# Eye Movement Guidance in Familiar Visual Scenes: A Role for Scene Specific Location Priors in Search

by

Barbara Hidalgo-Sotelo

B.S. Electrical Engineering
B.S. Biology
University of Texas at Austin, 2003



Submitted to the Department of Brain and Cognitive Sciences in
partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Cognitive Science
at the
MASSASCHUSETTS INSTITUTE OF TECHNOLOGY
September 2010



Author .....
Department of Brain and Cognitive Sciences
August 9. 2010

Certified By .....
Aude Oliva, PhD
Associate Professor of Brain and Cognitive Science
Thesis Supervisor

Accepted By .....
Earl Miller, PhD
Picower Professor of Neuroscience
Director, Brain and Cognitive Sciences Graduate Program

# Eye Movement Guidance in Familiar Visual Scenes: A Role for Scene Specific Location Priors in Search

by
Barbara Hidalgo-Sotelo

Submitted to the Department of Brain and Cognitive Sciences
on August 9, 2010 in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Cognitive Science

## ABSTRACT

Ecologically relevant search typically requires making rapid and strategic eye movements in complex, cluttered environments. Attention allocation is known to be influenced by low level image features, visual scene context, and top down task constraints. Scene specific context develops when observers repeatedly search the same environment (e.g. one's workplace or home) and this often leads to faster search performance. How does prior experience influence the deployment of eye movements when searching a familiar scene? One challenge lies in distinguishing between the roles of scene specific experience and general scene knowledge. Chapter 1 investigates eye guidance in novel scenes by comparing how well several models of search guidance predict fixation locations, and establishes a benchmark for inter-observer fixation agreement. Chapters 2 and 3 explore spatial and temporal characteristics of eye guidance from scene specific location priors. Chapter 2 describes comparative map analysis, a novel technique for analyzing spatial patterns in eye movement data, and reveals that past history influences fixation selection in three search experiments. In Chapter 3, two experiments use a response-deadline approach to investigate the time course of memory-based search guidance. Altogether, these results describe how using long-term memory of scene specific representations can effectively guide the eyes to informative regions when searching a familiar scene.

Thesis supervisor: Aude Oliva
Title: Associate Professor in Brain and Cognitive Sciences

# Table of Contents

## Acknowledgements

First of all, I want to thank my advisor, Aude Oliva, for more reasons than can be described here. From my first visit to the lab, her passion, creativity, and devotion to pursing science were immediately inspiring. Throughout my graduate school experience, she has taught, encouraged, and supported my intellectual curiosities as well as my overall health and well being. Without the help of her guidance, endless patience, and belief in my abilities, I would not be here today. For this, I will be eternally grateful.

I would also like to thank my thesis committee members, Molly Potter, Jeremy Wolfe, and George Alvarez. My research has undoubtedly benefited from their helpful comments and advice. Somehow, each conversation with these individuals gave me a sense that I could and should continue my pursue my research efforts, even when I had doubts. I am extremely lucky to benefit from the intellectual rigor that my entire committee has brought to my research program.

Very importantly, my life would not have been the same without my labmates. I had no idea, when I arrived at MIT, the difference that having such remarkable lab colleagues (friends) would make in my life. Whether I was giving a lab meeting, asking for methods or statistics advice, or simply wanting to bounce off ideas, my labmates were an amazing resource. Also, they are simply a lot of fun. For this, I would like to thank the members of CVCL, past and present: Tim Brady, Emmanuelle Boloix, Michelle Greene, Krista Ehinger, Olivier Joubert, Talia Konkle, and Soojin Park.

Life also would not have been as enjoyable without the presence of a truly extraordinary set of friends in the department (and partners of friends in the department). For being such smart, funny, and supportive friends, I simply must thank Dominique and Tracy Pritchett, Jenn Olejarczyk, Retsina Meyer, Reuben Goodman, Todd Thompson, Ethan Meyers, John Kramer, Vikash Mansinghka, and Jim Munch. I am also grateful to my partner, Ed Vul, for making life much more fun by having someone awesome to share it with.

Lastly, I would like to acknowledge my family. To my little sister, Alex Hidalgo-Sotelo, whose positiveness of spirit makes my life richer every day. Finally, I would not be the person I am today without the role that my incredible parents, Rosalba Sotelo and Domingo Hidalgo, have played in my life. Someday, I would like to have a fraction of their strength and wisdom.

CHAPTER 1

# Introduction

As humans, our perception of the world is mediated by sensory and cognitive constraints that limit the extent to which information can be acquired with a single gaze. Ecologically important activities such as navigation and search rely on a series of saccadic eye movements to inspect and interpret the visual scene. The capacity for efficient visual search, therefore, is an important part of adaptive behavior. Decades of study have produced a rich body of knowledge about visual search behavior in artificial displays and, correspondingly, been used to develop models of attentional processing (e.g. Duncan and Humphreys, 1989; Triesman & Gelade, 1980; Wolfe, 1994). Recently, real world scenes have also been used to model the guidance of attention (Chapter 1; Hwang, Higgins, Pomplun, 2009; Itti & Koch, 2000; Torralba, Oliva, Castelhano, & Henerson, 2006; Zelinsky, 2008).

Before moving your eyes to a region of space, attention is deployed to that region first (Deubel & Schneider, 1996). Two recent developments have been central to improving our understanding of how attention operates in naturalistic scenes: first, the ability to readily record the center gaze with eye tracking technology (e.g. Hayhoe & Ballard, 2005; Henderson, 2003) and, second, advances in computational approaches to scene understanding (e.g. Itti & Koch, 2000; Oliva & Torralba, 2001). Contemporary models of eye guidance in visual search incorporate mechanisms for bottom up feature based guidance, as well as top down task constraints, although models vary in the implementations of each component (Chikkerur, Tan, Serre, & Poggio, 2010; Kannan, Tong, Zhang, & Cottrell, 2009; Torralba et al, 2006).

Laboratory based visual search studies typically instruct participants to localize an object that may be present in the display (or indicate its absence) while recording the time required to make a response. Reaction times (RTs) are considered indicative of the difficulty of search (Duncan & Humphreys, 1998) and are highly correlated with the number of fixations made in the display (Zelinsky & Sheinberg, 1995). RTs in search experiments can be influenced by any processing stage between the retina and the hand (Wolfe, Oliva, Horowitz, Butcher, & Bompas, 2002). Traditionally, the efficiency of visual search has been studied by measuring how steeply RT rises when items are added to the display (Wolfe, 1998). Efficient visual searches have little or no cost of adding items to a search display, for example looking for a red target among blue distractors, while more difficult searches involve longer RTs and steeper slopes when the set size of the display increases (Wolfe, 1998).

Investigating search guidance in realistic scenes represents a challenge for this approach, because set size is a difficult construct to define in a world of overlapping surfaces and objects. Furthermore, a lifetime of experience navigating and searching real world environments is not wasted by the visual system. Experience with different types of scenes- e.g. outdoor places like city streets, indoor places like kitchens- can bias attention based on perceived situational regularities. Early search fixations- even the initial saccade- are direct toward a contextually consistent region of the scene, for example toward horizontal surfaces when looking for a mug or

vertical surfaces when looking for a painting (Torralba et al, 2006). As highlighted above, it is well established that eye fixations will vary, among other things, depending on the object of search and the scene context (Buswell, 1935; Yarbus, 1967). Does eye guidance also depend on past experience with *specific* scene contexts?

Before being studying the role of scene-specific experience, it is important to have a sense how strongly general scene context (e.g. knowing to look for mugs on waist-level surfaces) guides eye movements across different observers and different environments. Chapter 2 investigates eye guidance during search of novel real world scenes. In this experiment, we recorded 14 observers eye movements as they performed a search task (person detection) in 912 outdoor scenes. Interestingly, observers were highly consistent in the regions fixated during search, even when the target was absent from the scene. These eye movements were then used to evaluate computational models of search guidance from three sources: saliency, target features, and scene context. Each of these models independently outperformed a cross-image control in predicting human fixations. Models that *combined* sources of guidance ultimately predicted 94% of human agreement, with the scene context component providing the most explanatory power. None of the models, however, could reach the precision and fidelity of an attentional map defined by human fixations. In addition to providing a benchmark for inter-observer fixation agreement, eye movements from these observers were used in the spatial eye movement analyses of Chapter 3.

Scene specific location priors develop when particular environments are searched repeatedly, thereby associating a scene's identity (e.g. one's own office) with the target's location (or absence). This phenomenon has been studied extensively in the contextual cuing paradigm (Chun & Jiang, 1998) albeit primarily in artificial displays (but see Brockmole, Castelhano, & Henderson, 2006; Brockmole & Henderson 2006a, 2006b; Ehinger & Brockmole, 2008). Search in these studies, however, typically entails looking for a letter-target somewhere within the display. In contrast, everyday search tasks involve looking for objects associated with real world contexts. In chapters 2 and 3, I outline a distinct role for scene-specific experience in guiding the eyes during search for a context-consistent target object (e.g. person in urban scenes, a book in indoor scenes).

In chapter 3, I investigate the role of the person, place, and past in guiding gaze in familiar environments. At the level of eye fixations, it not known whether a person's *specific* search experience influences attentional selection. Eye movements are notoriously variable: people often foveate different places when searching for the same target in the same scene (Mannan, Ruddock, Wooding, 1997). Do individual differences in fixation locations influence how the scene is subsequently examined? Here, I introduce a method, *comparative map analysis*, for analyzing spatial patterns in eye movement data. This analysis was used to quantify the consistency of fixated locations within the same observer and between observers during search of real world scenes. The results of three independent search experiments show a remarkably high degree of similarity in the locations fixated by the same observer across multiple searches of a given scene. Critically, the role of observer-specific guidance was shown to be distinct from other sources of guidance such as global scene context and familiarity with the scene. This is interpreted as evidence for a uniquely informative role of an individual's search experience on attentional guidance in a familiar scene.

The time course of memory retrieval is the topic of chapter 4, in particular, the speed of using scene specific location priors to guide visual search. As suggested above, repeated searching of a specific environment strengthens its representation in memory as the scene's identity becomes increasingly predictive of a target's location. Retrieving these scene specific location priors, however, may not always occur on each instance of search of a familiar environment. To what extent does *time* increase the probability of retrieving and using scene specific memory to guide search? In chapter 4, I introduce an experimental paradigm- the *delayed search* approach- to evaluate speed-accuracy tradeoffs in oculomotor behavior when searching novel and familiar scenes. Two experiments tested the hypothesis that scene specific memory can influence attentional mechanisms in a temporally predictable manner. Specifically, I propose that longer time intervals increase the effectiveness of using memory to help guide attention to the target. Experiment 1 was a people-search task of outdoor scenes in which observers fixated for a typical or an "extended" duration on the scene before making an initial saccade. Interestingly, steeper learning curves and overall lower reaction times resulted when observers were delayed for an extended interval. Experiment 2 was a book-search task of indoor scenes in which observers were delayed for a variable amount of time on any given search trial of novel or familiar scenes. Surprisingly, longer delays improve search performance on both novel and familiar scenes. Eye movement data, however, indicate that longer delays improve the probability of directing overt attention directly to the target. The main conclusion of this chapter is that achieving stronger memory retrieval before initiating search enhances the efficacy of attentional guidance in familiar environments. Overall, this thesis supports the idea that scene specific experience is associated with unique spatial and temporal characteristics that help guide the eyes to informative regions of a familiar scene.

## References

Brockmole J.M., Castelhano, M.S., & Henderson, J.M. (2006). Contextual cueing in naturalistic scenes: Global and local contexts. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 32, 699-706.

Brockmole, J.R., & Henderson, J.M. (2006a). Using real-world scenes as contextual cues for search. *Visual Cognition*, 13, 99-108.

Brockmole, J.R., & Henderson, J.M. (2006b). Recognition and attention guidance during contextual cuing in real-world scenes: Evidence from eye movements. *The Quarterly Journal of Experimental Psychology*, 59, 1177-1187.

Buswell, G. T. (1935). *How people look at pictures*. Oxford: Oxford University Press.

Chikkerur, S., Tan, C., Serre, T., & Poggio, T. (2010). What and Where: A Bayesian Inference Theory of Attention, *Vision Research*, June 16, 2010

Chun, M.M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36,* 28-71

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36,* 1827–1837.

Duncan, J., & Humphreys, G.W. (1989). Visual search and stimulus similarity. *Psychological Review*, 96, 433-458.

Ehinger, K.A., & Brockmole, J.R. (2008). The role of color in visual search in real-world scenes: Evidence from contextual cuing. *Perception & Psychophysics*, 70, 1366-1378.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. Trends in Cognitive Science, 4, 188-194.

Henderson, J. M. (2003). Human gaze control in real-world scene perception. *Trends in Cognitive Sciences*, 7, 498-504.

Hwang, A.D., Higgins, E.C., Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision,* 9, 1-18.

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

Kanan, C., Tong, M.H., Zhang, L., & Cottrell, G.W. (2009). SUN: Top-down saliency using natural statistics. *Visual Cognition, 17*, 979-1003.

Logan, G.D. (1988). Towards an instance theory of automatization. *Psychological review*, 95, 992-527.

