural images. Although these models are quite different from each other, one commonality is they all have some conception of

an 'activation map,' where with higher activation are more likely to be fixated. These models fall into several different classes. The first class is saccadic targeting models which begin with the theory that eye-movements are directed towards search elements that contain features similar to the target (e.g., Eckstein et al., 2006; Findlay, 1997; Pomplun, 2006; Rao, Zelinsky, Hayhoe, & Ballard, 2002; Zelinsky, 2008; Zelinsky et al., 2013). Naturally, these models tend to place an emphasis on target-relevant features. The second class of models are those that emphasize scene context as a source of information (e.g., Torralba et al., 2006; Ehinger et al., 2009). The final class of models are uncertainty reduction models, which select fixations based on how much they will reduce the uncertainty about the target location (e.g., Najemnik & Geisler, 2005, 2008; Vincent, 2011). Similar to the saccadic targeting models, uncertainty reduction models place a large emphasis on target-relevant features. The main distinction between these two approaches is how they select fixations. Saccadic targeting models operate by fixating near highly probable target locations, while uncertainty reduction models fixate locations that will maximally reduce the model's uncertainty about the target location. Models that exclusively place an emphasis on saliency as a source of information (e.g., Itti & Koch, 2000; Nakayama & Martini, 2011) will not be discussed here since they are not normative and can't account for human search behavior.

1.3.1 The Target Acquisition Model

One of the more successful saccadic targeting models is the Target-Acquisition Model (TAM) (Zelinsky, 2008; Zelinsky et al., 2013). The TAM primarily takes advantage of target-relevant information. It also attempts to model the degraded quality of peripheral information that results from our foveated visual system. In order to search for the target, the TAM passes through several phases of processing. First, it computes an activation map of image locations where greater activation corresponds to greater likelihood of the target being present. The activation map is then thresholded and averaged. Based on the revised activation map, the TAM then generates an eye-movement towards a highly probable target location. Once a search element is brought within the region defined as the TAM's fovea it can either be rejected as a distractor or reported as the target. This process iterates until either all search elements have been examined and rejected, or the target is found.

The authors evaluated the TAM's performance by comparing it with human observers on several tasks: search in natural scenes (helicopters or tanks in landscape scenes), search for natural targets in sparse backgrounds (teddy bears in crib scenes) and artificial targets (Q's and O's on a monochromatic background). Subjects were first given either an exact image of the target they were searching for (Zelinsky, 2008) or were simply told the category of object to search for (Zelinsky et al., 2013). In this paper I will restrict my discussion to results obtained from the first two tasks since the third task is not naturalistic.

The TAM is able to produce some features of human scan paths, including fixating between search elements and not fixating previous visited locations. Fixating between search elements is a natural consequence of how the activation map is spatially averaged. However, avoiding previously visited locations is included as a heuristic in the model and is therefore not normative. When the TAM fixates a location and rejects it as a distractor that location is tagged with Gaussian distributed inhibition. If the TAM were normative this feature would not need to be added and should instead fall out of the model.

Additionally, the TAM still falls far short of actually matching humans in search. One shortcoming is that the TAM produces significantly more detection errors when compared with human observers, meaning humans outperform the TAM. This indicates that the TAM is either too constrained, or does not have access to the same information observers are using. Additionally, the TAM is only able to qualitatively match the scan paths of a subset of observers, and tends to make larger amplitude saccades. The mismatch between model and human scan paths is likely the result of using a common visibility map across all participants. Two human observers will not have the same visibility map, so assuming a common one can result in a significant mismatch with other observers. It is difficult to determine whether the TAM's tendency towards large amplitude saccades is due to the visibility map used or the imposition of a minimum saccade amplitude. It may be that the visibility map used by the authors results in a shallower falloff than actual human observers experience, meaning the model would have less blurring in the periphery than humans. Alternatively, the minimum saccade amplitude for the model may be

too high relative to human observers.

A final consideration is that the TAM does not include the effect of scene context, which can be quantified and is known to be informative in visual search. Scene context would likely have been a good source of information in the natural image search (helicopters or tanks in terrain). It's exclusion from the TAM may have hurt the model's performance in this condition. However, in searches for natural targets on sparse backgrounds scene context most likely would not be very informative due to the sparsity of the search displays. Due to its flaws, the TAM leaves significant room for improvement.

1.3.2 Contextual Guidance Model

The Contextual Guidance Model (CGM) is one of the few models that explicitly includes the effect of scene context (Torralba et al., 2006). Although it also uses saliency as a source of information I will discuss it here because the majority of its performance is derived from its use of a gist measure of scene context as a source of information. In its original conception, the CGM quantifies saliency and scene context as probabilistic sources of information (Torralba et al., 2006). Using probabilistic sources of information allows the authors to apply the normative framework provided by Bayes rule to optimally combine saliency and scene context information.

In the CGM, saliency corresponds to the likelihood and a gist representation of scene context corresponds to the prior distribution. The CGM then

constrains the possible fixation locations based on the resulting posterior distribution. The CGM's posterior distribution functions in a similar manner to the TAM's activation map where highly probable locations are selected for fixation. Locations with a high posterior probability tend be ones that are both salient and likely to contain the target based on the scene gist.

To evaluate the performance of the contextual guidance model (Torralba et al., 2006) conducted a visual search experiment. Participants were tasked with finding all the pedestrians, mugs, or paintings in a set of natural images. Targets could be present or absent, and there could also be more than one target present. Torralba (2006) simulated the CGM on the same images and compared the pattern of observer fixations to the regions selected by the CGM. The CGM was able to predict the first 1-2 fixations well above chance levels, regardless of target type (pedestrian, mug or painting) although there should be some skepticism regarding this result. Rather than selecting a single location for the next proposed fixation (as the TAM does), the CGM selects 20% of the image. If participants' fixations are within this 20% of the image then they are counted as a correctly predicted. When the CGM is limited to selecting only 10% of the image performance drops precipitously (Ehinger et al., 2009). Even when 20% of the image is selected the CGM still does a poor job for certain classes of targets (e.g. mugs).

Additionally, there is no analysis examining whether the CGM can predict features of human scan paths. It is highly unlikely that the CGM would fixate between search elements or avoid refixating previously visited located. This is

because the CGM simply highlights a section of an image and has no means for ruling out previously fixated locations. This means the CGM would be unable to prevent itself from examining previously fixated locations. Similarly the CGM would not necessarily select locations between search elements for fixation since it selects salient image regions, which tend to be objects. It also turns out that bulk of the model's performance is a product of the use of scene gist. Analysis of model predictions with only saliency (no scene gist) consistently do a poor job predicting human fixation locations (Torralba et al., 2006). Later implementations of the CGM (Ehinger et al., 2009) have attempted to rectify this deficiency by also including a term associated with target features in the form of a computer vision algorithm (Dalal & Triggs, 2005). In this implementation the likelihood comprises both saliency and target relevant features. This only produces marginally improved performance, so there is significant room for the CGM to be improved.

1.3.3 Uncertainty Reduction Models

A third class of models selects fixations based on how much they will reduce the observer's uncertainty about the target location. The ideal observer framework, which posits a theoretical machine that achieves the best possible performance on a specified task given known constraints, is one such model that has been applied to visual search tasks (Najemnik & Geisler, 2005, 2008; Vincent, 2011). Since the ideal observer makes predictions on how to achieve optimal performance it is an example of a normative model. The ideal ob-

server uses template matching constrained by a visibility map, similar to the approach taken by the TAM. However, there are two critical differences. The first is that the visibility maps used to constrain the model are measured for the ideal observer, but not for the TAM. The second difference is how fixations are selected. Rather than directing gaze towards search elements with target-like features, the ideal observer selects the location that will maximally increase the probability of localizing the target. The difference here is subtle, but it causes the ideal observer to occasionally make 'exclusion' fixations, fixations in image regions that have a high degree of uncertainty.

The first step in the ideal observer framework is to define the task so that the optimal decision rule can be derived. In Najemnik & Geisler (2005) the task was localizing a target signal (Gabor) in a field of 1/f noise. Although this task is not strictly natural, the background was naturalistic since the frequency spectrum of natural images falls off roughly 1/f. In this search task the optimal method of detecting the target is template matching through convolution. Since the visibility map in 1/f noise is not flat, the similarity measure generated by template matching must be weighted by the visibility map. The result of this computation is the posterior probability distribution of the target being at any location in the image. Again, the posterior probability distribution can be thought of as similar to the activation map in the TAM. For each subsequent fixation the ideal searcher again convolves the target template with the entire image and then integrates the result with the current posterior distribution. The optimal next fixation is then the location that will maximize the probability of correctly identifying the target, thereby also

maximally reducing uncertainty.

The ideal observer developed by Najemnik & Geisler (2005) was evaluated by comparing the fixation locations selected by the model with those of human observers. Firstly, the ideal observer replicated important features of human scan paths such as fixations falling between search elements and not fixating the same location twice. The ideal observer is also able to match the number of saccades humans require to find the target as well as the distribution of saccade amplitudes. Additionally, most of the ideal observer's fixations are concentrated in a donut-shaped region around the image center, with the greatest concentration occurring directly above and below the image center. Humans tend to fixate the same image regions, although with slightly greater frequency than the ideal would predict. Unlike the results from the TAM and CGM, results from simulating the ideal searcher don't warrant much skepticism. The lack of skepticism is motivated by the ideal observer framework being parameter free. This is important in that it allows us to draw more general conclusions about the sources of information humans use in search tasks since we can be assured that model performance is not overfit. Given that humans track the ideal relatively well, a good general conclusion is that in this task humans are using some form of template matching, and select fixation locations in a similar manner as the ideal observer.

One potential drawback of the ideal searcher is that the decision rule for selecting the next fixation location is quite complex. This calls into question how likely it is that humans are actually using this decision rule (Najemnik &

Geisler, 2008). To that end, the entropy-limit minimization (ELM) model was developed to provide a heuristic approximation to the ideal fixation selection rule (Najemnik & Geisler, 2009). Here, uncertainty is quantified in terms of Shannon entropy (Shannon & Weaver, 1949). This approach has been successful for tasks that are closely related to to classical visual search, including reading (Legge, Klitz, & Tjan, 1997) and object recognition (Renninger, Coughlan, Verghese, & Malik, 2005; Renninger, Verghese, & Coughlan, 2007). The only difference between the ELM model and ideal searcher is the selection of fixation locations on the basis of maximizing entropy reduction rather than the probability of correctly localizing the target.

Given the same task as the ideal observer, the ELM model proves to be a good approximation since it produces performance almost indistinguishable from that of the ideal (Najemnik & Geisler, 2009). The ELM model predicts almost exactly the same features of scan paths as the ideal observer, however the ideal tends to have a slightly greater bias towards the center of the image than the ELM model (Zhang & Eckstein, 2010). Since both models have the exact same prior and information this minor difference is a result of how they select fixations (the ideal seeks to maximize localization accuracy after the next fixation, whereas the ELM seeks to maximally reduce uncertainty regarding the target location).

One potential problem with the ideal observer and ELM model approaches is that they have not yet been extended to include the influence of scene context. Additionally, the ideal searcher has only been derived for tasks where

the target is a Gabor embedded in a region of 1/f noise. Although 1/f noise is naturalistic it is still a far cry from real world images. Unfortunately, extending the ideal observer or ELM model to natural scenes may be quite difficult. Using Gabors embedded in noise for stimuli allows us to derive a closed form expression for the ideal fixation selection rule, but in natural scenes this would not be possible. Instead, the ideal fixation selection rule would need to be approximated. Therefore, in order to integrate the influence of scene context into these uncertainty minimization models an approximation for the ideal fixation selection rule in natural images must first be developed.

1.4 Conclusion

In the previous sections I discussed constraints on visual processing and sources of information that are relevant for visual search tasks. I also described some of the more successful computational models for predicting fixations in these tasks. In the following section I will detail a Bayesian model of fixation selection in real-world scenes. The model incorporates scene context and target-relevant features, which are known important sources of information for visual search but have not been combined in any model of fixation selection. The model is constrained by a measured visibility map, a practice used in Najemnik & Geisler (2005; 2008; 2009) but not in Zelinsky (2013) or Torralba et al. (2006). Finally, the model is normative and is able to generate a sequence of fixations, unlike Torralba et al. (2006).

After detailing our Bayesian model of fixation selection, I will then present a visual search experiment conducted with human observers and compare model performance to human fixation patterns. Finally, I will discuss the implications of the results of the experiment and modeling simulations.