Mannan, S., Ruddock, K.H., & Wooding, D.S. (1997). Fixation patterns made during brief examination of two-dimensional images. *Perception*, 26, 1059-1072.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42,* 145-175.

Torralba, A., Oliva, A., Castelhano, M., & 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.

Triesman, A., & Gelade, G. (1980). A feature integration theory of attention. Cognitive Psychology, 12, 97-136.

Wolfe, J.M. (1994). Guided search 2.0. A revised model of visual search. *Psychonomic Bulletin and Review*, 1, 202-228.

Wolfe, J.M. (1998). What do 1,000,000 trials tell us about visual search? *Psychological Science, 9,* 33-39.

Wolfe, J.M., Oliva, A., Horowitz, T.S., Butcher, S.J., & Bompas, A. (2002). Segmentation of objects from backgrounds in visual search tasks. *Vision Research, 42,* 2985–3004.

Yarbus, A. (1967). *Eye movements and vision*. New York: Plenum.

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

Zelinsky, G., & Sheinberg, D. (1995). Why some search tasks take longer than others. *Studies in Information Processing*, 6, 325-336.

CHAPTER 2

# Modeling Search for People in 900 Scenes: A combined source model of eye guidance

Published as:
Ehinger, K.*, 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.

## Introduction

Daily human activities involve a preponderance of visually-guided actions, requiring observers to determine the presence and location of particular objects. How predictable are human search fixations? Can we model the mechanisms that guide visual search? Here, we present a dataset of 45,144 fixations recorded while observers searched 912 real-world scenes and evaluate the extent to which search behavior is (1) consistent across individuals and (2) predicted by computational models of visual search guidance.

Studies of free viewing have found that the regions selected for fixation vary greatly across observers (Andrews & Coppola, 1999; Einhauser, Rutishauser, Koch, 2008; Parkhurst & Neibur, 2003; Tatler, Baddeley, Gilchrist, 2006). However, the effect of behavioral goals on eye movement control has been known since the classic demonstrations by Buswell (1935) and Yarbus (1967) showing that observers' patterns of gaze depended critically on the task. Likewise, a central result emerging from studies of oculomotor behavior during ecological tasks (driving, e.g. Land & Lee, 1994; food preparation, e.g. Hayhoe, Shrivastava, Mruczek, & Pelz, 2003; sports, e.g. Land & McLeod, 2000) is the functional relation of gaze to one's momentary information processing needs (Hayhoe & Ballard, 2005).

In general, specifying a goal can serve as a referent for interpreting internal computations that occur during task execution. Visual search – locating a given target in the environment – is an example of a behavioral goal which produces consistent patterns of eye movements across observers. Figure 1 shows typical fixation patterns of observers searching for pedestrians in natural images. Different observers often fixate remarkably consistent scene regions, suggesting that it is possible to identify reliable, strategic mechanisms underlying visual search and to create computational models that predict human eye fixations.

Various mechanisms have been proposed which may contribute to attention guidance during visual search. Guidance by statistically unexpected, or salient, regions of a natural image has been explored in depth in both modeling and behavioral work (e.g., Bruce, & Tsotsos, 2005; Itti, Koch & Niebur, 1998; Koch & Ullman, 1985; Li, 2002; Rosenholtz, 1999; Torralba, 2003). Numerous studies have shown that regions where the local statistics differ from the background statistics are more likely to attract an observer's gaze: distinctive color, motion, orientation, or size constitute the most common salient attributes, at least in simple displays (for a review, Wolfe & Horowitz, 2004). Guidance by saliency may also contribute to early fixations on complex images (Bruce & Tsotsos, 2005; Harel, Koch & Perona, 2006; Itti & Koch, 2000; Parkhurst, Law & Niebur, 2002; van Zoest, Donk, & Theeuwes, 2004), particularly when the scene context is not informative (Parkhurst et al., 2002; Peters, Iyer, Itti & Koch, 2005) or during free viewing. In natural images, it is interesting to note that objects are typically more salient than their background (Torralba, Oliva, Castelhano, & Henderson, 2006; Elazary & Itti, 2008), so oculomotor guidance processes may use saliency as a heuristic to fixate objects in the scene rather than the background.

In addition to bottom-up guidance by saliency, there is a top-down component to visual attention that is modulated by task. During search, observers can selectively attend to the scene regions most likely to contain the target. In classical search tasks, target features are an

ubiquitous source of guidance (Treisman & Gelade, 1980; Wolfe, Cave & Franzel, 1998; Wolfe, 1994, 2007; Zelinsky, 2008): for example, when observers search for a red target, attention is rapidly deployed towards red objects in the scene. Although a natural object, such as a pedestrian, has no single defining feature, it still has statistically reliable properties (upright form, round head, straight body) that could be selected by visual attention. In fact, there is considerable evidence for target-driven attentional guidance in real world search tasks (Einhauser et al, 2008; Pomplun, 2006; Rao, Zelinsky, Hayhoe, & Ballard, 2002; Rodriguez-Sanchez, Simine & Tsotsos, 2007; Tsotsos, Culhane, Wai, Lai, Davis & Nuflo, 1995; Zelinsky, 2008).

Another top-down component which applies in ecological search tasks is scene context. Statistical regularities of natural scenes provide rich cues to target location and appearance (Eckstein, Drescher & Shimozaki, 2006; Hoiem, Efros, & Hebert, 2006; Torralba & Oliva, 2002, 2003; Oliva & Torralba, 2007). Within a glance, global information can provide useful information about spatial layout and scene category (Joubert, Rousselet, Fize & Fabre-Thorpe, 2007; Greene & Oliva, 2009; Renninger & Malik, 2004; McCotter, Gosselin, Sowden, & Schyns, 2005; Rousselet, Joubert, & Fabre-Thorpe, 2005; Schyns & Oliva, 1994). Categorical scene information informs a viewer of which objects are likely to be in the scene and where (Bar, 2004; Biederman, Mezzanotte, & Rabinowitz, 1982; De Graef, 1990; Friedman, 1979; Henderson, Weeks & Hollingworth, 1999; Loftus & Mackworth, 1978). Furthermore, global features can be extracted quickly enough to influence early search mechanisms and fixations (Castelhano & Henderson, 2007; Chaumon, Drouet & Tallon-Baudry, 2008; Neider & Zelinky, 2006; Torralba et al., 2006; Zelinsky & Schmidt, this issue).

In the present work, we recorded eye movements as observers searched for a target object (a person) in over 900 natural scenes and evaluated the predictive value of several computational models of search. The purpose of this modeling effort was to study search guidance, that is, where observers look while deciding whether a scene contains a target. We modeled three sources of guidance: bottom-up visual saliency, learned visual features of the target's appearance, and a learned relationship between target location and scene context. Informativeness of these models was assessed by comparing the regions selected by each model to human search fixations, particularly in target absent scenes (which provide the most straightforward and rigorous comparison).

The diversity and size of our dataset (14 observers' fixations on 912 urban scenes)1 provides a challenge for computational models of attentional guidance in real world scenes. Intelligent search behavior requires an understanding of scenes, objects and the relationships between them. Although humans perform this task intuitively and efficiently, modeling visual search is challenging from a computational viewpoint. The combined model presented here achieves 94% of human agreement on our database, however a comprehensive understanding of human search behavior environments stands to benefit from mutual interest by cognitive and computer vision scientists alike.

[1] The complete dataset and analysis tools will be made available at the authors' website.

## Experimental Method

*Participants*. Fourteen (14) observers (18-40 years old with normal acuity) were paid for their participation ($15/hour). They gave informed consent and passed the eyetracking calibration test.

*Apparatus*. Eye movements were recorded at 240 Hz using an ISCAN RK-464 video-based eyetracker. Observers sat at 75 cm from the display monitor, 65 cm from the eyetracking camera, with their head centered and stabilized in a headrest. The position of the right eye was tracked and viewing conditions were binocular. Stimuli were presented on a 21" CRT monitor with a resolution of 1024 by 768 pixels and a refresh rate of 100 Hz. Presentation of the stimuli was controlled with Matlab and Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). The following calibration procedure was performed at the beginning of the experiment and repeated following breaks. Participants sequentially fixated 5 static targets positioned at 0° (center) and at 10° of eccentricity. Subsequently, the accuracy of the calibration was tested at each of 9 locations evenly distributed across the screen, including the 5 calibrated locations plus 4 targets at +/- 5.25° horizontally and vertically from center. Estimated fixation position had to be within 0.75° of visual angle for all 9 points, otherwise the experiment halted and the observer was re-calibrated.

*Stimuli*. The scenes consisted of 912 color pictures of urban environments, half containing a pedestrian (target present) and half without (target absent). Images were of resolution 800 by 600 pixels, subtending 23.5 ° by 17.7 ° of visual angle. When present, pedestrians subtended on average 0.9 ° by 1.8 ° (corresponding to roughly 31 by 64 pixels). For the target present images, targets were spatially distributed across the image periphery (target locations ranged from 2.7 ° to 13 ° from the screen center; median eccentricity was 8.6 °), and were located in each quadrant of the screen with approximately equal frequency [2].

*Procedure*. Participants were instructed to decide as quickly as possible whether or not a person was present in the scene. Responses were registered via the keyboard, which terminated the image presentation. Reaction time and eye movements were recorded. The first block consisted of the same 48 images for all participants, and was used as a practice block to verify that the eyes could be tracked accurately. The experiment was composed of 19 blocks of 48 trials each and 50% target prevalence within each block. Eyetracking calibration was checked after each block to ensure tracking accuracy within 0.75 ° of each calibration target. Each participant performed 912 experimental trials, resulting in an experiment duration of 1 hour.

*Eye movement analysis*. Fixations were identified on smoothed eye position data, averaging the raw data over a moving window of 8 data points (33 ms). Beginning and end positions of saccades were detected using an algorithm implementing an acceleration criterion (Araujo, Kowler, & Pavel, 2001). Specifically, the velocity was calculated for two overlapping 17 ms intervals; the onset of the second interval was 4.17 ms after the first. The acceleration threshold was a velocity change of 6 °/s between the two intervals. Saccade onset was defined as the time when acceleration exceeded threshold and the saccade terminated when acceleration dropped

[2] See additional figures on authors' website for distribution of targets and fixations across all images in the database

below threshold. Fixations were defined as the periods between successive saccades. Saccades occurring within 50 ms of each other were considered to be continuous.

## Human Eye Movements Result

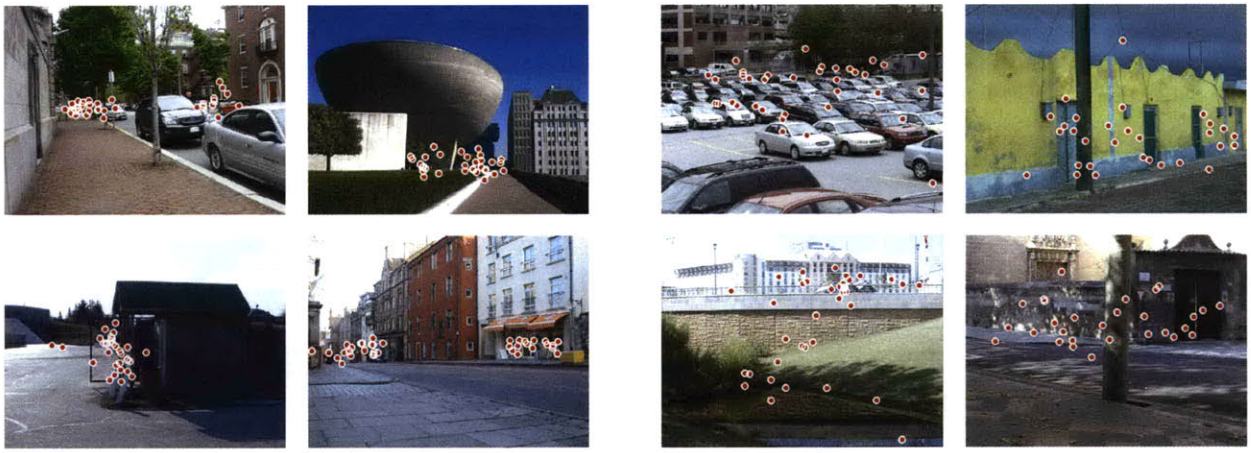
### Accuracy and Eye Movement Statistics

On average, participants' correct responses when the target was present (hits) was 87%. The false alarm rate (fa) in target absent scenes was 3%. On correct trials, observers' mean reaction time was 1050 ms (one standard error of the mean or s.e.m = 18) for target present and 1517 ms (one s.e.m = 14) for target absent. Observers made an average of 3.5 fixations (excluding the initial central fixation but including fixations on the target) in target present scenes and 5.1 fixations in target absent scenes. The duration of "search fixations" exclusively (i.e. exploratory fixations excluding initial central fixation and those landing on the target) averaged 147 ms on target present trials and 225 ms on target absent trials. Observers spent an average of 428 ms fixating the target-person in the image before indicating a response.

We focused our modeling efforts on predicting locations of the first *three* fixations in each scene (but very similar results were obtained when we included all fixations). We introduce below the measures used to compare search model's predictions and humans' fixations.

### Agreement among Observers

How much eye movement variability exists when different observers look at the same image and perform the same task? First, we computed the regularity, or agreement among locations fixated by separate observers (Mannan, Ruddock, Wooding, 1995; Tatler, Baddeley, Gilchrist, 2005). As in Torralba et al (2006), a measure of inter-observer agreement was obtained for each image by using the fixations generated by all-except-one observers. The "observer-defined" image region was created by assigning a value of 1 to each fixated pixel and 0 to all other pixels, then applying a Gaussian blur (cutoff frequency = 8 cycles per image, about 1° visual angle). The observer-defined region was then used to predict fixations of the excluded observer. For each image, this process was iterated for all observers. Thus, this measure reflected how consistently different observers selected similar regions to. Figure 1 shows examples of target absent scenes with high and low values of inter-observer agreement.



(a) High inter-observer agreement (b) Low inter-observer agreement

Figure 1. Examples of target absent scenes with (a) high and (b) low inter-observer agreement. Red dots represent the first 3 fixations from each observer.

Not all of the agreement between observers is driven by the image, however – human fixations exhibit regularities that distinguish them from randomly-selected image locations. In this issue, Tatler & Vincent present compelling evidence that robust oculomotor biases constrain fixation selection independently of visual information or task (see also Tatler 2007). Qualitatively, we observe in our dataset that the corners of the image, and the top and bottom edges, were less frequently fixated than regions near the image center. We therefore derived a measure to quantify the proportion of inter-observer agreement that was independent of the particular scene's content (see also Foulsham & Underwood, 2008; Henderson, Brockmole, Castelhano, Mack, 2007). Our "cross-image control" was obtained using the procedure described above, with the variation that the observer-defined region for one image was used to predict the excluded observer's fixations from a *different* image selected at random.

The Receiver Operating Characteristic (ROC) curves for inter-observer agreement and the cross-image control are shown in Figure 2. These curves show the proportion of fixations that fall within the fixation-defined map (detection rate) in relation to the proportion of the image area selected by the map (false alarm rate). In the following, we report the area under the curve (AUC), which corresponds to the probability that the model will rank an actual fixation location more highly than a non-fixated location, with a value ranging from 0.5 (chance performance) to 1 (perfect performance) (Harel et al, 2006; Renninger, Verghese & Coughlan 2007; Tatler et al 2005).

The results in Figure 2 show a high degree of inter-observer agreement, indicating high consistency in the regions fixated by different observers for both target absent scenes (AUC = 0.93) and target present scenes (AUC = 0.95). Overall, inter-observer agreement was higher in target present than in target absent scenes ($t(805) = 11.6$, $p<0.0001$), most likely because fixating the target was the primary goal of the search. These human agreement curves represent the upper bound on performance, against which the computational models will be compared. Furthermore, the cross-image consistency produced an AUC of 0.68 and 0.62 for target absent and present

scenes respectively (random chance: AUC = 0.5). The cross-image control line represents the proportion of human agreement due to oculomotor biases and other biases in the stimuli set, and serves as the lower bound on the performance of the models.

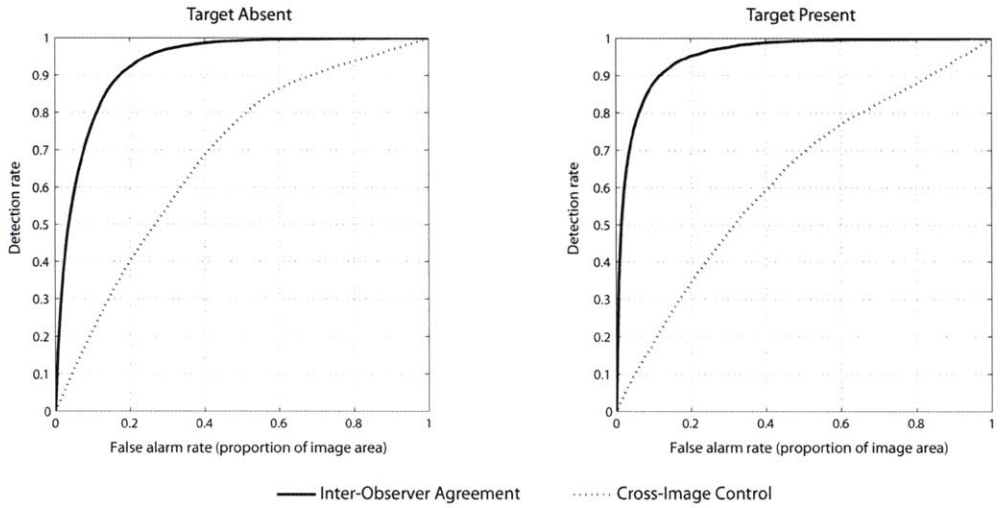



Figure 2. Inter-observer agreement and cross-image control for target absent (left) and present (right) scenes. The false alarm rate, on the x-axis, corresponds to the proportion of the image selected by the model.

## Modeling Methods

Here we used the framework of visual search guidance from Torralba (2003) and Torralba et al (2006). In this framework, the attentional map (M), which will be used to predict the locations fixated by human observers, is computed by combining three sources of information: image saliency at each location ($M_S$), a model of guidance by target features ($M_T$), and a model of guidance by the scene context ($M_C$).

$$M(x,y) = M_S(x,y)^{\gamma 1} M_T(x,y)^{\gamma 2} M_C(x,y)^{\gamma 3} \quad (1)$$

The exponents ($\gamma 1$, $\gamma 2$, $\gamma 3$), which will act like weights if we take the logarithm of Equation 1, are constants which are required when combining distributions with high-dimensional inputs that were independently trained, to ensure that the combined distribution is not dominated by one source (the procedure for selecting the exponents is described below). Together, these three components ($M_S$, $M_T$ and $M_C$) make up the combined attentional map (M).

Figure 3 illustrates a scene with its corresponding saliency, target features, and scene context maps, as well as a combined map integrating the three sources of guidance. Each model makes different predictions, represented as a surface map, of the regions that are likely to be fixated. The best model should capture as many fixations as possible within as finely-constrained a region as possible. In the following sections, we evaluate the performance of each of the three models individually, and then combined models.

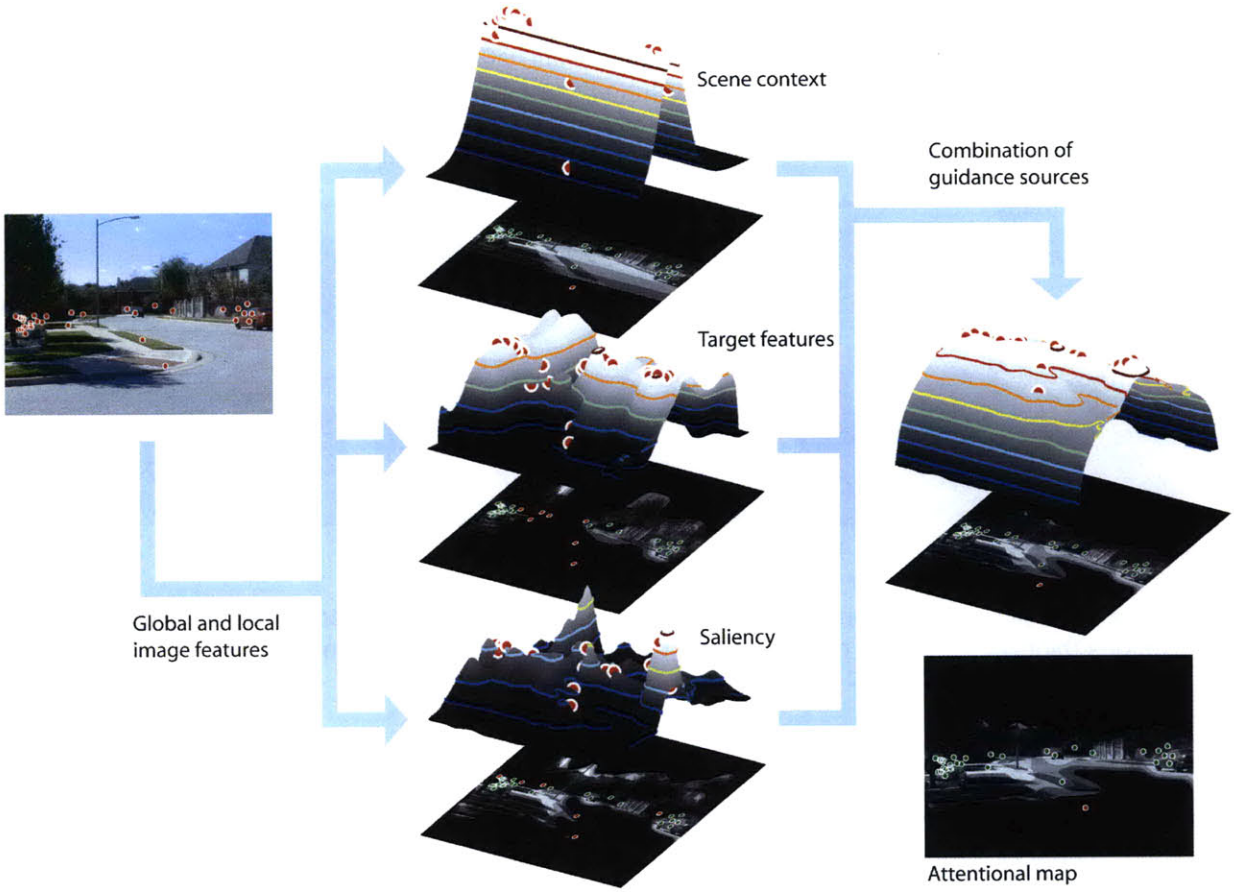



Figure 3: Illustration of an image, the computational maps for three sources of guidance, and the overall, combined attentional map. The flattened maps show the image regions selected when the model is thresholded at 30% of the image. Fixations within the selected region are shown in green.

## Guidance by Saliency

Computational models of saliency are generally based on one principle: they use a mixture of local image features (e.g. color and orientation at various spatial scales) to determine regions that are local outliers given the statistical distribution of features across a larger region of the image. The hypothesis underlying these models is that locations whose properties differ from neighboring regions or the image as a whole are the most informative. Indeed, rare image features in an image are more likely to be diagnostic of objects (Elazary & Itti, 2008; Torralba et al., 2006), whereas repetitive image features or large homogenous regions are unlikely to be object-like (Rosenholtz, Li & Nakano, 2007; Bravo & Farid, 2004).

Computing saliency involves estimating the distribution of local features in the image. Here we used the statistical saliency model described in Torralba et al (2006), including the use of an

independent validation set to determine an appropriate value for the exponent[3]. The independent validation set was composed of 50 target present and 50 target absent scenes selected randomly from the 912 experimental images and excluded from all other analyses. Figure 4 shows maps of the best and worst predictions of the saliency model on our stimuli set.

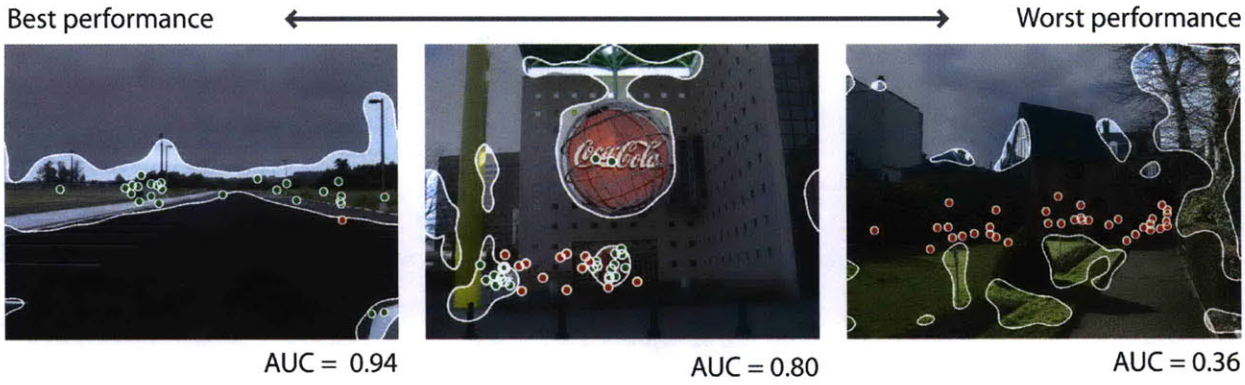



Figure 4. Saliency maps from images with the best, mid-range, and worst performance in predicting fixations (highlighted region represents the highest X% selected by the model).

## Guidance by Target Features

To date, the most ubiquitous source of search guidance are target features (for reviews, Wolfe, 2007; Zelinsky, 2008). Identifying the relevant features of an object's appearance remains a difficult issue, although recent computer vision approaches have reached excellent performance for some object classes (i.e. faces, Ullman, Vidal-Naquet, & Sali, 2002; cars, Papageorgiou & Poggio, 2000; pedestrians, Dalal & Triggs, 2005; cars, bicycles and pedestrians, Serre, Wolf, Bileschi, Riesenhuber & Poggio, 2007; Torralba, Fergus & Freeman, 2008). Here, we used the person detector developed by Dalal & Triggs (2005) and Dalal, Triggs, & Schmid (2006) to model target features, as their code is available online[4], and gives state of the art detection performance at a reasonable speed.