Chapter 2

A Bayesian Model of Fixation Selection

Here I present a Bayesian model of fixation selection for categorical target searches in images of real-world scenes. The model uses two sources of information: Scene context and target-relevant features. Scene context acts as the prior distribution and is measured with GIST features, while target-relevant features act as the likelihood function and are measured using Histograms of Oriented Gradients (HOG). The model is constrained by a visibility map which adds noise to the likelihood function. New fixations were selected using either a MAP or ELM decision rule. In the following sections, I describe the computation of the prior distribution, likelihood function and posterior distribution. I also detail the decision rules for the MAP and ELM observers and how the posterior is integrated across fixations. For details about the training

the likelihood function and prior distribution refer to Section 6.1, Section 6.2 and Section 6.3.

2.1 Computing the Likelihood Function

The effect of target relevant features was modeled using Histograms of Oriented Gradients (HOG). HOG features provide a fine-grained representation of the dominant orientations and scales over a specified subsection (referred to as a window) of an image. If the window closely matches the size of the target then HOG features provide a explicit representation of the target. All of the targets in this experiment were at most $2.25° \times 1.13°$ so we selected used a window size of $3.0° \times 1.5°$ to extract HOG features. This window size was selected to match the buffer between pedestrian and window edge used in Dalal & Triggs (2005). To compute the likelihood function for an image we extracted the HOG features within $3.0° \times 1.5°$ windows centered at each pixel in the image. The extracted features were projected onto the 230 most informative principal components, producing the HOG representation H for each pixel in the image. Next we computed the likelihood ratio as in Equation 2.1:

$$LR = \frac{\mathcal{L}\left(\boldsymbol{X} = 1|\boldsymbol{H}\right)}{\mathcal{L}\left(\boldsymbol{X} = 0|\boldsymbol{H}\right)} \tag{2.1}$$

where $\mathcal{L}\left(\boldsymbol{X} = 1|\boldsymbol{H}\right)$ is the likelihood of the target being present at each location in the image $\boldsymbol{X}$ given the HOG features there, and $\mathcal{L}\left(\boldsymbol{X} = 0|\boldsymbol{H}\right)$ is

the likelihood of the target not being present at each location in the image $\boldsymbol{X}$ given the HOG features there. This equation can be re-written as:

$$LR = \frac{p\left(\boldsymbol{H}|\boldsymbol{X}=1\right)}{p\left(\boldsymbol{H}|\boldsymbol{X}=0\right)} \tag{2.2}$$

where the numerator represents the probability of obtaining the HOG features at a location given that the target was present and the denominator represents the probability of obtaining the HOG features given that the target was not present.

We represented the likelihood functions for the signal and noise distributions multivariate-Gaussians. The mean μ and variance σ^2 of these distributions were fit using maximum likelihood (for more details see Apendix!!!!!!). This equation can be re-written as Equation 2.3:

$$LR = \frac{\mathcal{N}\left(\boldsymbol{H};\mu_s,\sigma_s^2\right)}{\mathcal{N}\left(\boldsymbol{H};\mu_n,\sigma_n^2\right)} \tag{2.3}$$

where μ_s and σ_s^2 are the mean and variance of the target present distribution, and μ_n and σ_n^2 are the mean and variance of the target absent distribution. This likelihood ratio was then used as the likelihood function for the image. This is equivalent to classifying each pixel as a target or distractor. Pixels with values greater than 1 are classified as targets, and those with larger likelihood ratios are more likely to be targets. Examples of the likelihood ratio for several images used in the natural search can be seen in Figure 2.1. In general, the

Figure 2.1: Results of applying the HOG classifier to several images used in the natural search task. Purple pixels correspond to those with a likelihood ratio greater than 1.

likelihood function accurately classifies the target but also classifies too many distractors as targets.

2.2 Learning the Scene Context Prior Distribution

The effect of scene context was modeled using GIST features (Oliva & Torralba, 2006). GIST features provide a coarse representation of the dominant orientations and scales in an image by pooling together the responses of filter outputs. Since the GIST representation of an image is a coarse representation features of the target will only be explicitly represented if the target is very large. In our case the target always took up 2.25° of visual angle or less and was therefore too small. However, the GIST of an image can still be predictive of target position. This is because scenes that have similar GIST representations tend to be similarly structured. Since target position is at least somewhat dependent on image structure targets are likely to appear in similar locations for images with similar GIST representations. Mixture models can take advantage of this fact by 'clustering' GIST representations according to their similarity. This allows for different prior distributions according to the scene context. We take a similar approach here, representing the prior distribution as a mixture of regressions as in Torralba et al. (2006). Equation 2.4 shows the equation for the prior distribution:

$$p\left(\boldsymbol{X}, G\right) = \sum_{n=1}^{M} p\left(n\right) p\left(G|n\right) p\left(\boldsymbol{X}|G, n\right) \tag{2.4}$$

where $\boldsymbol{X}$ is the same as before, G is the GIST representation of the im-

age and n indexes the mixture component. The first term $p(n)$ is the mixing coefficient, or the prior probability of each mixture component. The second term $p(G|n)$ is the probability of the GIST representation given the mixture component. Finally, the last term $p(\boldsymbol{X}|G,n)$ represents the probability of the target being at any location $\boldsymbol{X}$ given the GIST representation of the image and the mixture component. The mixing coefficient was represented using a multinomial distribution and the other two terms were represented as multivariate Gaussians. We now re-write Equation 2.4 as Equation 2.5:

$$p(\boldsymbol{X}, G) = \sum_{n=1}^{M} \pi_n \mathcal{N}(G; \zeta_n, \Lambda_n) \mathcal{N}(\boldsymbol{X}; \mu_n, \Sigma_n) \tag{2.5}$$

where π_n is the mixing coefficient, ζ_n and Λ_n are the mean and covariance of the distribution GIST features for component n, and μ_n and Σ_n are the mean and covariance of the distribution of target locations for component n. There is an improvement in predictive performance if the the distribution of target locations is made to also depend linearly on the GIST representation G. The mean of $p(\boldsymbol{X}|G,n)$ is then given by Equation 2.6:

$$\Omega_n = \mu_n + \boldsymbol{W}_n G \tag{2.6}$$

where $\boldsymbol{W}_n$ is the regression matrix for mixture component n. Using this form allows target location to depend linearly on the GIST representation of the image.

The scene context prior was fit using Expectation Maximization (EM) (Dempster, Laird, & Rubin, 1977). EM consists of two steps performed iteratively, an expectation step and a maximization step. In the expectation step, the expectation of the posterior distribution is computed for the current parameter values. Then in the maximization step we compute the parameters that maximize the expectation. Since the likelihood is maximized at each iteration, EM guarantees that the likelihood will increase monotonically and therefore converge to a maxima. Examples of the prior distribution for several images used in the natural search experiment can be seen in Figure 2.2. Overall the scene context prior selects relatively reasonable locations in the image.

To select the number of regression components, we performed 10-fold cross validation on the training images, which helps prevent over-fitting. The metric used to select the number of components was the proportion of times the target was in the top K of prior probability. We evaluated this for the top 5%, 10% and 20% of the prior probability. The results of this analysis can be seen in Figure 2.3. Localization accuracy continues to increase with model complexity, but only marginally after 10 regression components. Therefore, we used 10 regression components to represent the scene context prior.

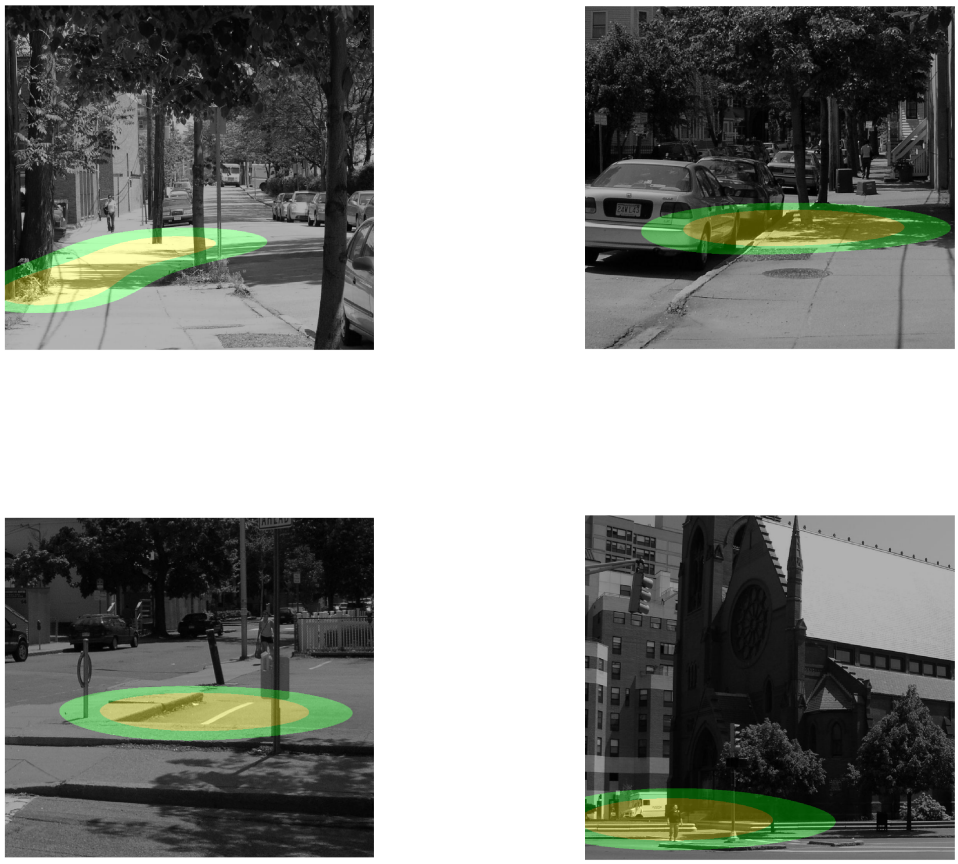

Figure 2.2: The GIST prior distribution for several images used in the natural search task. Yellow regions correspond to the top 5% of the prior probability and green corresponds to the top 10%.

2.3 Computing the Posterior

With the likelihood function and prior distribution computed, we turn to the posterior distribution give in Equation 2.7:

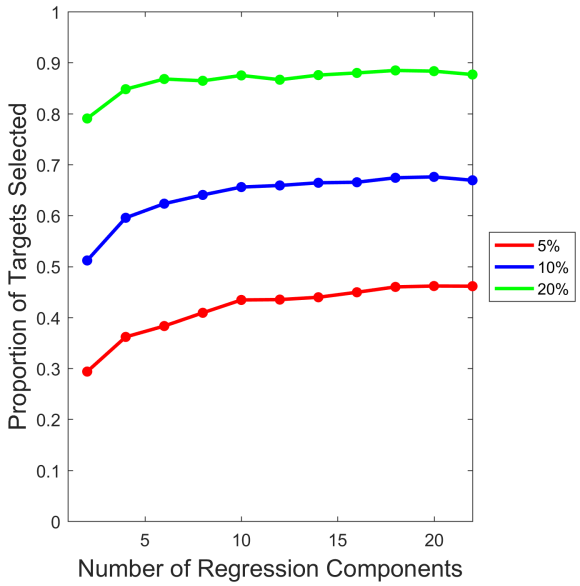

Figure 2.3: Proportion of targets selected by the top 5%, 10% or 20% of the GIST prior distribution as a function of the number of regression components.

$$P = p\left(\boldsymbol{X}|G\right)\mathcal{L}\left(\boldsymbol{X}|\boldsymbol{H}\right) \tag{2.7}$$

where P is the posterior distribution. The first term $p\left(\boldsymbol{X}|G\right)$ reflects the influence of scene context and gives the probability of the target being at any location in the image $\boldsymbol{X}$ given its GIST representation. The second term $\mathcal{L}\left(\boldsymbol{X}|\boldsymbol{H}\right)$ represents the likelihood of the target appearing at any location in the image $\boldsymbol{X}$ given the HOG features at that location. The posterior distribution can also be re-written as Equation 2.8:

$$P\left(\boldsymbol{X}|G,\boldsymbol{H}\right) = p\left(\boldsymbol{X},G\right)\frac{p\left(\boldsymbol{H}|\boldsymbol{X}=1\right)}{p\left(\boldsymbol{H}|\boldsymbol{X}=0\right)} \tag{2.8}$$

Figure 2.4: The posterior distribution for several images used in the natural search task. Purple pixels correspond to those with a positive log-posterior probability.

Examples of the posterior distribution for several images used in the natural search experiment can be seen in Figure 2.4. After incorporating the influence of scene context the classifies fewer distractors as targets but retain much of the ability to correctly classify targets.

2.4 Decision Rules

After computing the posterior distribution the Bayesian observer must select a location to fixate. Two model observers were used, one that made fixations on the basis of a MAP rule and one that made fixations on the basis of an ELM rule. Here we describe the computation of these decision rules.

To select a new fixation (denoted $k\left(T+1\right)$) the ELM observer considers each possible next fixation and the expected resulting visibility map. Here, k represents a location in the image and T represents the fixation number. The location $k_{ELM}\left(T+1\right)$ that will maximally reduce it's uncertainty about the target's location is then selected for fixation. This is computed according to Equation 2.9.

$$k_{ELM}\left(T+1\right)=\underset{k(T+1)}{\arg\max}\sum_{i=1}^{N}P_{i}\left(T\right)d_{i}^{\prime2}\left(k,T+1\right) \tag{2.9}$$

where $P_i\left(T\right)$ is the posterior at location i and $d_i^{\prime 2}\left(k,T+1\right)$ is the visibility map at fixation $T+1$. Equation 2.9 is exactly the ELM decision rule presented in Najemnik & Geisler (2009).

Conversely, the MAP observer simply selects the maximum of the posterior distribution for fixation $k_{MAP}\left(T+1\right)$ and does not consider any changes in the visibility map. The decision rule for the MAP observer is given in Equation 2.10.

$$k_{MAP}\left(T+1\right)=\arg\max_{i}P_{i}\left(T\right) \tag{2.10}$$

2.5 Updating the Posterior

After a new location has been fixated, it is necessary to re-compute the posterior distribution. To update the posterior across fixations at time T, we use the sequential updating framework developed by (Najemnik & Geisler, 2009) and presented in Equation 2.11.

$$P_{T}\left(\boldsymbol{X}|G,\boldsymbol{H}\right)=p\left(\boldsymbol{X},G\right)\sum_{t=1}^{T-1}d_{t}'^{2}\mathcal{L}\left(\boldsymbol{X}|\boldsymbol{H}\right) \tag{2.11}$$

This updating rule has the effect of weighting the likelihood function according to how detectable the target is at each location. Locations where the target is more detectable are given more weight since the fidelity of information is greater (uncertainty is lower), while locations where the target is less detectable will be pushed towards 0, indicating greater uncertainty.

Chapter 3

Methods

3.1 Participants

Data were collected from five participants. All participants had normal or corrected-to-normal vision and were paid $10 per hour for their participation.

3.2 Apparatus

Stimuli were presented on a 22in Philips 202P4 CRT monitor at a resolution of 1280 x 1024 pixels and a frame rate of 100 Hz. Participants were seated 70 cm from the display such that the display subtended 24.1° x 30.1° of visual angle. Stimuli were generated and presented using MATLAB software (Mathworks) and the Psychophysics Toolbox extensions (Brainard, 1997). Head position

was fixed using a forehead and chin rest. Participants' right-eye was tracked using an Eyelink 1000 infrared eye tracker (SR Research, Kanata, Ontario, Canada). Gaze location was sampled from the eye tracker at 500 Hz in the detection task and 1000 Hz in the search task.

3.3 Stimuli

All images of real world scenes used in this experiment, including cropped images of pedestriations, were obtained from the LabelMe Database (Russell, Torralba, Murphy, & Freeman, 2008).

3.3.1 Detection Task Stimuli

The stimulus for the detection task consisted of a random crop taken from an image of a natural scene in the LabelMe database. The random crop was then added to a circular region of $1/f$ filtered noise (the noise mask). The diameter of the natural scene patch was 6°. If the patch contained a pedestrian, the pedestrian size was approximately 2.25° x 1.5°. The noise mask had a diameter of 24° and 7.5% RMS contrast. The area surrounding the noise mask was set to the mean luminance of the display ($40cd/m^2$). Example signal and noise trials can be seen in Figure 3.1.

Before adding the natural scene patch to the noise mask we first multiplied the crop by a hanning window centered on the patch itself:

$$\bar{I} = w\,(x, y)\, I\,(x, y) \tag{3.1}$$

where I is the natural scene patch, $w\,(x, y)$ is the value of the hanning window at pixel coordinates (x, y) and $\bar{I}$ is the resulting image patch. The value of $w\,(x, y)$ was entirely dependent on the radius of the hanning window ρ, which in this case was equal to the radius of the natural scene patch. For pixels where $\sqrt{x^2 + y^2} < \rho$ the value of the hanning window is:

$$w\,(x, y) = 0.5 + 0.5 \cos\left(\pi \sqrt{x^2 + y^2}/\rho\right) \tag{3.2}$$

For all other locations the value of the hanning window is set to 0. Prior to adding the natural image patch to the noise mask, the noise mask was also multiplied by an inverse hanning window with the same radius ρ and centered on the location of the natural scene patch. The inverse hanning window is defined as:

$$\bar{I} = (1 - w\,(x, y))\, I\,(x, y) \tag{3.3}$$

This process removes edge effects as it causes the luminance of the pixels in the natural image patch to approach the mean with increasing distance from the center. It also increases the strength of the noise mask as distance from the center increases. Sample target present and target absent stimuli can be

Figure 3.1: Examples of signal and noise trials used in the detection task.

seen in Figure 3.1.

3.3.2 Search Task Stimuli

The stimulus for the search task was a cropped image of a natural scene from the LabelMe database. The diameter of the cropped natural scene patch was 24°. If the patch contained a pedestrian, the pedestrian size was approximately 2.25° x 1.5°. The area surrounding the noise mask was set to the mean luminance of the display ($40cd/m^2$).

For target present images, the location of the crop was chosen randomly from a set of valid locations for that image. We defined the set of valid locations as those that maintained a buffer of 1.88° between the upper and lower edges of the pedestrian, and 1.41° between the left and right edges and the pedestrian. Examples of target present images are show in Figure 3.2. To select a random crop for target absent images, we first selected an image size from the distribution of target present images. The target absent image was re-sized to match the size of the sampled target present image, and then a

Figure 3.2: Example target present images used in the natural search task. Targets are outlined with a red box, which did not appear in the actual experiment.

crop was chosen randomly from the same set of valid locations. Examples of target absent images are shown in Figure 3.3.

3.4 Procedure

Participants ran 10 blocks of 100 trials each for 1000 total trials during each laboratory visit for the detection task. In the search task participants ran 9 blocks of 50 trials each for 450 total trials during each laboratory visit. At

Figure 3.3: Example target absent images used in the natural search task.

the start of each block, participants completed a five-point (detection task) or nine-point (search task) calibration routine. The calibration was repeated until the average test-retest measurement error across all gaze points fell below 0.25°.

3.4.1 Detection Task

Detection performance was measured before the search task. Prior to the start of each trial participants fixated a central marker. A stroked circle cued the location to where the natural scene patch would appear. On each trial it was

randomly determined whether a target scene patch (pedestrian) or noise scene patch (background) would be presented. If a target scene patch was presented the pedestrian was always centered on the cued location. Participants initiated each trial with a button press. After a stimulus onset asynchrony (SOA) of $100 - 400$ ms (chosen randomly) one 250 ms interval of the stimulus was presented. Participants then indicated whether a target was present or not with a button press. Auditory feedback was also presented. If participants blinked or moved their eyes at any time during the trial they were notified and the trial was discarded. Natural scene patches were always presented along the horizontal meridian, moving from the center of the display to the right. The eccentricity was either $0°$, $3°$, $6°$, $9°$ or $12°$, with trials blocked by eccentricity.

At the start of a block of trials, participants completed one practice trial in which the natural scene patch was presented at its original contrast (approximately 20% RMS contrast). Data from this trial was not recorded. For the remaining trials in the block, target contrast was selected using an interleaved, adaptive procedure (Kontsevich & Tyler, 1999). To avoid clipping, the contrast of a patch was never increased above it's initial value. Participants completed the detection task in 2-3 sessions.

Target patches were selected from a set of 481 natural scene patches and their left-right reflections, for a total of 962 target patches. Noise patches were selected from a set of 634 natural scene patches and their left-right reflections, for a total of 1268 noise patches. Scene patches were never presented more than once at each eccentricity, but were sometimes presented at different

eccentricities.

3.4.2 Search Task

Prior to the start of each trial participants fixated a marker that was 14° to the left or right of center. On each trial it was randomly determined whether the target would be present or absent from the cropped scene. Participants initiated each trial with a button press. After a SOA of $100 - 400$ ms (chosen randomly) the stimulus was a presented from 2000 ms. During this time, participants were free to move their eyes around the image. Participants could report at any time whether the target was present or absent using a button press. If they did not respond within 2000 ms, the stimulus was removed from the screen. Even if the stimulus was removed subjects still responded. Auditory feedback was also presented. If participants blinked or moved their eyes at any time during the SOA they were notified and the trial was discarded.

At the beginning of each session, participants completed a practice block of five trials for them to get accustomed to the procedure. All participants completed the search task in one session.

Target present scenes were selected from a set of 222 images. Target absent scenes were selected from a set of 347 images. Scenes were presented at most once.

3.5 Simulation of the ELM and MAP Observers

The ELM and MAP observers were simulated on all target present images that the human observers ran on. Model observers were not simulated on target absent images as there was no way to assess their performance without a target. The first step in the simulation process was to compute the posterior distribution. We first extracted the GIST and HOG features from the image. We then computed the prior probability of the target being at any pixel in the image according to the previously measured prior distribution, and the likelihood according to the likelihood ratio of the signal and noise distribution of HOG features. The likelihood and prior were then combined to produce the posterior distribution as in Equation 2.7. The maximum of the posterior distribution only coincided with the bounding box of the target on 77% of the test images. As the goal of this experiment was to make normative claims about human search, we chose to only simulated the model observers on the images where the maximum of the posterior coincided with the target.

At the beginning of each simulated trial, we randomly selected a point 14° to the left or right of center as for the human observers. The next step was to apply the visibility map to the likelihood function. Due to the computation time required to simulate the ELM observer the visibility map was fit by aggregating the data from all human observers in the detection task. The d' at each point in the image was computed, and a corresponding sample of noise ϵ was drawn from the distribution defined in Equation 4.4. We then multiplied the likelihood function by the sample of Gaussian noise. Once the

noise was applied to the likelihood function, the posterior distribution was re-computed. Finally, the ELM and MAP observers selected the next fixation location according to Equation 2.9 and Equation 2.10, respectively. After selecting a new fixation location, the visibility map and posterior were updated again. Search was terminated either when the maximum of the posterior distribution coincided with the bounding box of the target or when 16 fixations had been simulated. The threshold of 16 fixations was chosen because it was the maximum number of fixations made by a human observer across all target present trials. Therefore, it represents an upper bound on the number of fixations that a human observer would be able to make in a given trial.

Chapter 4

Results

4.1 Estimating the Visibility Map

Visual sensitivity for the human observers was characterized using a visibility map, which specifies the signal-to-noise ratio, or d', as a function of target contrast and c and eccentricity ρ. Due to the computational complexity of simulating the ELM observer, we computed a common visibility map by aggregating all human observers' detection data. The visibility map was computed by taking the standard normal integral of the observers' expected accuracy given in Equation 4.1:

$$d'(c,\rho) = \sqrt{2}\Phi^{-1}\left[PC(c,\rho)\right] \tag{4.1}$$

where Φ represents the standard normal integral and $PC\,(c, \rho)$ represents the observer's expected detection accuracy given the target contrast c and psychometric function at eccentricity ρ. Detection accuracy was modeled using a cumulative Weibull function defined in Equation 4.2:

$$PC\,(c, \rho) = 1 - 0.5e^{-\frac{c}{\alpha_\rho}^{\beta}} \tag{4.2}$$

where β controls the steepness of the psychometric function and α_ρ is the contrast threshold at eccentricity ρ. Individual estimates of β did not differ significantly as a function of eccentricity, so we assumed a common steepness parameter.