*Implementation of the DT person detector.* The Dalal &Triggs (DT) detector is a classifier-based detector that uses a scanning window approach to explore the image at all locations and scales. The classifier extracts a set of features from that window and applies a linear Support Vector Machine (SVM) to classify the window as belonging to the target or background classes. The features are a grid of Histograms of Oriented Gradients (HOG) descriptors. The detector is sensitive to the gross structure of an upright human figure but relatively tolerant to variation in the pose of the arms and legs. We trained various implementations of the DT detector with different training set sizes and scanning window sizes, but here we report the only the results from the implementation which ultimately gave the best performance on our validation set[5]. This implementation used a scanning window of 32 x 64 pixels and was trained on 2000 upright, unoccluded pedestrians, along with their left-right reflections. Pedestrians were cropped from

[3] In our validation set, the best exponent for the saliency map was 0.025, which is within the optimal range of 0.01-0.3 found by Torralba et al (2006).

[4] See people detector code at http://pascal.inrialpes.fr/soft/olt/

[5] See the authors' website for details and results from the other implementations.

images in the LabelMe database (Russell, Torralba, Murphy, & Freeman, 2008) and reduced in size to fill three-quarters of the height of the detection window. Negative training examples consisted of 30 randomly-selected 32 x 64 pixel patches from 2000 images of outdoor scenes which did not contain people. None of the experimental stimuli were used as training images. The training process was as described in Dalal & Triggs (2005).

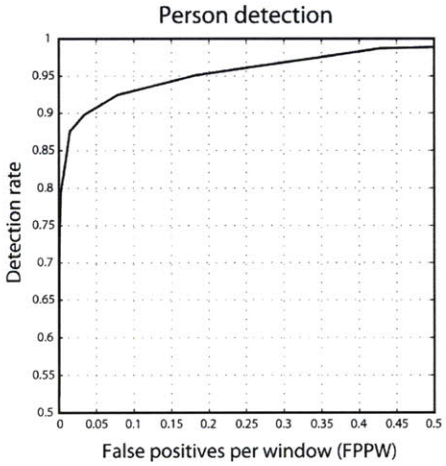



Figure 5. The detection curve of the best implementation of the DT pedestrian detector (trained on 2000 examples of window size 64x32 pixels) on our database of 456 target present images.

The detector was tested on our stimuli set with cropped, resized pedestrians from our target present scenes serving as positive test examples and 32 x 64 pixel windows from our target absent scenes serving as negative test examples. Figure 5 shows the detection performance of our selected DT model implementation[6]. This implementation gave over 90% correct detections at a false positive rate of 10%, confirming the reliability of the DT detector on our database. Although this performance might be considered low given the exceptional performance of the DT detector on other image sets, the scenes used for our search task were particularly challenging: targets were small, often occluded, and embedded in high clutter. It is worth nothing that our goal was not to detect target-people in the dataset, but to use a reliable object detector as a *predictor* of human search fixations.

*Target features map.* To generate target features maps for each image, the detector was run using a sliding window that moved across the image in steps of 8 pixels. Multiscale detection was achieved by iteratively reducing the image by 20% and rerunning the sliding window detector; this process was repeated until the image height was less than the height of the detector window (see Dalal & Triggs, 2005, for details). This meant that each pixel was involved in many detection windows, and therefore the detector returned many values for each pixel. We created the object detector map ($M_T$) by assigning to each pixel the highest detection score returned for that pixel (from any detection window at any scale). As with the saliency map, the resulting object detector map was raised to an exponent (0.025, determined by iteratively varying the exponent to obtain the best performance on the validation set) and then blurred by applying a

[6] See the authors' website for the detection curves of the other model implementations.

Gaussian filter with 50% cut-off frequency at an 8 cycles/image. Figure 6 shows maps of the best and worst predictions of the target features model on our stimuli set.

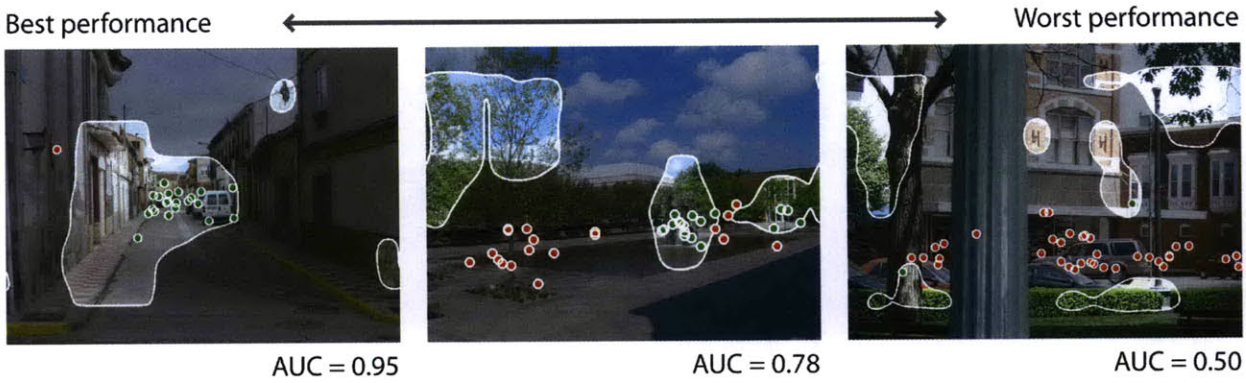



Figure 6. Target features maps from images with the best, mid-range, and worst performance in predicting fixations.

**Guidance by Scene Context Features**

The mandatory role of scene context in object detection and search has been acknowledged for decades (for reviews, Bar, 2004; Chun, 2003; Oliva & Torralba, 2007). However, formal models of scene context guidance face the same problem as models of object appearance: they require knowledge about how humans represent visual scenes. Several models of scene recognition have been proposed in recent years (Bosch, Zisserman, & Muñoz, 2008; Fei-Fei & Perona, 2005; Grossberg, & Huang, in press; Lazebnik, Schmidt, & Ponce, 2006; Oliva & Torralba, 2001; Vogel & Schiele, 2007; Renninger & Malik, 2004), with most of the approaches summarizing an image's "global" features by pooling responses from low-level filters at multiple scales and orientations sampled over regions in the image.

Our model of scene context implements a top-down constraint that selects "relevant" image regions for a search task. Top-down constraints in a people-search task, for example, would select regions corresponding to sidewalks but not sky or trees. As in Oliva and Torralba (2001), we adopted a representation of the image using a set of "global features" that provide a holistic description of the spatial organization of spatial frequencies and orientations in the image. The implementation was identical to the description in Torralba et al (2006), with the exception that the scene context model incorporated a finer spatial analysis (i.e. an 8x8 grid of non-overlapping windows) and was trained on more images (1880 images). From each training image, we produced 10 random crops of 320x240 pixels to generate a training set with a uniform distribution of target locations. As in Torralba et al (2006), the model learned the associations between the global features of an image and the location of the target. The trained computational context model compared the global scene features of a *novel* image with learned global scene features to predict the image region most highly associated with the presence of a pedestrian. This region is represented by a horizontal line at the height predicted by the model. Figure 7 shows maps of the best and worst predictions of the scene context model on our stimuli set.


Figure 7. Scene context maps from images with the best, mid-range, and worst performance in predicting fixations.

There are cases where the scene context model failed to predict human fixations simply because it selected the wrong region (see Figures 7 and 8). In these cases, it would be interesting to see whether performance could be improved by a "context oracle" which knows the true context region in each image. It is possible to approximate contextual "ground truth" for an image by asking observers to indicate the best possible context region in each scene (Droll & Eckstein, 2008). With this information, we can establish an upper bound on the performance of a model based solely on scene context.

*Evaluating the ground truth of Scene Context: a "Context Oracle."* Seven new participants marked the context region for pedestrians in each scene in the database. The instructions were to imagine pedestrians in the most plausible places in the scene and to position a horizontal bar at the height where the heads would be. Participants were encouraged to use cues such as the horizon, the heights of doorways, and the heights of cars and signs in order to make the most accurate estimate of human head height. Image presentation was randomized and self-paced. Each participant's results served as an individual "context model" which identified the contextually relevant location for a pedestrian for each scene. The "context oracle" was created by pooling responses from all observers. Context oracle maps (Figure 8), were created by applying a Gaussian blur to the horizontal line selected by each observer, and then summing the maps produced by all participants.

(a) Scene context model (b) Context oracle

Figure 8: Example of a target-absent image showing its computationally-defined scene context map (left) and an empirically-defined context oracle map (right). Fixations within the top 30% of each model's selected region are shown in green.

**Guidance by a Combined Model of Attention**

The three models were combined by multiplying the weighted maps as shown in Equation 1. The weights ($\gamma 1 = 0.1$, $\gamma 2 = 0.85$, $\gamma 3 = 0.05$) were selected by testing various weights in the range [0,1] to find the combination which gave the best performance on the validation set. Examples of combined source model maps are shown in Figure 9.

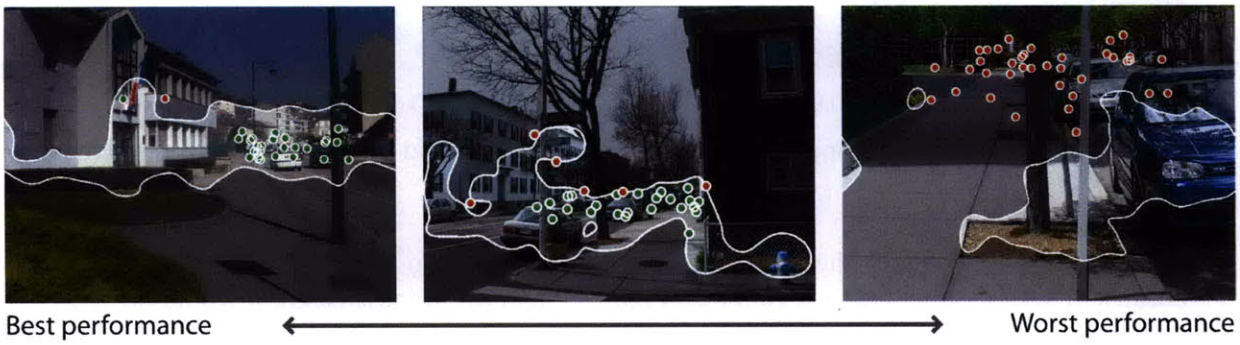



Figure 9. The combined source maps from images with the best, mid-range, and worst performance in predicting fixations.

## Modeling Results

The ROC curves for all models are shown in Figure 10 and the performances are given in Table 1. Averaging across target absent and target present scenes, the scene context model predicted fixated regions with greater accuracy (AUC = 0.845) than models of saliency (0.795) or target features (0.811) alone. A combination of the three sources of guidance, however, resulted in greater overall accuracy (0.895) than any single source model, with the overall highest performance given by a model that integrated saliency and target features with the "context oracle" model of scene context (0.899). Relative to human agreement, the purely computational combined model achieved 94% of the AUC for human agreement in both target present and absent scenes. When the context oracle was substituted for the scene context model, the combined model achieved on average 96% of the AUC of human agreement.

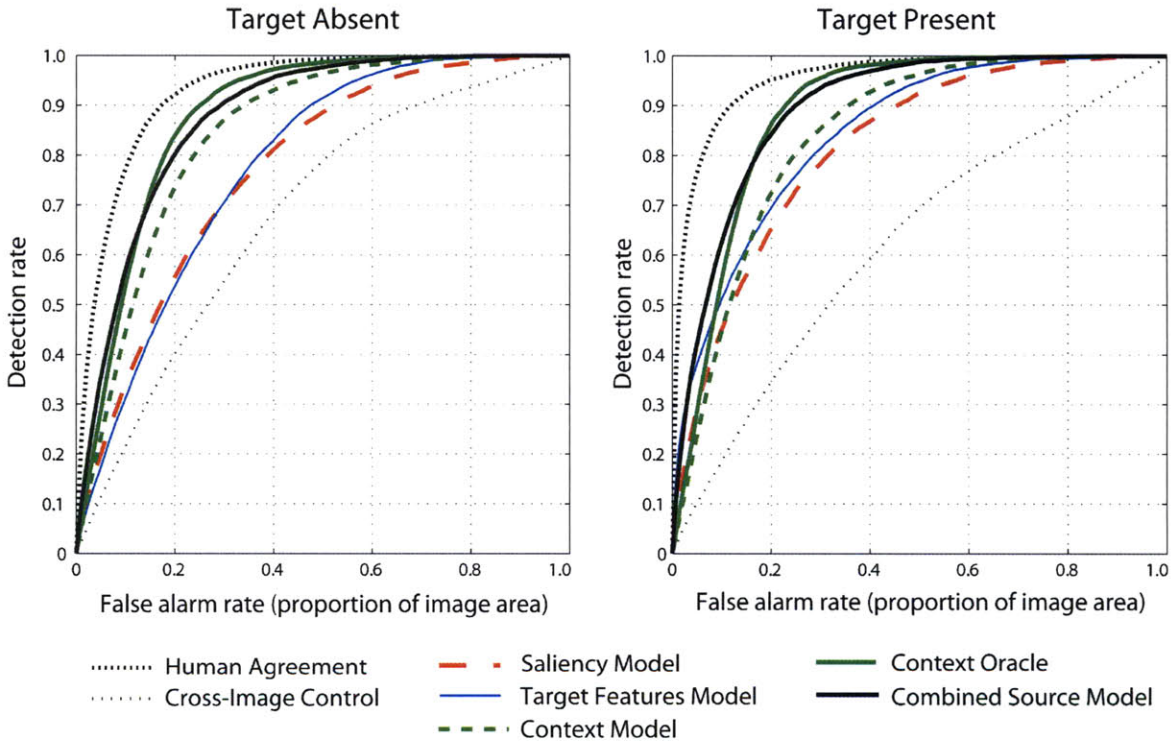



Figure 10: ROC curves for various models. The ROC curves for human agreement and human chance level consistency correspond respectively to the upper bounds and lower bounds of performances against which models will be compared.