Since participants made saccades in the search experiment, it was necessary to compute a continuous representation of the visibility map. To that end, we modeled contrast thresholds α_ρ as a log-linear function of eccentricity:

$$\alpha_\rho = \alpha_0 e^{\tau\rho} \tag{4.3}$$

where α_0 is the contrast threshold at the fovea and τ is a log slope parameter that controls the increase in contrast threshold as a function of eccentricity. This function has been shown to accurately describe the rise in contrast thresholds with increasing eccentricity for various tasks (Peli, Yang, & Goldstein, 1991). The resulting psychometric function has three parameters: α_0, β and τ. These parameters were jointly fit to the detection data using maximum

likelihood. Combined parameter estimates, as well as individual estimates, can be seen in Table 4.1. Overall there was little variability in parameter estimates across participants.

Table 4.1: Visibility Map Parameter Estimates for Each Human Observer and for the Combined Visibility Map

Observer	β	α_0	τ
PE	0.9396	0.0238	0.2108
AWA	0.6855	0.0195	0.3689
MS	1.0923	0.0282	0.1468
CG	0.5521	0.0150	0.3199
RS	0.7918	0.0268	0.1871
Combined	0.7442	0.0237	0.2232

To determine how much noise to add to the likelihood function we first computed the d' of the likelihood function over all training examples. We then estimated the amount of Gaussian distributed noise that needed to be added to the log-likelihood function to reduce classification performance to the d' for human observers at the given contrast and eccentricity. The noise was a randomly sampled Gaussian distributed variable as in Equation 4.4:

$$\epsilon \sim \mathcal{N}(0, \delta) \tag{4.4}$$

To estimate the variance δ for each potential d' we simulated $10,000$ random draws of ϵ per training example and computed the resulting d' of the likelihood function. We then selected the value of δ that produced the minimum deviation from the desired d'. This allows us to compute a noise-map, which consists of

random draws of ϵ at each location in the image. The value of ϵ was dependent on the contrast and eccentricity of the pixel relative to the currently fixated location. Contrast was computed over the same $3.0^\circ \times 1.5^\circ$ window used to compute HOG features. $\boldsymbol{\epsilon}$ was then added to the log-likelihood function could then to produce the constrained log-likelihood function and the posterior could be re-computed.

4.2 Model Localization Accuracy

We computed the localization accuracy of each component of the Bayesian model. This allowed us to quantify the performance of the model, as well as evaluate how much of the performance is due to each component. To this end, we computed the proportion of times the center of the target was in the top 5%, 10% or 20% of prior probability, posterior probability and likelihood function. Localization accuracy was computed using 10-fold cross-validation to ensure that estimated model accuracy was not a result of over-fitting. The results of this analysis can be seen in Figure 4.1. The combined posterior distribution achieves the highest localization accuracy, indicating that both scene context and target-relevant features are informative for this task. Interestingly, the localization accuracy of target-relevant features alone was quite close to the posterior distribution. However, the localization accuracy of scene context alone was much lower. This means that target-relevant features are providing the bulk of the information, but scene context does provide some additional

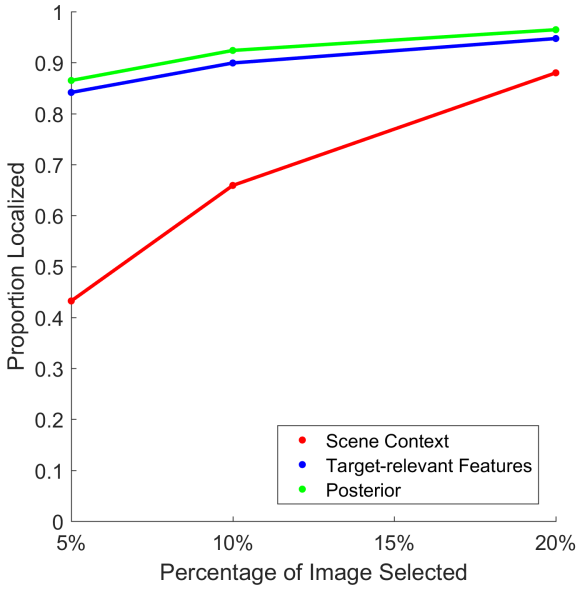

Figure 4.1: Proportion of training image targets selected in the top 5%, 10% or 20% of the prior distribution, likelihood function, and posterior distribution.

information.

We also computed localization accuracy for each image used in the search experiment. The results of this analysis are presented in Figure 4.2. We find marginally reduced accuracy for the posterior and target relevant features alone, but the localization accuracy of scene context alone decreases significantly. This indicates that the trained GIST prior generalizes poorly. This could be for several reasons, which will be discussed in detail in Section 5.4

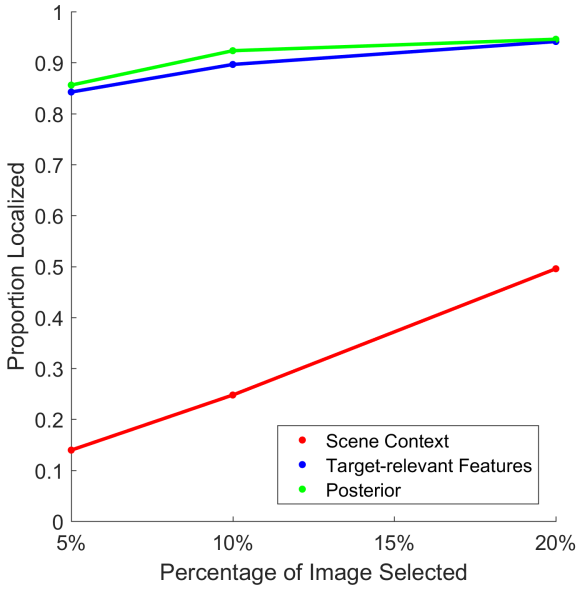

Figure 4.2: Proportion of search experiment targets selected by the top 5%, 10% or 20% of the prior distribution, likelihood function, and posterior distribution.

4.3 Human Search Performance

In this section we analyze the fixations made by human observers in the natural search task. The most important metrics in determining the performance of the human observers are the proportion correct, mean number of fixations, and mean response time. We also compute the mean saccade amplitude and mean fixation duration for each human observer. For all analyses, the first fixation was discarded as it was off-image and therefore would give observers minimal information, and arbitrarily increase mean saccade amplitudes.

The proportion correct for the human observers is presented in Figure 4.3. All human observers were more accurate on target absent trials than on target

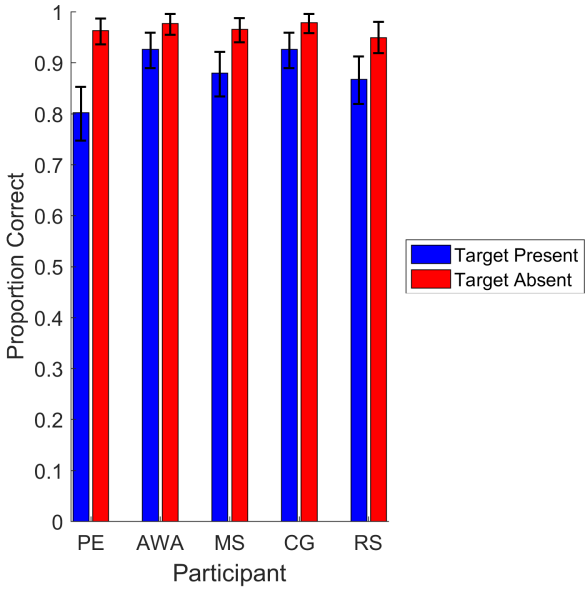

Figure 4.3: Proportion correct for each participant on target present and target absent trials. Error bars represent 95% confidence intervals.

present trials. This difference was significant for observers PE, MS and RS, but not for the other observers. Proportion correct on target present trials ranged from 80.18% to 92.63%, while accuracy on target absent trials ranged from 94.90% to 97.89%. Some of the target present trials included images where significant portions of the target were either occluded or shaded, which may explain why participants were less accurate.

The number of fixations each human observer made before responding is presented in Figure 4.4. All human observers made fewer fixations on target present trials than on target absent trials. This difference was significant for all observers except one (PE). The average number of fixations each observer made on target present trials ranged from 4.43 fixations to 7.36 fixations. For

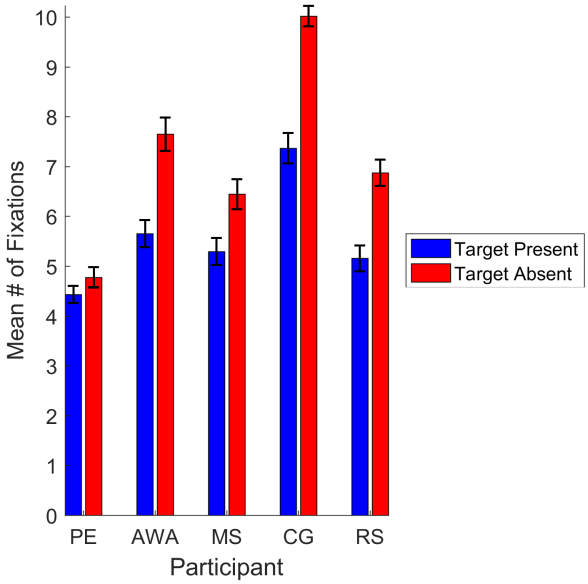

Figure 4.4: Average number of fixations for each participant on target present and target absent trials. Error bars represent 95% confidence intervals.

target absent trials, the average number of fixations ranged from 4.78 fixations to 10.02 fixations. Observers that made more fixations tended to have higher accuracy, regardless of whether the target was present or not.

A similar pattern appeared in the average response time for each human observer, presented in Figure 4.5. Response time was defined as the time between stimulus presentation and when the stimulus was removed. If the observer used more than the 2000 ms of search time then the response time was 2000 ms since the stimulus was removed then. All human observers responded faster on target present trials than on target absent trials. This difference was significant for all observers. The response time for target present trials ranged from 804 ms to 1570 ms, while the response time for target absent trials ranged

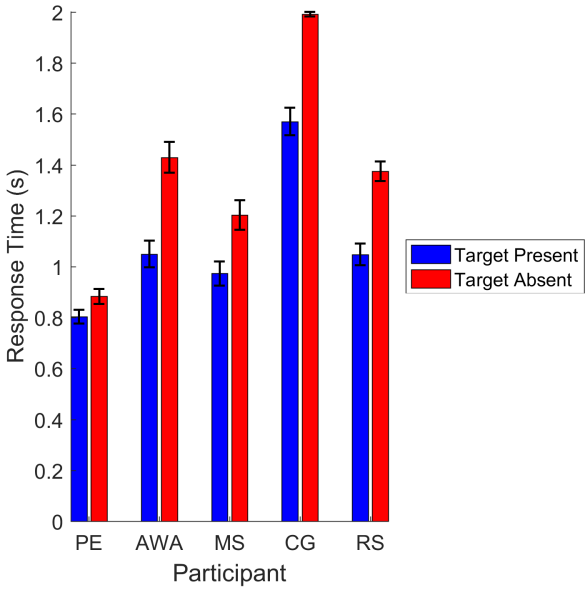

Figure 4.5: Average response time for each participant on target present and target absent trials. Error bars represent 95% confidence intervals.

from 883 ms to 1993 ms. Again, observers with higher response times tended to be more accurate and make more fixations. Observers that had higher response times also made more fixations, which is expected. Interestingly, observer CG chose to use the full search time on almost every target absent image, as opposed to only 1500 ms on the target present images.

The average saccade amplitudes for each human observer are presented in Figure 4.6. The amplitude of a saccade was computed by taking the Euclidean distance between neighboring fixations in a sequence. The average saccade amplitude was significantly smaller on target present trials than on target absent trials for all human observers. For target present trials, the average saccade amplitude ranged from 2.83° to 3.74°, but for target absent trials, the

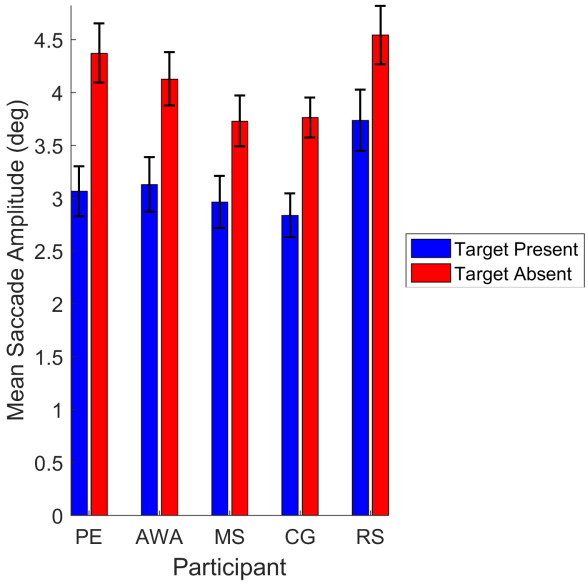

Figure 4.6: Average saccade amplitude for each participant on target present and target absent trials. Error bars represent 95% confidence intervals.

average saccade amplitude ranged from 3.73° to 4.54°. This discrepancy could be because observers made more fixations on target absent than on target present trials. In fact, When the order of fixation this effect disappeared, indicating that the cause of this difference between target present and target absent trials was due to the late sequence larger amplitude saccades on target absent trials.