### Saliency and target features models

The saliency model had the lowest overall performance, with an AUC of 0.77 and 0.82 in target absent and present scenes. This performance is within the range of values given by other saliency models predicting fixations in free viewing tasks (AUC of 0.727 for Itti et al., 1998; 0.767 for Bruce & Tsotsos, 2006; see also Harel et al., 2005).

The best example shown in Figure 4 is typical of the type of scene in which the saliency model performs very well. The saliency model does best in scenes with large homogenous regions (sky, road), and in which most of the salient features coincide with the region where observers might reasonably expect to find the target. This illustrates the difficulty in determining how saliency influences eye movement guidance: in many cases, the salient regions of a real-world scene are also the most contextually relevant regions. In fact, recent studies suggest that the correlation between saliency and observer's fixation selection may be an artifact of correlations between salience and higher-level information (Einhauser et al, 2008; Foulsham & Underwood, 2008; Henderson et al, 2007; Stirk & Underwood, 2007; Tatler, 2007). The saliency model can also give very poor predictions of human fixations in some scenes, as shown by the example in Figure 4. In a search task, saliency alone is a rather unreliable source of guidance because saliency is often created by an accidental feature (such as a reflection or a differently-colored gap between two objects) that does not necessarily correspond to an informative region.

In target present scenes, not surprisingly, the target features model (AUC = 0.85) performed significantly better than the saliency model ($t(404) = 4.753$, $p < .001$). In target absent scenes, however, the target features model (AUC = 0.78) did not perform significantly above the saliency model ($t(405) < 1$). Interestingly, both models were significantly correlated with each other (0.37, $p < .001$), suggesting that scenes for which the saliency model was able to predict fixations well tended to be scenes in which the target features model also predicted fixations well.

Figure 5 shows target absent images for which the target features model gave the best and worst predictions. Similar to the saliency model, the target model tended to perform best when most of the objects were concentrated within the contextually relevant region for a pedestrian. Also like the saliency model, the target features performed poorly when it selected accidental, non-object features of the image (such as tree branches that happened to overlap in a vaguely human-like shape). It is important to note that the performance of the target features model is not due solely to fixations **on** the target. In the target absent scenes, there was no target to find, yet the target features model was still able to predict human fixations significantly above the level of the cross-image control. Even in target present scenes, replacing predictions of the target features model with the *true* location of the target (a "target oracle") did not explain the target model's performance on this dataset[7].

**Context models**

Overall, scene context was the most accurate single source of guidance in this search task. The computational model of scene context predicted fixation locations with an overall accuracy of 0.85 and 0.84 in target absent and present scenes respectively. The scene context model performed significantly better than the target features model in target absent scenes ($t(405) = 11.122$, $p < .001$), although the two models did not significantly differ in target present scenes ($t(404)<1$).

[7] See the authors' website for a comparison of the ROC curves of the target features model and the target oracle.

In the majority of our scenes, the computational scene context model gave a very good approximation of the location of search fixations. The first and second images in Figure 7 show the model's best and median performance, respectively, for target absent scenes. In fact, the context model failed to predict fixated regions (i.e., had an AUC below the mean AUC of the cross-image control) in only 26 target absent scenes and 24 target present scenes. Typical failures are shown in Figures 7 and 8: in a few scenes, the model incorrectly identifies the relationship between scene layout and probable target location. In order to get around this problem and get a sense of the true predictive power of a context-only model of search guidance, we used the "context oracle." The empirically-determined context oracle should be able to distinguish between cases in which the context model fails because it fails to identify the appropriate context region, and cases in which it fails because human fixations were largely outside the context region.

Overall performance of the context oracle was 0.88 and 0.89 for target absent and target present images respectively. The context oracle performed significantly better than the computational model of scene context in target absent ($t(405) = 8.265$, $p < .001$) and target present ($t(404) = 8.861$, $p < .001$) scenes. Unlike any of the computational models, the context oracle performed above chance on all images of the dataset; at worst, it performed at about the level of the average AUC for the cross-image control (0.68 for target absent scenes). Examples of these failures are shown in Figure 11.

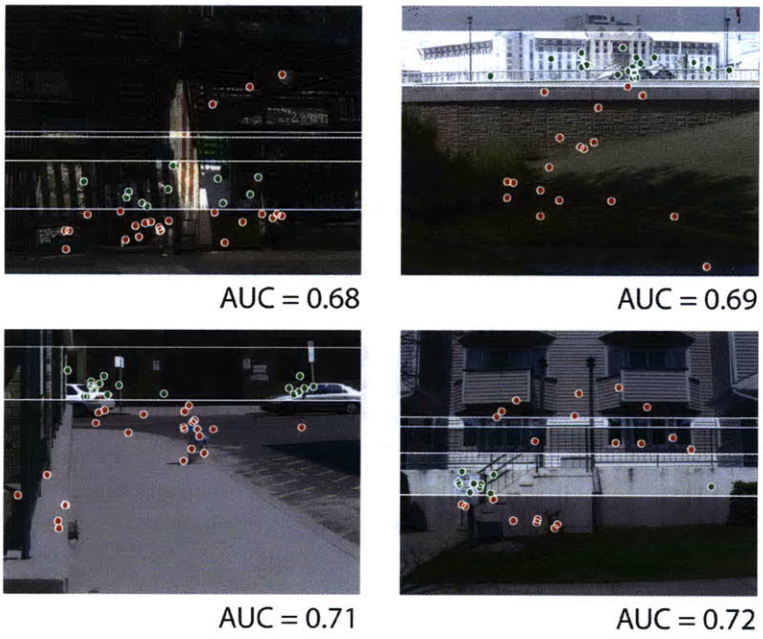



Figure 11: Examples of images for which the context oracle performs poorly.

**Combined source models**

A combined source model that integrated saliency, target features, and scene context outperformed all of the single-source models, with an overall AUC of 0.88 in target absent scenes and 0.90 in target present scenes (see Table 1). The combined guidance model performed better than the best single-source model (scene context) in both target absent ($t(405) = 10.450$, $p < .001$) and target present ($t(404) = 13.501$, $p < .001$) scenes.

Across the image set, performance of the combined model was strongly correlated with that of the scene context model ($r = 0.80$, $p < .001$ in target absent scenes). The combined model was also moderately correlated with the saliency model ($r = 0.51$, $p < .001$ in target absent scenes), and the target features model correlated weakly ($r = 0.25$, $p < 0.001$ in target absent scenes). Taken together, this suggests that the success or failure of the combined model depended largely on the success or failure of its scene context component, and less on the other two components.

In order to analyze the combined model in greater detail, we also tested partial models that were missing one of the three sources of guidance (see Table 1). Removing the saliency component of the combined model produced a small but significant drop in performance in target absent ($t(405) = 6.922$, $p < .001$) and target present ($t(404) = 2.668$, $p < .01$) scenes. Likewise, removing the target features component of the model also produced a small but significant drop in performance in target absent ($t(405) = 5.440$, $p < .001$) and target present ($t(404) = 10.980$, $p < .001$) scenes. The high significance value of these extremely small drops in performance is somewhat deceptive; the reasons for this are addressed in the general discussion. Notably, the largest drop in performance resulted when the scene context component was removed from the combined model (target absent: $t(405) = 17.381$, $p < .001$; target present: $t(404) = 6.759$, $p < .001$).

Interestingly, the combined source model performed very similarly to the empirically-defined context oracle. The difference between these two models was not significant in target absent ($t(405) = -1.233$, $p = .218$) or target present ($t(404) = 2.346$, $p = .019$) scenes.

Finally, the high performance of the context oracle motivated us to substitute it for the scene context component of the combined model, to see whether performance could be boosted even further. Indeed, substituting the context oracle for computational scene context improved performance in both target absent ($t(405) = 5.565$, $p < .001$) and target present ($t(404) = 3.461$, $p = .001$) scenes. The resulting hybrid model was almost entirely driven by the context oracle, as suggested by its very high correlation with the context oracle ($r = 0.97$, $p < .001$ in target absent scenes).

TABLE 1
Summary of performance of human observers, single source models, and combined source of guidance models

| | *Area under curve* | *Performance at 20% threshold* | *Performance at 10% threshold* |
|---|---|---|---|
| Target-absent scenes | | | |
| Human agreement | .930 | .923 | .775 |
| Cross-image control | .683 | .404 | .217 |
| Saliency model | .773 | .558 | .342 |
| Target features model | .778 | .539 | .313 |
| Scene context model | .845 | .738 | .448 |
| Context oracle | .881 | .842 | .547 |
| Saliency × Target features | .814 | .633 | .399 |
| Context × Saliency | .876 | .801 | .570 |
| Context × Target features | .861 | .784 | .493 |
| Combined source model | .877 | .804 | .574 |
| Combined model, using context oracle | .893 | .852 | .605 |
| Target-present scenes | | | |
| Human agreement | .955 | .952 | .880 |
| Cross-image control | .622 | .346 | .186 |
| Saliency model | .818 | .658 | .454 |
| Target features model | .845 | .697 | .515 |
| Scene context model | .844 | .727 | .451 |
| Context oracle | .889 | .867 | .562 |
| Saliency × Target features | .872 | .773 | .586 |
| Context × Saliency | .894 | .840 | .621 |
| Context × Target features | .890 | .824 | .606 |
| Combined source model | .896 | .845 | .629 |
| Combined model, using context oracle | .906 | .886 | .646 |

## General Discussion

We assembled a large dataset of 912 real world scenes and recorded eye movements from observers performing a visual search task. The scene regions fixated were very consistent across different observers, regardless of whether the target was present or absent in the scene. Motivated by the regularity of search behavior, we implemented computational models for several proposed methods of search guidance and evaluated how well these models predicted observers' fixation locations. Over the entire database, the scene context model generated better predictions on target absent scenes (it was the best single map in 276 out of the 406 scenes) than saliency (71 scenes) or target features (59 scenes) models. Even in target present scenes, scene context provided better predictions (191 of 405 scenes) than saliency (72 scenes) but only slightly more than target features (142 scenes). Ultimately, combining models of attentional guidance predicted 94% of human agreement, with the scene context component providing the most explanatory power.

Although the combined model is reasonably accurate at predicting human fixations, there is still room for improvement. Moving forward, even small improvements in model specificity will represent a significant achievement. Our data shows that human observers are reasonable predictors of fixations even as map selectivity increases: 94% and 83% accuracy for selected region sizes of 20% and 10% respectively. In contrast, the accuracy of all models fell off drastically as map selectivity increased and a region size of roughly 40% is needed for the combined source model to achieve the same detection rate as human observers. Figure 12 illustrates this gap between the best computational model and human performance: observers' fixations are tightly clustered in very specific regions, but the model selects a much more general region containing many non-fixated objects. In the following, we offer several approaches that may contribute to an improved representation of search guidance in real-world scenes.

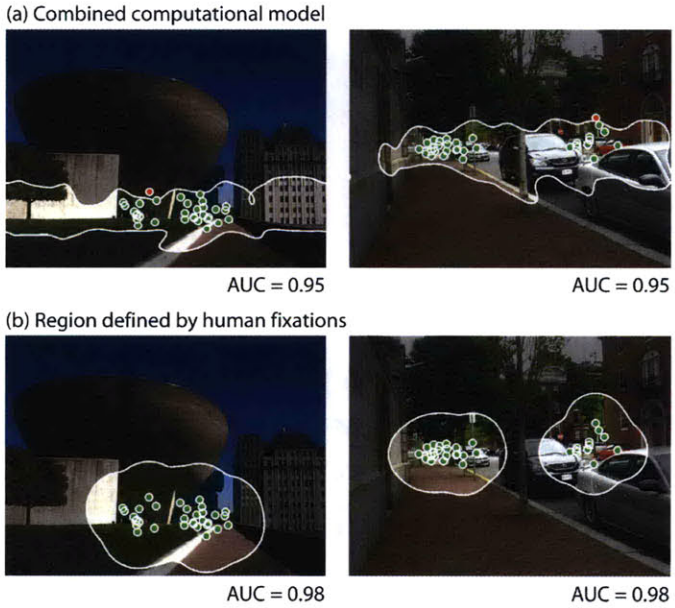



Figure 12. Illustration of the discrepancy between model and human performance. Human fixations are concentrated in a very small region of the image (bottom). The combined guidance model (top) selects this region, but also selects areas that human observers do not fixate

Figure 11 shows the worst performance of the context oracle for target absent scenes. Why was contextual guidance insufficient for predicting the fixated regions of these scenes? One reason may be that our model of the context region did not adequately represent the real context region in certain complex scenes. We modeled the context region as a single height in the image plane, which is appropriate for most images (typically pedestrians appear on the ground plane and nowhere else). However, when the scenes contain multiple surfaces (such as balconies, ramps, and stairs) at different heights, the simplified model tends to fail. Improving the implementation of scene context to reflect that observers have expectations associated with multiple scene regions may reduce the discrepancy between model predictions and where observers look.