The average fixation duration for each human observer is presented in Figure 4.7. For each human observer, the average fixation duration was greater on target present trials than on target absent trials, but this difference was only significant for subject CG. On target present trials the average fixation duration ranged from 146 ms to 174 ms, while on target absent trials the

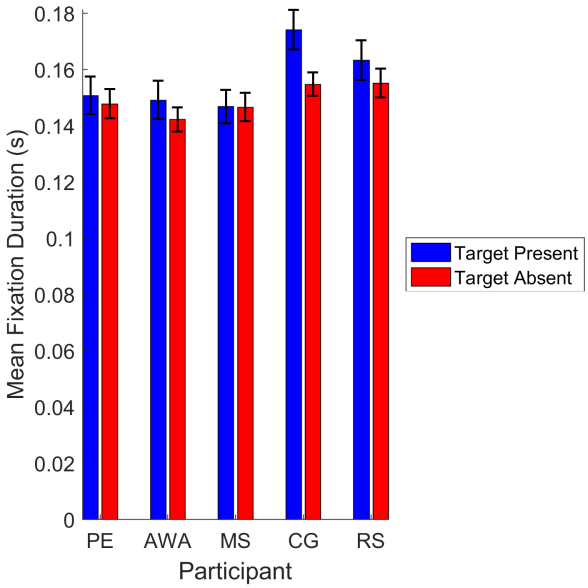

Figure 4.7: Average fixation duration for each participant on target present and target absent trials. Error bars represent 95% confidence intervals.

average fixation duration ranged from 142 ms to 155 ms.

The results presented in this section indicate that human observers were more accurate on target absent than target present trials, but made their decision faster for target present trials. There were only minor differences between the average saccade amplitude and average fixation duration, meaning that the features of each observers' saccades were roughly similar regardless of whether the target was present or absent. Therefore, although subjects were less accurate and faster on target present trials, it seems there were no differences in how they selected their fixations.

4.4 Model Simulation Results

The ELM and MAP observer were simulated on all images used in the search experiment where the maximum of the posterior coincided with the bounding box of the target. This only accounted for 77% of the images used in the search task. We chose to only simulate the model observers on these images because the goal of this study was to make normative claims about humans should search. Similar to the human observers, we computed localization accuracy, average number of fixations per trial and mean saccade amplitude. Mean fixation duration and response time were not computed, as there was no time component to the simulated searches.

4.4.1 Model Search Performance

Localization accuracy was computed as the proportion of trials where the maximum of the visibility map constrained posterior distribution coincided with the bounding box of the human on at least one fixation. As a baseline, the localization accuracy of ELM and MAP observers was compared to several sampling based decision rules:

- Sampling with replacement from a uniform distribution
- Sampling with replacement from the prior distribution
- Sampling with replacement from the likelihood function

- Sampling with replacement from the likelihood function with a visibility map
- Sampling with replacement from the posterior distribution
- Sampling with replacement from the posterior distribution with a visibility map

All of the sampling based decision rules performed extremely poorly, with localization accuracy ranging from from 5% to 13%. By contrast, the MAP and ELM observers correctly localized the target on 43% and 80% of images they were simulated on. The large gulf in performance between the sampling rules and the MAP and ELM decision rules indicates that the decision rules are a critical component of the model's performance. Additionally, the ELM observer was significantly more accurate than the MAP observer. However, both model observers were less accurate than the human observers, as can be seen in Figure 4.8. The ELM observer was significantly less accurate than two human observers (AWA and CG), while the MAP observer was significantly less accurate than all human observers.

When looking at the features of the fixations themselves, we found that the ELM observer made fewer, larger amplitude saccades than the MAP observer. Both differences can be explained by the fact that the MAP observer almost always became stuck at one location if it didn't find the target on one the first few fixations. This artificially reduced the mean saccade amplitude for the MAP observer. Additionally, because the MAP observer found the target less frequently it necessarily made more fixations since it had more trials where it reached the upper limit of 16 simulated fixations. Comparing these results

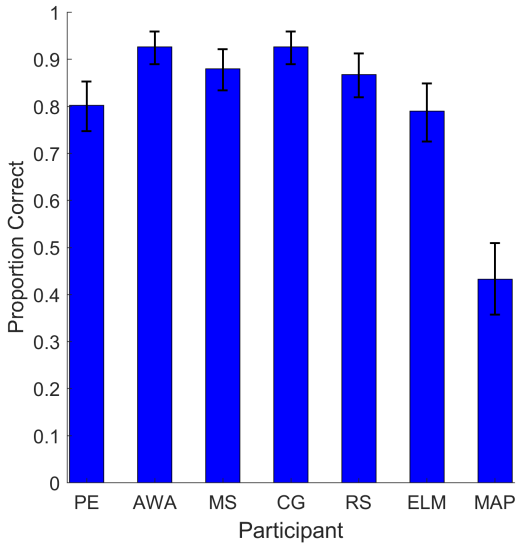

Figure 4.8: Proportion correct for each participant and model observer on trials where the posterior correctly localized the target. Error bars represent 95% confidence intervals.

to the human observers in Figure 4.9, we found that the ELM and MAP observer both made more fixations than all but one human observer (CG). The ELM and MAP observer were also significantly more variable in the number of fixations they made. This is because the models either found the target quickly or were unable to find the target and therefore simulated 16 fixations. In terms of saccade amplitude, we found that the mean saccade amplitude was lower for the MAP observer and higher for the ELM observer, relative to the human observers. These results are presented in Figure 4.10. Both of these results are expected. The MAP observer should make smaller saccades on average as it frequently selected the same location for fixation when it couldn't find the target. Larger average saccade amplitudes are expected for the ELM observer are consistent with previous work on the topic (Najemnik

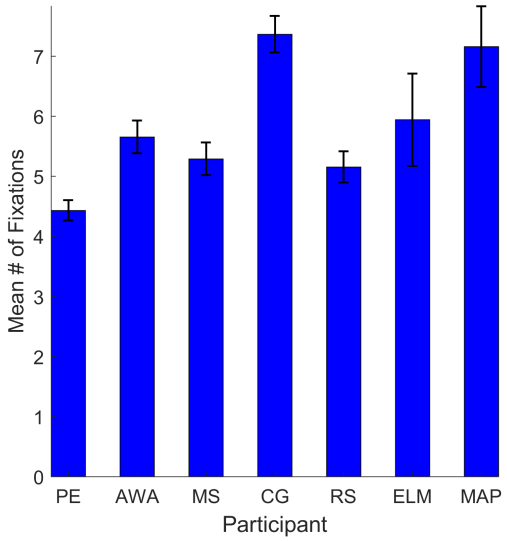

Figure 4.9: Average number of fixations per trial where the posterior correctly localized the target for each participant and model observer. Error bars represent 95% confidence intervals.

& Geisler, 2005, 2008). Taken together, these results indicate that the MAP observer was outperformed by the ELM observer, and both model observers were outperformed by the humans.

The goal of this paper is to make normative claims regarding how humans should search. Therefore, we also computed the mean number of fixations and mean saccade amplitude for the MAP and ELM observers using only the trials where they successfully found the target. These analyses are presented in Figure 4.11 and Figure 4.12. When the either model observer found the target they did so using fewer fixations than the human observers. In fact, the MAP observer now made significantly fewer fixations than the ELM observer, meaning it found that target faster. In terms of saccade amplitudes, the MAP observer made significantly larger saccades but the ELM observer remained

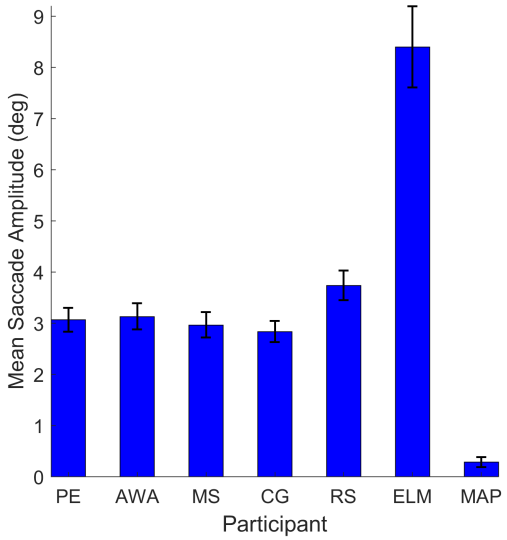

Figure 4.10: Average saccade amplitude on trials where the posterior correctly localized the target for each participant and model observer. Error bars represent 95% confidence intervals.

virtually unchanged. Taken together these results indicate that when the model observers found the target they did so using fewer fixations and larger saccades than the human observers.

4.4.2 Similarity Between Model and Human Observer Fixation Sequences

To quantify how similar the ELM and MAP observers' fixation sequences were to the human observers a natural choice is to compute the average Euclidean distance or L^2 norm between correspoding fixations in model and human observer fixation sequences. However, since the number of fixations for a given image differs between the models and human observers we needed a method

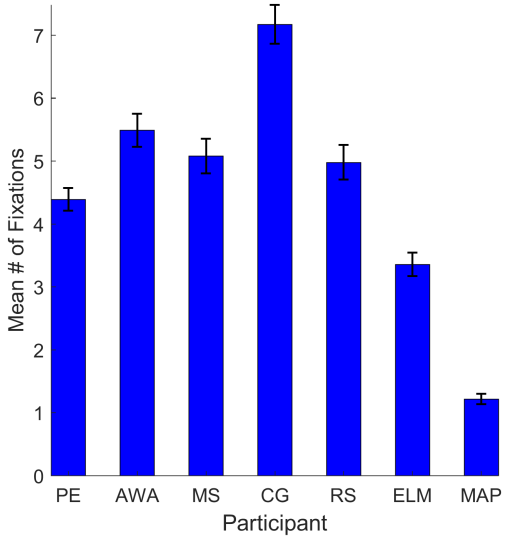

Figure 4.11: Average number of fixations per trial for each participant and model observer when the target was found. Error bars represent 95% confidence intervals.

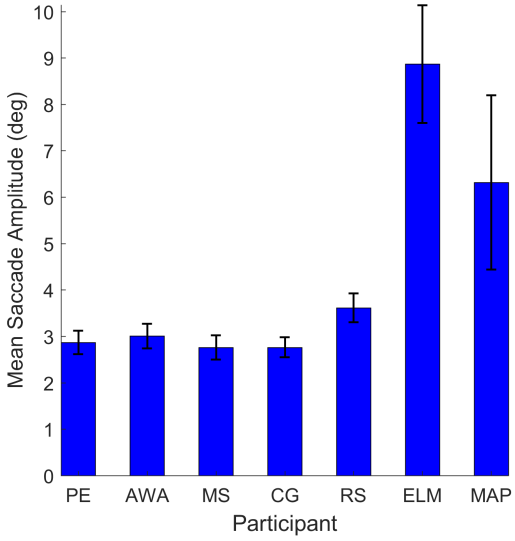

Figure 4.12: Average saccade amplitude for each participant and model observer when the target was found. Error bars represent 95% confidence intervals.

to align the fixation sequences. We take advantage of a modified form of the Needleman-Wuncsh algorithm to accomplish this task. The Needleman-Wuncsh algoritm has previously been used to align protein sequences in bioinformatics research (Needleman & Wunsch, 1970) and as part of the Scan-Match algorithm for computing fixation sequence similarity (Cristino, Mathot, Theeuwes, & Gilchrist, 2010; Mathot, Cristino, Gilchrist, & Theeuwes, 2012).

The Needleman-Wuncsh algorithm represents the problem of protein sequence alignment as a dynamic programming problem. It evaluates each possible alignment of the two sequences and gives a score to each. In a given alignment each protein in a sequence A will either be paired with a protein in sequence B or is left unpaired. If a pair of proteins are the same they are considered a 'match,' otherwise they are considered a 'mismatch.' Interestingly, leaving a protein unpaired (a 'gap') can sometimes be advantageous as it can setup a better match later in the sequence. In the standard Needleman-Wuncsh algorithm there are two free parameters: the score for a match and the score for a gap or mismatch. The algorithm then computes the score for all possible alignments and selects the minimum cost alignment as the one that is most likely. The decision for the value of these free parameters is subjective, so domain knowledge is usually required to select reasonable values.

Our modification of the Needleman-Wuncsh algorithm removes the need for a subjective decision about the score for a match or mismatch. Rather than having sequences of categorical variables (proteins), our sequences are fixations represented in Euclidean coordinates. Therefore, we simply used the

Euclidean distance between fixations that are paired together as the score. This leaves only the cost of a gap as a free parameter. Rather than selecting an arbitrary number for the cost of a gap we allowed it's value to range from 1° to 34° to evaluate its effect on the alignment score. The gap cost was capped at 34° since the images in the search experiment were 24° × 24°, and $34 = \sqrt{24^2 + 24^2}$ is therefore the maximum Euclidean distance between two on-image fixations. The modified Needleman-Wuncsh algorithm simultaneously aligns the fixation sequences and computes their similarity to each other, so used the smallest score from this algorithm as our similarity metric.

The average similarity score for the smallest gap cost (1°) and largest gap cost (34° are plotted on Figure 4.13. Regardless of gap cost, we find that the human observers' fixation sequences are more similar to the ELM observer than the MAP observer. This is reflected in the lower similarity score for every human observer when paired with the ELM observer. However, when only considering the images where the MAP and ELM observer found the target this relationship flips and the MAP observer now has the lower similarity score regardless of gap cost. Similarity scores for each model observer and human observer are plotted on Figure 4.14 for gap costs of 1° and 34°. The similarity scores for each model and human observer also decrease, meaning that the human and model observers are more similar on images where the model observers found the target.

Finally, we also computed the average similarity score between each human observer and the ELM or MAP observer, the average similarity score between

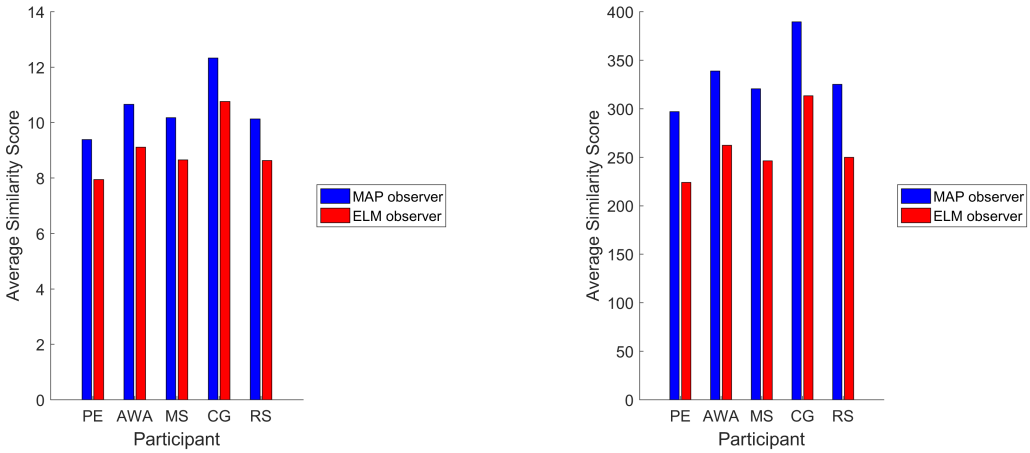

Figure 4.13: Average similarity score for each observer compared to the MAP and ELM observers when using a gap cost of 1° (left) and 34° (right).

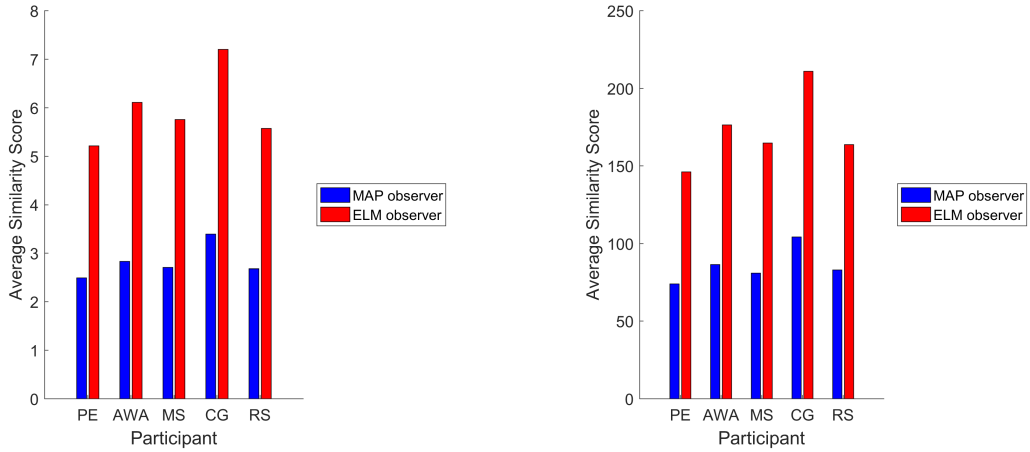

Figure 4.14: Average similarity score for each observer compared to the MAP and ELM observers when using a gap cost of 1° (left) and 34° (right) on trials where the target was correctly localized.

each human observer, and the average similarity score between the ELM and MAP observers. The results of this analysis for gap costs 1° and 34° are presented in Figure 4.15. Unsurprisingly, we found that the human observers are more similar to the ELM observer on average than the MAP observer,

Figure 4.15: Average similarity score for each aggregated pairing when using a gap cost of 1° (left) and 34° (right).

regardless of gap cost. Additionally, we find that the MAP and ELM observer were the most similar to each other out of all the pairings. This is expected since the two model observers used the exact same posterior distribution to select fixations. The most surprising aspect is that on average the human observers are less similar to each other than they are to the model observers. Most of these relationships hold true when we only consider trials where the target was correctly localized. These results are presented in Figure 4.16 for gap costs 1° and 34°. The one exception, of course, is that the human observers become more similar to the MAP observer than the ELM observer.

Examples of representative model and human observer fixations sequences are presented in Figure 4.17. These examples show the larger amplitude saccades made by the model observers, as well as instances where the MAP observer became stuck. Additionally, this examples also show the ELM observer making exclusion fixations, or fixations to highly uncertain image regions.

Figure 4.16: Average similarity score for each aggregated pairing when using a gap cost of 1° (left) and 34° (right) on trials where the target was correctly localized.

4.5 Conclusion

In this section we presented the results of the natural search task and compared the fixations simulated by the MAP and ELM observers with those of the human observers. Human observers were more accurate on target absent trials than target present trials, but also were slower as measured by number of fixations and response time. The MAP and ELM observer were less accurate than human observers, however when they did find the target they did so much faster than human observers did. We also computed the similarity between fixation sequences generated by human and model observers. Human observers were more similar to the ELM observer over all trials, but when we only analyzed trials where the target was correctly localized they were more similar to the MAP observer. Interestingly, we also found that human observers are less similar to each other than they are to the model observers.

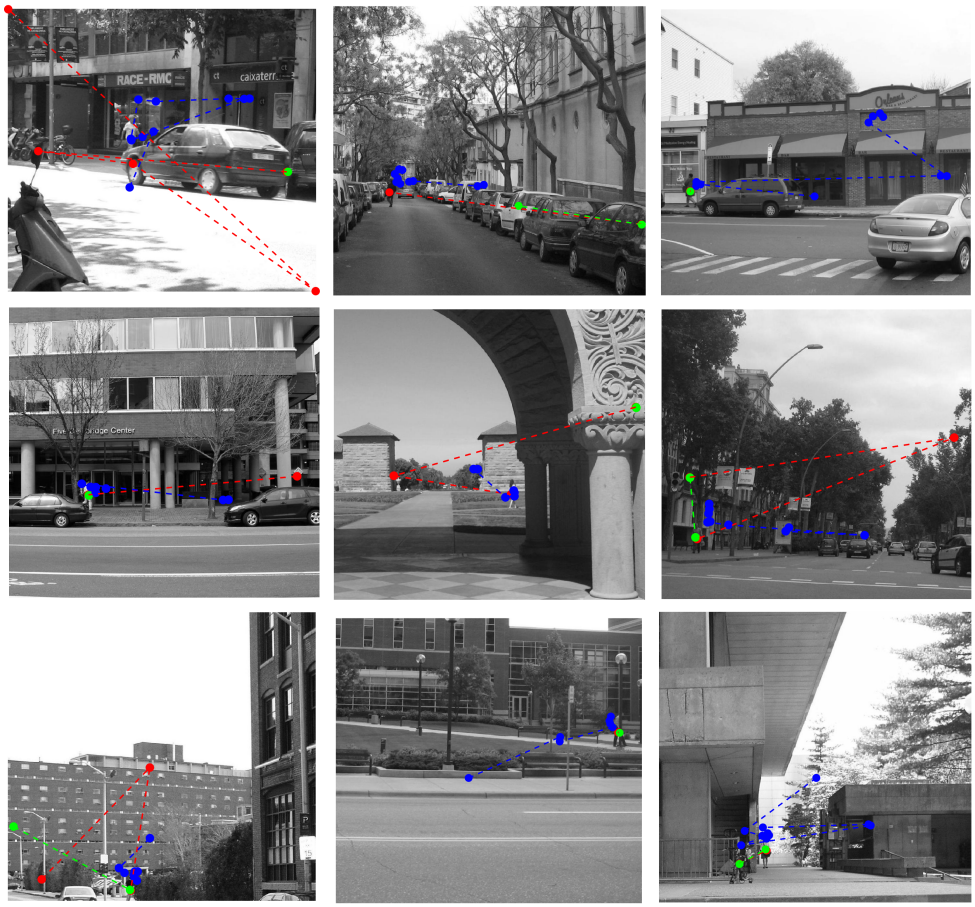

Figure 4.17: Representative fixation sequences generated by MAP (green), ELM (red) and human observers (blue) on target present images in the natural search experiment.

Chapter 5

Discussion

In the current paper we developed a Bayesian model of fixation selection for images of real world scenes. This model used of two sources of information: scene context, which is used for the prior distribution, and target-relevant features, which are used for the likelihood function. Unlike previous approaches that use these sources of information, we constrained the posterior by adding noise to the likelihood function via a measured visibility map. Additionally, the model actually selected image locations for fixation, on the basis of either a MAP or ELM decision rule. We then compared fixations simulated by the model observers to those made by human observes in a natural search task. In the following section, we discuss the implications of this work as well as its limitations and suggestions for future research.

5.1 Human Search Data

This study replicated many of the results of previous work on visual search. First, we found that human observers make more fixations and take longer to respond on target absent trials compared with target present trials, similar to others (Wolfe, 1998; Townsend, 1990; Townsend & Wenger, 2004). We also found evidence for a speed-accuracy trade-off as participants who made more fixations and took longer to respond had higher accuracy in the search task. Localization accuracy was lower for target present trials, which is most likely because some targets were quite difficult to find (e.g. they were significantly occluded or shaded). Finally, there was no difference between fixation duration or saccade amplitude on target present and target absent trials. This may mean that human observers select their fixations in a similar manner, regardless of whether a target is present or not. This result is intuitive since the human observers don't know a priori whether a target is present or not. Overall, human search data are consistent with the established literature on visual search.

5.2 Sources of Information for Fixation Selection

The Bayesian model developed in this paper used two sources of information to guide fixations: scene context and target relevant features. Scene context

was measured using GIST features and modeled as a mixture of linear regression, and target-relevant features were measured using HOG and modeled as a multivariate Gaussian distribution. Information from scene context was used as for the prior distribution, and information from target-relevant features was used for the likelihood function. We simulated two models, one that selected fixations on the basis of a MAP decision rule and one that selected fixations on the basis of an ELM decision rule. Model accuracy in the search task was lower than for the human observers, regardless of decision rule. Human observers also found the target faster on average than the model observers. This could either mean that scene context and target-relevant features are not normative sources of information or that the representations used for scene context (GIST) and target-relevant features (HOG) don't represent this information with enough fidelity.