In addition, observers may be guided by contextual information beyond what is represented here. It is important to note that scene context can be represented with a number of approaches. Associations between the target and other *objects* in the scene, for example, may also contribute to search guidance (Kumar & Hebert, 2005; Rabinovich, Vedaldi, Galleguillos, Wiewiora & Belongie, 2007; Torralba, Murphy & Freeman, 2004, 2007). In our search task, for example, the presence of a person may be more strongly associated with a doorway than a garbage can. The role of semantic influences in search guidance remains an interesting and open question. In this issue, Zelinsky & Schmidt explore an intermediate between search of semantically meaningful scenes and search in which observers lack expectations of target location. They find evidence that scene segmentation and flexible semantic cues can be used very rapidly to bias search to regions associated with the target (see also Eckstein et al, 2006; Neider & Zelinsky, 2006).

Scene context seems to provide the most accurate predictions in this task, which provokes the question: Is scene context *typically* the dominant source of guidance in real world search tasks? Similarly, how well do the findings of this study generalize to search for other object classes? Our search task may be biased toward context-guided search in the following ways. First, observers may have been biased to adopt a context-based strategy rather than relying on target features simply because the target pedestrians were generally very small (less than 1% of image area) and often occluded, so a search strategy based mainly on target features might have produced more false alarms than detections. Second, the large database tested here represented both semantically-consistent associations (pedestrians were supported by surfaces; Biederman et al, 1982) and location-consistent associations (pedestrians were located on ground surfaces). As a result, even when the target was absent from the scene, viewers expected to find their target within the context region, and therefore the scene context model predicted fixations more effectively than the target-features or saliency models. Searching scenes in which the target location violated these prior expectations (e.g. person on a cloud or rooftop) might bias the pattern of fixations such that emphasis on different sources of guidance is different from the weights on the current model.

A fully generalizeable model of search behavior may need to incorporate flexible weights on the individual sources of search guidance. Consider the example of searching for a pen in an office. Looking for a pen from the doorway may induce strategies based on convenient object relations, such as looking first to a desk, which is both strongly associated with the target and easy to discriminate from background objects. On the other hand, looking for a pen while standing in front of the desk may cause facilitate the use of other strategies, such as searching for

pen-like features. It follows that the features of the target may vary in informativeness as an observer navigates through their environment. A counting task, for example, may enhance the importance of a target features model (Kanan, Tong, Zhang, & Cottrell, 2009). The implications for the combined source model of guidance are that, not only would the model benefit from an improved representation of target features (e.g. Zelinsky, 2008) or saliency (Kanan et al, 2009) or context, but the weights themselves may need to be flexible, depending on constraints not currently modeled.

In short, there is much room for further exploration: we need to investigate a variety of natural scene search tasks in order to fully understand the sources of guidance that drive attention and how they interact. It is important to acknowledge that we have chosen to implement only one of several possible representations of image saliency, target features, or scene context. Therefore, performance of the individual guidance models discussed in this paper may vary with different computational approaches. Our aim, nevertheless, is to set a performance benchmark for how accurately a model representing combined sources of guidance can predict where human observers will fixate during natural search tasks.

## Concluding Remarks

We present a model of search guidance that combines saliency, target features, and scene context, and accounts for 94% of the agreement between human observers searching for targets in over 900 scenes. In this people-search task, the scene context model proves to be the single most important component driving the high performance of the combined source model. None of the models, however, fully capture the selectivity of the observer-defined map. A comprehensive understanding of search behavior may require that future models capture mechanisms that underlie the tight clustering of search fixations.

## Acknowledgments

K.E. was funded by a Singleton graduate research fellowship. B.H.S is funded by a National Science Foundation Graduate Research Fellowship. This work was also funded by an NSF CAREER award (0546262) and a NSF contract (0705677) to A.O., as well as an NSF CAREER award to A.T (0747120). Supplementary information available on the following website: http://cvcl.mit.edu/SearchModels. Correspondence may be sent to any authors: K.E. (ehinger@mit.edu), B.H.S (bhs@mit.edu), A.T (torralba@csail.mit.edu), A.O (oliva@mit.edu).

## References

Andrews, T. J. & Coppola, D. M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments, *Vision Research*, 39, 2947-2953.

Araujo, C., Kowler, E., & Pavel, M. (2001). Eye movements during visual search: The cost of choosing the optimal path. *Vision Research, 41*, 3613-3625.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, 5, 617-629.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14, 143-177.

Bosch, A., Zisserman, A., & Muñoz, X. (2008). Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30, 712-727.

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

Bruce, N. & Tsotsos, J. K. (2006). Saliency Based on Information Maximization. *Advances in Neural Information Processing Systems*, 18, 155-162.

Buswell, G. T. (1935). *How people look at pictures*. Oxford: Oxford University Press.

Castelhano, M. S., & Henderson, J. M. (2007). Initial Scene Representations Facilitate Eye Movement Guidance in Visual Search. *Journal of Experimental Psychology: Human Perception and Performance*, 33, 753-763.

Chaumon, M., Drouet, V., & Tallon-Baudry, C. (2008). Unconscious associative memory affects visual processing before 100 ms. *Journal of Vision*, 8(3):10, 1-10.

Chen, X., & Zelinsky, G. J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research*, 46, 4118-4133.

Chun, M. M. (2003). Scene perception and memory. In D. E. Irwin & B. H. Ross (Eds.) *The psychology of learning and motivation: Advances in research and theory*, Vol. 42 (pp. 79-108). San Diego, CA: Academic Press.

Dalal, N, & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 2, 886-893.

Dalal, N, Triggs, B., & Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. *European Conference on Computer Vision*, 2, 428-441.

De Graef, P., Christiaens, D., & d'Ydevalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, 52, 317-329.

Droll, J., & Eckstein, M. (2008). Expected object position of two hundred fifty observers predicts first fixations of seventy seven separate observers during search, *Journal of Vision*, 8(6), 320.

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional Cues in Real Scenes, Saccadic Targeting and Bayesian Priors. *Psychological Science*, 17, 973-980

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, 8(2):2, 1-19.

Elazary, L., and Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, 8(3):3, 1-15.

Fei Fei, L., & Perona, P. (2005). A Bayesian Hierarchical model for learning natural scene categories. *IEEE Proceedings in Computer Vision and Pattern Recognition*, 2, 524-531.

Fei Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1), 1-29.

Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encoding and memory of gist. *Journal of Experimental Psychology: General*, 108, 316-355.

Grossberg, S., & Huang, T-R. (in press) ARTSCENE: A neural system for natural scene classification. *Journal of Vision.*

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

Hayhoe, M., Shrivastava, A., Mruczek R., & Pelz J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, 3, 49-63.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in Neural Information Processing Systems*, 19, 545-552.

Henderson, J. M. (2003). Human gaze control in real-world scene perception. *Trends in Cognitive Sciences*, 7, 498-504.

Henderson, J. M., Brockmole J. R., Castelhano M. S., & Mack, M. (2007). Visual saliency does not account for eye movement during visual search in real-world scenes. In van R. Gompel, M. Fischer, W. Murray, & R. Hill (Eds.) *Eye Movement Research: Insights into Mind and Brain* (pp. 537-562). Oxford: Elsevier.

Henderson, J. M., Weeks, P. A. Jr., & Hollingworth, A. (1999). Effects of semantic consistency on eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 25, 210-228.

Hoiem, D., Efros, A. A., & Hebert, M. (2006). Putting objects in perspective. *IEEE Conference on Computer Vision and Pattern Recognition*, 2, 2137-2144.

Findlay, J. M. (2004). Eye scanning and visual search. In J. M. Henderson & F. Ferreira (Eds.) *The interface of language, vision and action: Eye movements and the visual world* (pp. 135-150). New York: Psychology Press.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, 8(2):6, 1-17.

Greene, M. R., & Oliva, A. (2009). Recognition of Natural Scenes from Global Properties: Seeing the Forest Without Representing the Trees. *Cognitive Psychology*, 58(2), 137-179.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Vision*, 20(11):12-54.

Itti, L., & Koch, C. (2001). Computational Modeling of Visual Attention. *Nature Reviews Neuroscience*, 2, 194-203.

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

Joubert, O., Rousselet, G., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research*, 47, 3286-3297.

Koch, C., & Ullman, S. (1985). Shifts in visual attention: Towards the underlying circuitry. *Human Neurobiology* 4, 219-227.

Kumar, S. & Hebert, M. (2005). A hierarchical field framework for unified context-based classification. *IEEE International Conference on Computer Vision (ICCV)*, 2, 1284-1291.

Land, M. F. & Lee, D. N. (1994). Where we look when we steer. *Nature*, 369, 742-744.

Land, M. F. & McLeod, P. (2000). From eye movements to actions: How batsmen hit the ball. *Nature Neuroscience*, 3, 1340-1345.

Lazebnik, S., Schmidt, C., & Ponce, J. (2006). Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. *IEEE Conference on Computer Vision and Pattern Recognition*, 2, 2169- 2178.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6(1), 9-16.

Loftus, G. R. & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 565-572.

Mannan, S., Ruddock, K. H., & Wooding, D. S. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. *Spatial Vision*, 9, 363-386.

McCotter, M., Gosselin, F., Sowden, P., & Schyns, P. G. (2005). The use of visual information in natural scenes. *Visual Cognition*, 12, 938-953.

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

Noton, D., & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science*, 171(3968), 308-311.

Oliva, A. & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42, 145-175.

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

Oliva, A. & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12), 520-527.

Papageorgiou, C. & Poggio, T. (2000). A trainable system for object detection. *International Journal of Computer Vision*, 38(1), 15-33.

Parkhurst, D. J. & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision*, 16(2), 125-154.

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

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies, *Spatial Vision*, 10, 437-442.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, 45, 2397–2416.

Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., & Belongie, S. (2007). Objects in Context. *IEEE International Conference on Computer Vision (ICCV)*, 1-8.

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

Rayner, K. (1998) Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372-422.

Renninger, L. W. & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research*, 44, 2301–2311.

Renninger, L. W., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision*, 7(3):6, 1-17

Rodriguez-Sanchez, A. J., Simine, E., & Tsotsos, J. K. (2007). Attention and visual search. *International Journal of Neural Systems*, 17(4), 275-288.

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, 39, 3157-3163.

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, 7(2):17, 1-22.

Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*, 12, 852-877.

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

Schyns, P. G. & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, 5, 195-200.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M. & Poggio, T. (2007). Object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3), 411-426.

Stirk, J. A., & Underwood, G. (2007). Low-level visual saliency does not predict change detection in natural scenes. *Journal of Vision*, 7(10):3, 1-10

Tatler, B.W., (2007). The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4, 1-17.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45(5), 643-659.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, 46(12), 1857-1862.

Turano, K. A., Geruschat, D. R., & Baker, F. H. (2003). Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research*, 43, 333-346.

Torralba, A. (2003a). Modeling global scene factors in attention. *Journal of Optical Society of America A. Special Issue on Bayesian and Statistical Approaches to Vision*, 20(7), 1407-1418

Torralba, A. (2003b). Contextual priming for object detection. *International Journal of Computer Vision*, 53(2), 169-191.

Torralba, A., & Oliva, A. (2002). Depth estimation from image structure. *IEEE Pattern Analysis and. Machine Intelligence*, 24, 1226-1238.

Torralba, A., Fergus, R., & Freeman, W. T. (2008). 80 million tiny images: a large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.30, 1958-1970.

Torralba, A., Murphy, K. P., & Freeman, W. T. (2005). Contextual Models for Object Detection using Boosted Random Fields. *Advances in Neural Information Processing Systems*, 17, 1401-1408.

Torralba, A., Murphy, K. P., & Freeman, W. T. (2007). Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5), 854-869.

Torralba, A., Oliva, A., Castelhano, M., & 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.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, 12, 97-136.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y. H., Davis, N., & Nuflo, F. (1995). Modeling visual-attention via selective tuning. *Artificial Intelligence*, 78, 507-545.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification, *Nature Neuroscience*, 5, 682-687.

van Zoest, W., Donk, M., & Theeuwes, J. (2004). The role of stimulus-driven and goal-driven control in saccadic visual selection. *Journal of Experimental Psychology: Human Perception and Performance*, 30, 746–759.

Viviani, P. (1990). Eye movements in visual search: cognitive, perceptual and motor control aspects. *Reviews Oculomotor Research*, 4, 353-393.

Vogel, J., & Schiele, B. (2007). Semantic scene modeling and retrieval for content-based image retrieval. *International Journal of Computer Vision*, 72(2), 133-157.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, 1, 202-228.

Wolfe, J. M. (2007). Guided Search 4.0: Current Progress with a model of visual search. In W. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 99-119). New York: Oxford Press.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided Search: An alternative to the Feature Integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 419-43

Wolfe, J. M. & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6), 495-501.

Yarbus, A. (1967). *Eye movements and vision*. New York: Plenum.