To make truly normative claims, we only considered trials where the model observers found the target. On these trials, the model observers found the target significantly faster than the human observers, regardless of whether the ELM or MAP decision rule were used. Based on these results, we can say that scene context and target-relevant features most likely are normative sources of information. This fact indicates that the poor localization performance may be a result of GIST and HOG features not representing their respective information sources accurately enough. In sum, target-relevant features and scene context are normative sources of information, but better methods for representing these information sources must be developed.

5.3 Decision Rules

We evaluated whether human observers should use a MAP decision rule or ELM decision rule to find categorical targets in natural scenes. The results seemingly demonstrate that the ELM decision rule is the more normative approach, as it significantly outperformed the MAP decision rule in terms of both localization accuracy and speed. In fact, the MAP observer often became stuck in one image region and was unable to explore new parts of the image, which manifested itself as a low average saccade amplitude. This may be due to the integration rule for computing the posterior distribution. If the MAP observer made a saccade to a location with a high posterior probability, the only way for the MAP observer to select a new location would be for the reduction in noise from the visibility map to significantly reduce the posterior probability at the fixated location. If the posterior probability remained high after the new fixation the integration rule almost ensured that it will remain high. One possible solution would be to impose memory constraints, either to make the MAP observer memory-less so that there was no integration or to impose a decay function on previous fixations to weight them progressively less. However, due to the uncertainty surrounding how memory impacts visual search, and the inability to quantify its impact we chose not to impose memory constraints on the MAP or ELM observers. Additionally, there is no guarantee that imposing memory constraints would result in the posterior probability dropping. Another way to help the MAP observer become unstuck would be to apply inhibition of return to locations it had already visited. However, since

the purpose of this paper was to make normative claims, we chose not to use a heuristic like inhibition of return.

Broadly speaking, the fixations made by human observers are more similar to those made by the ELM observer than the MAP observer. The modified form of the Needleman-Wuncsh algorithm produced smaller scores when aligning the fixations made by the ELM and human observers than when aligning the MAP and human observers, regardless of the gap cost used in the algorithm. However, when we restricted our analysis to only images where the MAP and ELM observer localized the target we found that MAP decision rule is now the more normative approach since it found the target faster on average. Additionally, human observers were more similar to the MAP observer on these trials than they were to the ELM observer. Combined with the previous result, this indicates that human observers may flexibly switch between decision rules. Humans could use an ELM decision rule until the probability of the target being at a given location exceeds a threshold. Once this occurs, they could switch to a MAP rule to fixate the potential target location. Alternatively, humans could also start by selecting fixations according to a MAP decision rule, but then switch to an ELM decision rule when the maximum of the posterior distribution changes minimally from one fixation to the next. A hybrid decision rule would be consistent with the results of Verghese (2012), who found evidence that humans may flexibly switch between the two strategies based on task constraints. Future research will be required to determine if this is indeed the case.

5.4 Limitations and Future Directions

There are several relevant limitations of the current approach. The most obvious one is that the Bayesian model developed here was less accurate and slower than human observers. The two most likely possibilities are that humans use additional sources of information or that the representations of scene context and target-relevant features are missing information that is present in the humans' representations. To investigate the first possibility, future work could incorporate other sources of information such as object co-occurrence or saliency. Saliency seems unlikely to improve model performance, especially as it has been shown to not be a normative source of information. Object co-occurrence seems more useful, but aspects of this information source may already be captured by scene context.

The second point seems more fruitful. Poor representations of scene context and target-relevant features could be due to lack of representative training data or because the features themselves don't actually correlate with these sources of information. The most likely explanation is that the poor localization accuracy was due to both lack of training data and feature representations. The large discrepancy between accuracy of the scene context prior on training and search experiment images indicates that the model may have over-fit. However, reducing the complexity of the model by decreasing the number of regression components didn't improve localization accuracy. Therefore, it is likely that the training data used for the scene context prior was not representative enough of the search images. This is because the images used for

training data had an unconstrained number of pedestrians in them while images used in the search experiment could only have one. It may be that there are fundamental differences between images with few pedestrians and those with many. Unfortunately, the number of training images were quite small (around 1700), so restricting training to only single-target images would have reduced the size of the training set significantly. In terms of feature choices, other options for representing scene context are scant. Possibilities for target-relevant features include methods like Sift (Lowe, 1999) and convolutional neural networks (CNNs) (Hinton, Osindero, & Teh, 2006). Sift features are similar to HOG features, but CNNs may provide a better representation of target-relevant features. Most research with CNNs has been on classifying images (Krizhevsky, Sutskever, & Hinton, 2012; Simonyan & Zisserman, 2014) but they can also be used for target localization (Redmon, Divvala, Girshick, & Farhadi, 2016). CNNs have been shown to fit fine grained features of images so they may produce a better of target-relevant features. Improving methods for representing information sources and increasing the size of annotated, natural image databases are the most important areas to improve the power of computational models in natural images.

One other limitation was the implementation of the visibility maps. The relationship between eccentricity, contrast and detectability is relatively well understood for simple stimuli like Gabors in white noise. Unfortunately, the factors that influence detectability when it comes to natural images are less clear. Therefore, it is unlikely that the visibility maps estimated here take into account all of the factors that influence detection in natural scenes. A

better understanding of how visibility changes as a function of eccentricity in natural images, which factors influence it, and how to measure their influence would certainly improve the accuracy of visibility maps used here. Sebastian, Abrams & Gesiler (2017) conducted one such study that examined the impact of luminance, contrast, and similarity (as measured by cosine similarity) on detection thresholds in natural scenes. They found these three factors explained a significant portion of detection performance, but the target was a Gabor rather than a natural, categorical target. Future work is required to specify which factors influence the detectability of natural targets in natural scenes, and how they influence detectability.

Finally, the model presented here was missing one key component that is present in human visual search: a threshold for deciding when to stop searching because the target was found. In simulating the MAP and ELM observer we terminated the simulations when the maximum of the posterior distribution was inside the bounding box of the target. By contrast, when human observers are searching they use a decision threshold to decide when to stop. Once this threshold is passed the observer enters their response. Including a decision threshold would likely improve the performance of the ELM and MAP observers since they would sometimes mistakenly classify the wrong location as a target. Additionally, implementing a decision threshold would also allow us to compute the performance of the model observers on target absent images.

5.5 Conclusion

Here we described a Bayesian model of fixation selection for categorical target searches in images of natural scenes. The model uses scene context and target-relevant features as sources of information. Scene context information was used as the prior distribution and target-relevant features were used as the likelihood function. The likelihood function was also constrained by a visibility map, estimated using data from human observers in a detection task. Two models were simulated and compared to human observers on a natural search task: one that selected fixations using a MAP decision rule and one that made fixations using an ELM decision rule. We found evidence that target-relevant features and scene context are indeed normative sources of information for fixation selection. Additionally, it seems that human observers may switch between MAP and ELM decision rules rather than using one exclusively. Future work should continue to develop the tools required to develop improved computational models of visual search in real world scenes.

Chapter 6

Appendix

6.1 Images for GIST and HOG Computations

To compute the prior distribution and likelihood function it was necessary to estimate the distribution of GIST and HOG features in images of natural scenes. The GIST feature distribution was estimated using 1710 images from the LabelMe database (Russell et al., 2008). Many of these images had more than one pedestrian, producing a total of 3692 pedestrian locations across all training images. The LabelMe database was used since it was necessary to have ground truth labels of the true positions of pedestrians in the images. HOG feature distributions were estimated using 6255 positive examples and 25036 negative examples obtained from the LabelMe database and from Dalal & Triggs (2005). Positive examples were cropped images of pedestrians and

their left right reflections. Negative examples were cropped images of natural scenes where no pedestrian was present, and their left right reflections. All the images used for training were grayscale.

6.2 GIST and HOG Feature Computation

In this section we discuss how the HOG and GIST features were extracted from images of natural scenes. The basis for both features is to use linear filters to compute features over local (HOG) or global (GIST) regions of the image. Rather than using the raw HOG and GIST features we computed their principal components via principal component analysis (PCA). Using PCA allowed us to reduce the dimensionality of the HOG and GIST features and guaranteed that the features in our model would be orthogonal and Gaussian distributed. Therefore, using the principal components rather than raw features satisfies the assumptions implicit in multivariate Gaussian models.

6.2.1 GIST Feature Computation

We computed the GIST features in an image as in Torralba et al. (2006) and Ehinger et al. (2009). The image was first decomposed using a bank of steerable pyramid filters (Simoncelli & Freeman, 1995). We used filters at four scales and six orientations, for a total of 24 filters. Therefore, there are 24 filter responses per pixel in an image. Next we subsampled the filter responses

by dividing it into a 4 x 4 grid, resulting in 16 distinct blocks. Finally, the responses for each filter were averaged within each of the 16 blocks. This average operation smooths the output of the linear filters, producing a coarse representation of the scene. The resulting GIST representation G was a 384 dimensional vector (4 scales $\times$ 6 orientations $\times$ 16 blocks = 384).

6.2.2 HOG Feature Computation

The HOG features were extract in the same manner as Dalal & Triggs (2005). We first computed the gradient of the image. This was done using a simple [-1,0,1] derivative filter. The derivative filter was convolved in both the vertical and horizontal directions to compute the gradients in both the x and y directions. Next the image was divided into overlapping 'cells' of 8 x 8 pixels and a histogram of the gradients within each cell was computed. We used nine orientation bins, evenly spaced over the interval 0° to 180°. The weight added to each histogram bin was computed by bilinearly interpolating the magnitude of the gradients in both position and orientation. Using bilinear interpolation preserved some orientation tuning since intermediate gradients will contribute to two bins. Additionally, it ensured that the most important pixels, those at the center of the cell, were weighted more heavily.

In the next stage, cells were combined into 2 x 2 non-overlapping 'blocks,' such that each block consists of four cells (16 x 16 pixels). The histograms within each block were then concatenated and normalized. Finally, blocks were

combined to form 'detection windows.' Each detection window was centered on one pixel in the image, consisting of 4 x 8 non-overlapping blocks (64 x 128 pixels). The HOG representation H at each pixel was the collection of normalized block vectors aggregated over the detection window, resulting in a 3780 dimensional vector.

6.3 GIST and HOG Feature Selection

Both the GIST G and HOG H feature vectors were so high dimensional that representing them with multivariate Gaussian distributions resulted in a loss of machine precision. Therefore, we reduced the dimensionality of both of these representations using Principal Component Analysis (PCA). The output of PCA is a set of basis vectors (Principal Components) that can be linearly combined to reproduce any point in the original data set. Basis vectors are ordered according to their eigenvalue, which indicates how much variation in the original data can be explained by the corresponding basis vector. Additionally, all basis vectors are orthogonal to each other, meaning they represent independent pieces of information.

Typically, PCA is applied when one wants to reconstruct a data set in a lower dimensionality, while still preserving as much information as possible. The top 64 or 100 principal components are usually selected for these purposes since they explain the most variability. However, just because a principal component explains a large amount of variation does not mean that it is diagnostic

for the desired task. For example, in the case of HOG features we would like to be able to distinguish between two classes: targets and non-targets. It is possible, and even likely, that some principal components with large eigenvalues correspond to features that are highly variable for both targets and non-targets. Principal components like this will not be useful for differentiating the two classes. Alternatively, there will be some principal components that have small eigenvalues because their within class variability is low, even though their between class variability is high. Principal components like this would be useful for differentiating between the two classes, but would not be used if we only took the top 64 or 100 principal components.

Here, we take the approach of computing the most informative principal components for the search task. In the case of the GIST representations this means finding the principal components that are most predictive of target position, while for the HOG representations this means finding the principal components that are best for distinguishing between targets and non-targets. For both GIST and HOG representations we use 10-fold cross validation to select how many principal components to use in the model.

6.3.1 GIST Feature Selection

For the GIST representations G, we wanted to select the principal components that were most informative about target position. We assumed that target position was linearly dependent upon the GIST representation G such that:

$$\hat{y} = \boldsymbol{W} G \tag{6.1}$$

where $\hat{y}$ is the estimated target position and $\boldsymbol{W}$ is the regression matrix. For each principal component, we compute $\hat{\boldsymbol{W}}$, an estimate for $\boldsymbol{W}$, using:

$$\hat{\boldsymbol{W}} = \left(G^T G\right)^{-1} G^T y \tag{6.2}$$

where G^T is the transpose of the GIST principal component vector and y is the true target position in the image. We then computed a t-statistic for each GIST principal component:

$$t = \frac{SS_{reg}}{SS_{res}} \tag{6.3}$$

where SS_{reg} and SS_{res} are the sum of squares for the regression and residuals, respectively. Larger values of t correspond to principal components that explain more variability, and are therefore more informative about target location. The 32 principal components with the largest values of t were used to compute the prior distribution.