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

CHAPTER 3

# Search guidance by the person, place, and past: Comparative map analysis reveals an effect of observer specific experience on eye fixations

Partially published:
Hidalgo-Sotelo, B. & Oliva, A. (2010). Person, place, and past influence eye movements during visual search. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

# Introduction

An important feature of ecological visual search is that there are few truly novel, unfamiliar places in which a person is likely to search. Daily tasks often involve examining the same place repeatedly, such as the many occasions spent searching for a specific utensil in one's own kitchen. Locating the target in question benefits from both category based information (e.g. utensils are on countertops) and place specific information (e.g. in *this* kitchen, utensils hang over the stove). Combining these sources of information can happen automatically and often without awareness (Chun & Jiang, 1998). Previous work characterizing attentional deployment during search has neglected the role of a person's past experience with familiar environments. Does prior search experience, assessed using fixation locations, have a distinct role in guiding attention in familiar scenes?

Typically, searching an environment involves making eye movements and directing attentional resources to certain parts of the scene while ignoring others (Findlay & Gilchrist, 2005). This non-uniform sampling of the visual environment has inspired a great deal of interest in identifying what information the visual system uses to select fixation targets within a scene. An early observation by Yarbus (1967) was that the same visual scene (e.g. a painting depicting a scene) was inspected differently, according to the observer's cognitive goal. Later models of oculomotor guidance in natural scenes emphasized low level accounts, such as differences in visual feature content (local contrast, orientation) between fixated and unfixated scene regions (Itti & Koch, 2000; Parkhurst, Law & Niebur, 2002; Reinagel & Zador, 1999). A growing body of evidence suggests that observers *combine* high level information, such as learned target features and scene context, with low level image features to guide their gaze when searching in complex scenes (Chapter 1: Ehinger*, Hidalgo-Sotelo*, Torralba, & Oliva, 2009; Hwang, Higgins, Pomplun, 2009; Torralba, Oliva, Castelhano, Henderson, 2006; Zelinsky, 2008). Among these sources, the use of contextual information is most relevant to the focus of the present study.

Scene context refers to the fact that certain objects are more likely to occur in a particular environment than others: within the context of a kitchen, a whisk is more common than a hairbrush, while the opposite is true in the context of a bedroom. Numerous behavioral studies have documented an interactive role of context for identifying objects in scenes (Biederman, Mezzanotte, & Rabinowitz, 1982; Boyce, Pollatsek, & Rayner, 1989; Davenport & Potter, 2004; Palmer, 1975). Importantly, this contextual knowledge contains information about object *locations* as well as object identities (Gronau, Neta, & Bar, 2008; Oliva & Torralba, 2007). While it is not well understood how context is represented in different tasks, there is substantial evidence that observers use scene context as a heuristic for deciding where to search. These contextual influences, termed *category location priors*, refer to guidance by the spatial layout of a given scene category. A brief glimpse of a real world scene (about 200 ms) provides basic level category and spatial layout information that facilitates attention to relevant scene regions (Castelhano & Henderson, 2007; Henderson, Weeks, & Hollingworth, 1999; Hollingworth, 2009; Sanocki & Epstein, 1997). Even observers' first search fixation is typically directed towards a contextually relevant region (Eckstein, Drescher & Shimozaki, 2006; Neider & Zelinsky, 2006; Torralba et al, 2006), for example inspecting countertop surfaces when searching for a utensil in a kitchen.

Recognition, along with attentional guidance, contributes to the search process when observers search within a familiar context such as a personal office or home. Over time, these places have become associated with certain objects and their characteristic locations. These contextual influences, termed *scene specific location priors*, result from a person's experience of searching within a particular place. There is little evidence of how eye movements are affected by recognizing a familiar scene, and indeed whether memory can be used to guide gaze direction at all. The contextual cuing paradigm has been widely used to study scene specific location priors: when specific scenes are repeated with the target in a consistent location, observers gradually find targets faster in those scenes than in novel scenes, even when repeated scenes cannot be explicitly recognized (Chun & Jiang, 1998, 2003). Few studies have recorded observer's eye movements as they search and learn scene specific associations (Brockmole & Henderson, 2006; Hidalgo-Sotelo, Oliva, Torralba, 2005; Peterson & Kramer, 2001; Tseng & Li, 2004). The most consistent finding is that fewer saccades are made in repeated scenes, while other eye movement parameters, such as average fixation duration, time to initiate search, and saccade amplitude do not systematically change with learning.

Beyond a gross decrease in the number of fixations, does recognition affect the *spatial distribution* of scene regions that are fixated? Interestingly, Peterson & Kramer (2001) showed that searchers occasionally landed their initial search fixation on the target location in the repeated displays, implying that scene recognition occurred rapidly and guided attention directly to the target. More commonly, however, the scene was recognized *after* search had begun yet still resulted in fewer fixations on repeated scenes (Peterson & Kramer, 2001). This result suggests that searching a scene again and again increases memory driven search guidance, but that the strength of recognition is not wholly consistent across trials. As such, the question remains: how systematic is the relationship between scene specific location priors and eye guidance?

Distinguishing "experience based" influences from the myriad of sources that guide attention in natural scenes is tricky for several reasons. One challenge is that attention is strongly guided by information that does *not* depend on scene specific experience. For example, one study recently reported very high fixation agreement across observers, 93%, when participants searched for people in photographs of outdoor scenes, with or without a target (Chapter 1). To illustrate the regularities that exist in eye movements across and within observers, consider the kitchen image in Figure 1 and the different populations of fixations that have been projected onto the scene. Figure 1A shows fixations from 9 observers as they searched for a book in this kitchen scene: the high density of fixations along countertop and cabinet surfaces demonstrates that general scene information such as spatial layout and context guide where observers look. Another challenge in studying eye guidance is that systematic biases unrelated to the scene's content also influence gaze location. In Figure 1B, fixations were randomly sampled from observers' search of *other* scenes and have been projected onto the kitchen scene for illustrative purposes. The non-uniform, central bias in the fixation distribution is driven by oculomotor tendencies (Tatler, 2007; Tatler & Vincent, 2009) and photographer bias (Tseng, Carmi, Cameron, Munoz, & Itti, 2009). The result of these common guidance mechanisms and systematic biases is that observers may fixate similar scene regions, regardless of scene specific experience.

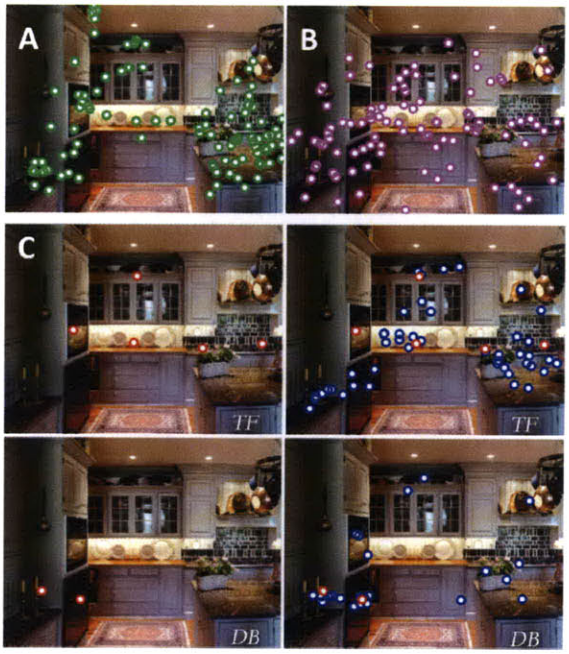



Figure 1: Regularities in eye movements while searching for books (Experiment 1). (A) Fixations from 9 observers searching for a book in this kitchen (green dots). Context and spatial layout constraints of the scene contribute to a non-uniform distribution of fixations, e.g. high fixation density along countertop surfaces in the foreground and background. (B) Fixations sampled from random scenes (not including this kitchen) were projected onto this scene (pink dots). Oculomotor and photographer bias contribute to a tendency for fixations to be centrally distributed, e.g. sparsely fixated top and bottom of the image. (C) Fixations from 2 observers who repeated many searches of this kitchen. Each observer's fixations are shown in 2 panels: Left- Initial search trial (red dots), Right- All subsequent search trials (blue dots). Individual differences in fixation patterns are evident, before and after learning.

Despite overall high agreement across observers, predicting where any one individual will look in a scene is subject to spatial uncertainty. Even when searching for the same object in the same scene, two independent observers may fixate different scene regions, illustrated in Figure 1C with the eye movements of observers *TF* and *DB*. The two panels on the left show each observer's fixations from the *first time* this kitchen scene was searched (red dots). Although each observer inspected different locations, the overall pattern of fixations is consistent with the locations fixated by an independent set of observers (in Figure 1A). The two panels on the right show each observer's fixations from the next seven search trials of this kitchen (blue dots). Interestingly, individual differences in fixation patterns can be seen in the initial search of the scene and they also persist across multiple search trials. This observation suggests that variation in the viewing patterns across observers may partially guide gaze location on subsequent searches of the scene. Such an effect would be masked by pooling over multiple observers **with** scene specific experience (Figure 1C, right panels) because the population as a whole would look similar to the population of observers **without** scene specific experience (Figure 1A). To reasonably estimate the influence of past experience, the search patterns of observers who have never viewed the scene should be contrasted with different observers who have previously searched the scene.

In this paper, we present (1) a novel data analysis approach termed *comparative map analysis* for evaluating how different sources of information contribute to the placement of individual fixation locations, and (2) the eye movement data from three independent experiments investigating the role of scene specific experience in visual search. **The main question underlying this work is whether a person's past experience- as measured by fixated locations- biases attentional selection when searching a familiar scene.** Each search experiment presented here was originally designed to investigate a different issue related to how observers search familiar scenes. Some dependent measures from these experiments have been previously published (Chapter 1; Hidalgo-Sotelo et al, 2005). Spatial characteristics of our participant's search fixations, however, have not previously been reported. Using comparative map analysis on these data sets, we show that eye fixations are guided by a person's past experience of searching in familiar scenes (experiments 1-3), regardless of the presence of a target in the familiar scene (experiments 2-3), and despite an inconsistent learning environment (experiment 3).

## Comparative Map Analysis

The approach we describe as comparative map analysis can be used to evaluate how well different distributions of fixations predict where observers will look in a scene. Critically, each fixation distribution is sampled from a different, strategically chosen, population of fixations. The resulting distributions are evaluated using Receiver Operating Characteristic curves to assess how precisely they distinguish between fixated and unfixated locations. In the present paper, this analysis was used to investigate whether an observer's experience, measured by fixated locations, plays a distinct role in guiding attention during search of familiar scenes. Toward this goal, we broadly study two classes of populations: *scene dependent fixations*, which are driven by the content of that specific scene, and *scene independent fixations*, which provide controls for sources of bias that are unrelated to the scene's content.

### Logic of the approach

Given the challenges outlined in the introduction, how can we isolate the bias resulting from an individual's experience searching a *specific* scene? The solution lies in strategically identifying fixation populations relevant to the question of interest. One population, for example, is defined by the fixation locations of different **novel** searchers of a given scene (i.e. observers who have never searched that particular scene). Another population is defined by the fixation locations of a **single observer** who has searched that particular scene on several instances. Whereas the first population represents the influence of (general) scene context on search, the second population reflects any specific idiosyncrasies of the observer's own examination of the scene. Fixation maps are created by sampling from these populations.

Each fixation map is used to predict the fixation locations of a separate trial, and its accuracy is quantified using an ROC curve. Signal detection measures such as the ROC curve are becoming increasingly used for evaluating eye fixation prediction (Bruce & Tsotsos, 2009; Einhaüser, Spain & Perona, 2008; Renninger, Verghese, Coughlan, 2007; Tatler, Baddeley, and Gilchrist,

2005). This approach is based on the logic that if a map (e.g. an image saliency map) can discriminate between fixated and control locations, then the information in the map is predictive of which locations will be fixated. The map's performance can be summarized by evaluating the area under the ROC curve. Accordingly, if there is no significant difference in the accuracy of two maps, then the underlying information is considered to be equally informative for predicting fixation locations.

Recent studies of attentional guidance have constructed control distributions by randomly sampling from populations of real fixations (e.g. Chapter 1; Renninger et al, 2007; Zhang, Tong, Marks, Shan & Cottrell, 2009). Drawing control distributions from real fixations creates an appropriate baseline condition with the same bias (Tatler et al, 2005). For example, when evaluating whether a saliency model predicted human fixations better than chance, Parkhurst & Niebur (2003) noted that comparison should be against saliency values drawn from actual fixated locations on randomly selected images (see also Henderson, Brockmole, Castelhano, and Mack, 2007). Similarly, Tatler and Vincent (2009) have used this technique to evaluate the predictive power of oculomotor biases.

Comparative map analysis extends this rationale by evaluating several populations that vary with respect to whether information about the "person," "past," and "place" is represented. To specify each population, we consider what information was available to observers who generated the fixations in that population. "Place" information, for instance, refers to how the scene's schema (i.e. spatial layout and semantic knowledge) guides observer's attention. Searching for an object in a new place, like someone else's office or bedroom, recruits this information to deploy attention across the scene. Depending on the task and content of the scene, place information constrains observers eye movements to a varying degree. Looking for a book in a very sparse scene such as a bathroom, for example, may compel gaze to the few object-containing scene regions, while searching within an untidy bedroom would provide fewer gaze constraints. Critically, what we term place information is distinct from information acquired through episodic experience. "Past" information refers to knowledge gained by repeated experience with a particular place; the speeded reaction times observed in contextual cuing (e.g. Chun & Jiang, 1998; Brockmole & Henderson, 2006) reflect the use of this information. Individual observers, however, may differ in where they look on any given search of a familiar or novel scene. "Person" information, therefore, refers to the scene regions fixated by a single observer during their search of a specific scene.