6.3.2 HOG Feature Selection

For the HOG representations H we wanted the best principal components classifying targets vs. non-targets. For each individual principal component, we computed the mean and variance of the signal (μ_s,σ_s^2) and noise (μ_n,σ_n^2) distributions over all training examples. For each test example we then computed the likelihood ratio:

$$LR = \frac{\mathcal{L}\left(\boldsymbol{X}=1|\boldsymbol{H}\right)}{\mathcal{L}\left(\boldsymbol{X}=0|\boldsymbol{H}\right)} \tag{6.4}$$

where $\mathcal{L}\left(\boldsymbol{H}|\boldsymbol{X}=1\right)$ is the likelihood of obtaining HOG features $\boldsymbol{H}$ given that a target was present ($\boldsymbol{X}=1$), and $\mathcal{L}\left(\boldsymbol{H}|\boldsymbol{X}=0\right)$ is the likelihood of obtaining $\boldsymbol{H}$ given that target was absent ($\boldsymbol{X}=0$). We can rewrite Equation 6.4 as:

$$LR = \frac{p\left(\boldsymbol{H}|\boldsymbol{X}=1\right)}{p\left(\boldsymbol{H}|\boldsymbol{X}=0\right)} \tag{6.5}$$

or alternatively as:

$$LR = \frac{\mathcal{N}\left(\boldsymbol{H};\mu_s,\sigma_s^2\right)}{\mathcal{N}\left(\boldsymbol{H};\mu_n,\sigma_n^2\right)} \tag{6.6}$$

where μ_s and σ_s^2 are the mean and variance of the target distribution, and μ_n and σ_n^2 are the mean and variance of the non-target distribution.

We then classified the sample as a target if the likelihood ratio is greater than 1 or a non-target if the likelihood ratio is 1 or less. Once this was done for all samples we computed the d' or signal-to-noise ratio (SNR) according to:

$$d' = \Phi^{-1}(hr) - \Phi^{-1}(fa) \tag{6.7}$$

where Φ^{-1} is the inverse of the cumulative normal integral, hr is the hit rate and fa is the false alarm rate on all samples. The 230 principal components with the largest d' were chosen for inclusion in the likelihood function. Using more principal components resulted in a loss of machine precision. To evaluate the performance of the likelihood function we used it to classify each training example using 10-fold cross-validation to prevent over-fitting. Classifying each training example we found a d' of 3.7, which indicates that the likelihood function differentiates target present and target absent image crops quite well.

References

Aivar, M. P., Hayhoe, M., Chizk, C. L., & Mruczek, R. E. B. (2005). Spatial memory and saccadic targeting in a natural task. *Journal of Vision*, *5*, 177-193.

Beutter, B. R., Eckstein, M. P., & Stone, L. S. (2003). Saccadic and perceptual performance in visual search tasks: I. contrast detection and discrimination. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, *20*, 1341-1355.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and towards structured representations. *Journal of Vision*, *11*, 1-34.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433-436.

Bravo, M. J., & Farid, H. (2009). The specificity of the search template. *Journal of Vision*, *9*, 1-9.

Carrasco, M., Evert, D. L., Chang, I., & Katz, S. M. (1995). The eccentricity effect: Target eccentricity affects performance on conjunction searches. *Perception & Psychophysics*, *57*, 1241-1261.

Carrasco, M., & Yeshurun, Y. (1998). The contribution of covert attention to the set-size and eccentricity effects in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 673-692.

Castelhano, M. S., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. *Attention, Perception & Psychophysics*, *72*, 1283-1297.

Castelhano, M. S., & Heaven, C. (2011). Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. *Psychonomic Bulletin & Review*, *18*, 890-896.

Coeffe, C., & O'Regan, J. K. (1987). Reducing the influence of non-target stimuli on saccade accuracy: predictability and latency effects. *Vision Research*, *27*, 227-240.

Cristino, F., Mathot, S., Theeuwes, J., & Gilchrist, I. D. (2010). Scanmatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, *42*, 692-700.

Dalal, N., & Triggs, B. (2005). Histogram of oriented gradients for human detection. In *Ieee conference on computer vision and pattern recognition* (p. 886-893). Los Alamitos, CA: IEEE Computer Society.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1-38.

Eckstein, M. P., & Abbey, C. K. (2001). Model observers for signal-known-statistically tasks (sks). *Proceedings of the SPIE Medical Imaging*, *4324*, 91-102.

Eckstein, M. P., Beutter, B. R., Pham, B. T., Shimozaki, S. S., & Stone, L. S. (2007). Similar neural representations of the target for saccades and perception during search. *Journal of Neuroscience*, *27*, 1266-1270.

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and bayesian priors. *Psychological Science: A Journal of the American Psychological Society (APS)*, *17*,

973-980.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, *17*, 945-978.

Einhauser, W., Rutishausher, U., & Kock, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*, 1-19.

Epelboim, J., & Suppes, P. (2001). The role of eye movements in solving geometry problems. *Vision Research*, *41*, 1561-1574.

Findlay, J. M. (1997). Saccade target selection during visual search. *Vision Research*, *37*, 617-631.

Fischer, B. (1973). Overlap of receptive field centers and representation of the visual field in cat's optic tract. *Vision Research*, *13*, 2113-2120.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? spatial and sequential aspects of fixation during encoding and recognition. *Journal of Vision*, *8*, 1-17.

Geisler, W. S., & Chou, K.-L. (1995). Separation of low-level and high-level factors in complex tasks: Visual search. *Psychological Review*, *102*, 356-378.

Ghahghaei, S., & Verghese, P. (2015). Efficient saccade planning requires time and clear choices. *Vision Research*, *113*, 125-136.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*, 188-194.

He, P., & Kowler, E. (1989). The role of location probability in the programming of saccades: Implications for "center-of-gravity" tendencies. *Vision Research*, *29*, 1165-1181.

Hidalgo-Sotelo, B., Oliva, A., & Torralba, A. (2005). Human learning of contextual priors for object search: Where does the time go? In *Computer vision and pattern recognition workshop* (p. 86). Los Alamitos, CA: IEEE Computer Society.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527-1554.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature*, *394*, 575-577.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489-1506.

Kaufman, L., & Richards, W. (1969). "center-of-gravity" tendencies for fixations and flow patterns. *Perception & Psychophysics*, *5*, 81-85.

Klein, R. M. (1988). Inhibitory tagging system facilitates visual search. *Nature*, *334*, 430-431.

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Science*, *4*, 138-147.

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*, 2729-2737.

Krizhevsky, A., Sutskever, S., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *NIPS Proceedings of the 25th International Conference on Neural Information Processing Systems*, *1*, 1097-1105.

Land, M. (2009). Vision, eye movements, and natural behavior. *Visual Neuroscience*, *26*, 51-62.

Land, M., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*, 3559-3566.

Legge, G. E., Klitz, T. S., & Tjan, B. S. (1997). Mr. chips: An ideal-observer model of reading. *Psychological Review*, *104*, 524-553.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the international conference on computer vision* (p. 1150-1157). Corfu, Greece: International Conference on Computer Vision.

Ludwig, C. J. H., Eckstein, M. P., & Beutter, B. R. (2007). Limited flexibility in the filter underlying saccadic targeting. *Vision Research*, *47*, 887-904.

Mack, S. C., & Eckstein, M. P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of Vision*, *11*.

Mathot, S., Cristino, F., Gilchrist, I. D., & Theeuwes, J. (2012). A simple way to estimate similarity between pairs of eye sequences. *Journal of Eye Movement Research*, *5*, 1-15.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.

Moore, C. M., & Egeth, H. (1998). How does feature-based attention affect visual processing? *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1296-1310.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*, 387-391.

Najemnik, J., & Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision*, *8*, 1-14.

Najemnik, J., & Geisler, W. S. (2009). Simple summation rule for optimal fixation selection in visual search. *Vision Research*, *49*, 1286-1294.

Nakayama, K., & Martini, P. (2011). Situating visual search. *Vision Research*, *51*, 1526-1537.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*, 443-453.

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*, 614-621.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23-36.

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research*, *40*, 1227-1268.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*, 107-123.

Peli, E., Yang, J., & Goldstein, R. B. (1991). Image invariance with changes in size: the role of pheripheral contrast thresholds. *Journal of the Optical Society of America. A, Optics and image science*, *2*, 1508-1532.

Pomplun, M. (2006). Saccadicselectivity in complex visual search displays. *Vision Research*, *46*, 1886-1900.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, *81*, 10-15.

Rajashekar, U., Bovick, A. C., & Cormack, L. K. (2006). Visual search in noise: Revealing the influence of structural cues by gaze-contingent classification image analysis. *Journal of Vision*, *6*, 379-386.

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research*, *42*, 1447-1463.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *2016 ieee conference on computer vision and patter recognition.* Los Alamitos, CA: IEEE Computer Society.

Renninger, L. W., Coughlan, J., Verghese, P., & Malik, J. (2005). An information maximization model of eye movements. *Advances in Neural Information Processing Systems*, *17*, 1121-1128.

Renninger, L. W., Verghese, P., & Coughlan, J. (2007). Where to look next? eye movements reduce local uncertainty. *Journal of Vision*, *7*, 1-17.

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, *7*, 1-22.

Russell, B. C., Torralba, A., Murphy, K. W., & Freeman, W. T. (2008). Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*, 157-173.

Scialfa, C. T., & Joffe, K. M. (1998). Response times and eye movements in feature and conjunction search as a function of target eccentricity.

Perception & Psychophysics, *60*, 1067-1082.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.

Shih, S. I., & Sperling, G. (1996). Is there feature-based attentional selection in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 758-779.

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *International conference on image processing* (p. 444-447). Washington, D.C.: IEEE Computer Society.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Computing Resource Repository*, *1409.1556*.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*, 1-23.

Tavassoli, A., Linde, A. C., I. Bovik, & Cormack, L. K. (2009). Eye movements selective for spatial frequency and orientation during active visual search. *Vision Research*, *49*, 173-181.

Tipper, S. P., Weaver, B., Jerreat, L. M., & Burak, A. L. (1994). Object-based and environment-based inhibition of return of visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 478-499.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*, 766-786.

Townsend, J. T. (1990). Seriel and parallel processing: Sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished. *Psychological Science*, *1*, 46-54.

Townsend, J. T., & Wenger, M. J. (2004). The serial-parallel dilemma: A case study in a linkage theory and method. *Psychonomic Bulletin & Review*, *11*, 391-418.

Verghese, P. (2012). Active search for multiple targets is inefficient. *Vision Research*, *74*, 61-71.

Vickery, T. J., King, L.-W., & Jiang, Y. (2005). Setting up the target template in visual search. *Journal of Vision*, *5*, 81-92.

Vincent, B. T. (2011). Covert visual search: Prior beliefs are optimally combined with sensory evidence. *Journal of Vision*, *11*, 1-15.

Vlaskamp, B. N. S., & Hooge, I. T. C. (2006). Crowding degrades saccadic search performance. *Vision Research*, *46*, 417-425.

Vlaskamp, B. N. S., Over, E. A. B., & Hooge, I. T. C. (2005). Saccadic search performance: The effect of element spacing. *Experimental Brain Research*, *167*, 246-259.

Wolfe, J. M. (1998). Visual search in continuous naturalistic stimuli. *Vision Research*, *34*, 1187-1195.

Wolfe, J. M., Horowitz, T. S., Kenner, N., Hyle, M., & Vasan, N. (2004). How fast can you change your mind? the speed of top-down guidance in visual search. *Vision Research*, *44*, 1411-1426.

Woodman, G. F., & Chun, M. M. (2006). The role of working memory and long-term memory in visual search. *Visual Cognition*, *14*, 808-830.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*, 787-835.

Zelinsky, G. J., Adeli, H., Peng, Y., & Samaras, D. (2013). Modelling eye movements in a categorical search task. *Phil Trans R Soc B*, *368*, 1-12.

Zhang, S., & Eckstein, M. P. (2010). Evolution and optimality of similar neural mechanisms for perception and action during search. *PLoS Computational Biology*, *6*.