The previous paragraph describes how scene content drives observers to look at certain parts of the scene, which constitutes the class of *scene dependent fixations*. Within this class, we specifically define: (1) fixations made by a single observer's repeated searches, (2) fixations of other familiar observers (i.e. searched the scene repeatedly), and (3) fixations of novel observers (i.e. searched the scene once). Importantly, these populations represent slightly different sources of information: self-consistency, learned scene knowledge and general scene context, respectively.

Control populations are crucial for assessing the relative informativeness of other regularities (e.g. oculomotor biases) in predicting the same eye movements. For this reason, we compare the above populations against *scene independent fixations*, which provide controls for sources of

bias that are unrelated to the scene's content. These populations are defined using: (4) fixations from the same observer on random scenes, and (5) fixations from different observers on random scenes. These populations reflect spatial biases in oculomotor behavior that manifest across the set of scenes (intra-observer and inter-observer biases respectively).

Two simple model-based populations (as opposed to sampling from empirical fixations) serve as controls to evaluate the extent to which a central Gaussian distribution (6) and uniform distribution (7) predict observers' fixations. The uniform distribution serves as the true measure of chance, while the central fixation bias in human eye movements (Tatler, 2007) suggests that a central Gaussian distribution may predict fixations above chance level.

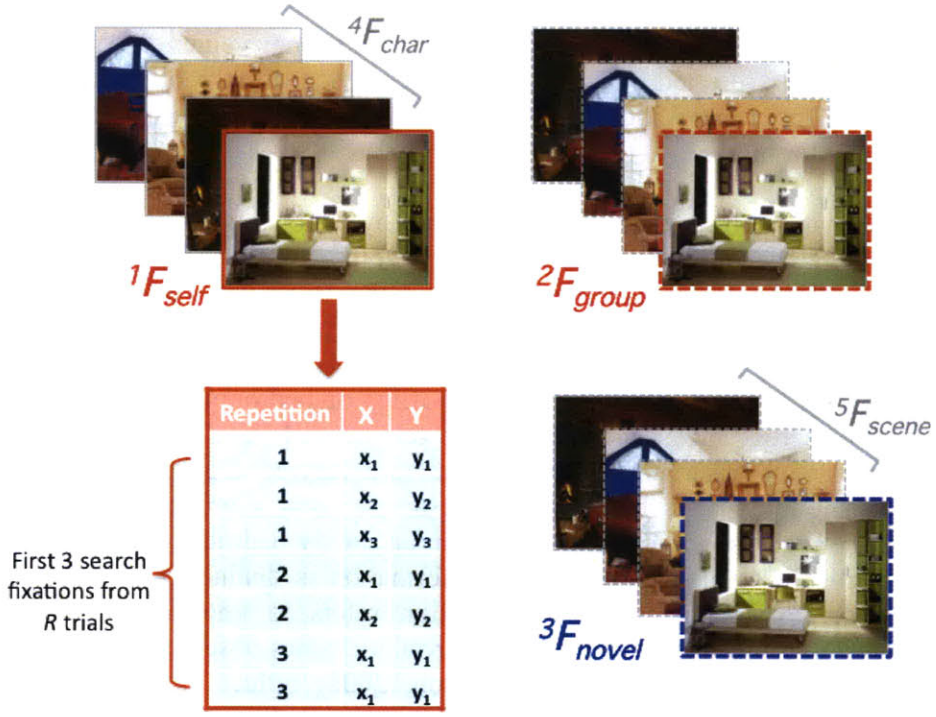



Figure 2: Schematic of comparative map analysis. This illustrates the source of fixation populations (1-5) and how they are sampled to create fixation maps that represent several influences on eye guidance. The following steps are performed iteratively for each of *R* trials: select one search trial (i.e. first 3 fixations of one trial) from $F_{self}$; use the remaining *N* fixations to create a prediction map for intra-observer similarity. Fixation maps for populations (2-5) are created by sampling *N* times from the corresponding distributions. Red (familiar observers) and blue (novel) outlines represent scene dependent populations. Dashed outlines indicate non-self fixation populations.

## Building fixation maps

Fixation maps were created for each of the above populations using the following procedure, shown schematically in Figure 2. First, we collected a list of the locations fixated by one observer in all repeated searches of a scene; trials in which the eye was lost or the observer failed to find the target object were not included. For each repeated search trial R, a self-consistency

fixation map (1) was built by excluding fixations from one search trial and using the remaining N fixations to define a prediction map. Next, the other fixation maps were created by sampling N times from the appropriate population of empirical fixations (2-5) or statistical model (6-7). This process was iterated for R repeated search trials, and the resulting fixation maps were used to predict the excluded trial's fixations (probe fixations).

Except where noted, the first 3 *search* fixations in each trial were used to build the fixation maps in the present analysis. Search fixations are defined as fixations made during active exploration of the scene, thus excluding fixations landing on the target and the initial central fixation. For each repeated search, the maps were compared in terms of how well they predicted the first 3 search fixations of an excluded trial. Given past findings that the consistency of fixation locations across observers decreases over time (Mannan, Ruddock, Wooding, 1997; Yarbus, 1967), we used the first 3 search fixations because it represented a time window appropriate for capturing the highest consistency across novel and repeated conditions.

**Evaluating fixation maps**

We used the Receiver Operator Characteristic to evaluate how well fixated and unfixated locations could be discriminated. The ROC curve is a common signal detection technique that represents the proportion of real fixations falling within a fixation map (detection rate) in relation to the proportion of the image area selected (false alarm rate) (e.g. Chapter 1; Renninger et al, 2007; Tatler et al, 2005). We used the area under the curve or AUC area (Green & Swets, 1966) to compare differences in prediction maps.

## Search Experiment 1

The purpose of this experiment was to investigate whether memory-guided visual search improves as a function of the amount of time to retrieve scene-specific location priors. Observers were given the task of finding a book in indoor scenes (e.g. kitchens, bedrooms) that were either searched once (Novel condition) or searched repeatedly (Familiar condition). Importantly, the book's location was unchanged in repeated presentations of the same scene, allowing observers to learn an association between a specific scene and the location of a book in that scene. Previous findings from our lab (Chapter 3) and others (Kunar, Flusberg, & Wolfe, 2008; Peterson & Kramer, 2001) suggest that a familiar spatial context may indeed guide attention to a learned target location, but that using this form of guidance is slow. We tested this question by introducing a variable stimulus onset asynchrony (SOA) between the *scene onset* (observers fixating centrally) and the *initial search fixation* on the scene.

A similar paradigm was used in a previous search study (Chapter 3, Experiment 1) in which observers were required to keep their eyes on a fixation cross superimposed on the middle of a scene for 1300 ms (long-delay group) or 300 ms (short-delay group) before initiating overt search to find the person in the scene. As in the present experiment, observers searched novel scenes and scenes that were repeated in each of 8 blocks. Our main question was whether prolonging central fixation on a familiar scene would enable participants to use their memory to guide search more effectively. Indeed, we observed that participants in the long-delay group had faster learning curves (RT x block) than participants who were delayed for only 300 seconds. This effect, we hypothesized, was driven by the fact that eye movements can be executed rapidly (latency to initiate the first saccade is typically 250-300 ms, e.g. Castelhano & Rayner, 2008), but a longer delay makes it more likely that scene specific associations will be retrieved from long term memory and used to the eyes to the target.

The present experiment was designed to test this hypothesis by varying the critical conditions, scene familiarity (novel or familiar) and retrieval time (SOA), within-subject. In an initial Learning phase, observers learned scene specific location priors for the repeated scenes. A subsequent Test phase presented each scene briefly (200 ms) followed by a variable SOA (ranging from 0.2 to 1.6 seconds). We predicted that there would be an interaction between scene familiarity and SOA, such that longer delays would predict shorter search times on familiar, but not novel, scenes. Since this variable was tested using a within-subject design, the eye movements from this search study have been collapsed across SOA levels for the purpose of this comparative map analysis.

### METHOD

*Participants.* Twenty two observers, ages 18-34, with normal acuity gave informed consent, passed an eyetracking calibration test, and were paid $15/hr for their participation.

*Materials.* Eye movements were collected using an ISCAN RK-464 video-based eyetracker with a sampling rate of 240 Hz. The stimuli were high resolution color photographs of indoor scenes presented on a 15" LCD monitor with a resolution of 1280 x 1024 px and refresh rate of 60 Hz.

Presentation of the stimuli was controlled with Matlab and Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). The target prevalence in the stimuli set was 100%: all scenes contained a target and, importantly, the target location never changed in a particular scene. To make the task challenging, book targets were small (1 to 2°) and spatially distributed across the periphery.

*Procedure.* The experiment consisted of a Learning phase followed by a Test phase. Observers were instructed, at the beginning of each phase, to find the book in each scene as quickly as possible. The purpose of the Learning phase was for participants to learn the location of a book in scenes that became familiar because they were searched once in each block. The purpose of the Test phase was to manipulate the amount of time between the *scene onset* (observers fixating centrally) and the *initial search fixation* on each scene; participants searched following a variable SOA (200, 400, 800, or 1600 ms) on a novel or familiar scene. Each phase was comprised of 4 search blocks: 24 repeated scenes and 8 novel scenes were randomly intermixed in each block (32 trials per block). Scenes were counterbalanced across observers with respect to the novel or repeated conditions.

The trial sequence was designed to be similar in Learning and Test phases, in order to habituate participants to the procedure of holding their gaze on a fixation cross. As shown in Figure 3, participants fixated a central fixation cross for 500 ms to begin the trial (gaze contingent). Next, the scene was presented with a blue fixation cross superimposed and participants were required to fixate the central cross for the duration of this interval (600 ms or 200 ms, Learning and Test phase respectively) without making a saccade, otherwise trial terminated. In the Learning phase, the fixation cross then briefly turned red (40 ms) and disappeared, signaling participants to actively explored the scene to find the book. In the Test phase, the initial scene presentation (200 ms) was followed by a variable delay on a gray screen, giving an overall SOA (delay plus the initial presentation time) of 200 ms, 400 ms, 800 ms, or 1600 ms; the *same scene* was then presented again and participants moved their eyes to find the target. Participants had a maximum of 8 s to respond via key press (Learning phase) or by fixating the target for 750 ms (Test phase). Feedback was given after each trial (reaction time displayed for 750 ms) to encourage observers to search speedily throughout the experiment. Short mandatory breaks were enforced in order to avoid eye fatigue. The entire experiment lasted approximately 50 min.

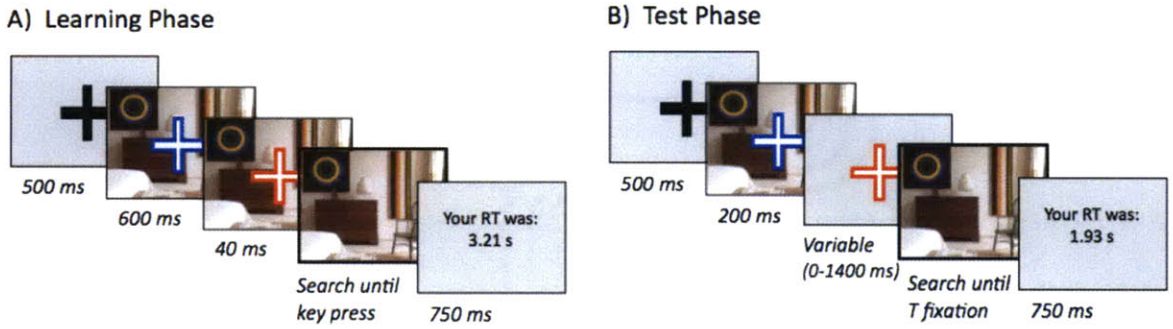



Figure 3. Trial sequence for each phase of Experiment 1. In the Learning Phase, participants learned the location of books in repeated scenes. In the Test Phase, a scene (novel or repeated) was briefly shown, followed by a variable SOA, then participants searched until the target was fixated. The size of the fixation cross is exaggerated for illustration: all fixation crosses were 2°x 2°.

Eyetracker calibration was critical for the gaze contingent aspects of the procedure, as well as to ensure accurate dependent measures (fixation locations). For this reason, calibration was checked at 9 locations evenly distributed across the screen after each search block; fixation position had to be within 0.75° of visual angle for all points, the experiment halted and the observer was recalibrated.

*Eye movement analysis.* Fixations were identified on smoothed eye position data, averaging the raw data over a moving window of eight data points (33 ms). Beginning and end positions of saccades were detected using an algorithm implementing an acceleration criterion (Araujo, Kowler, & Pavel, 2001). Specifically, the velocity was calculated for two overlapping 17 ms intervals; the onset of the second interval was 4.17 ms after the first. The acceleration threshold was a velocity change of 6 deg/s between the two intervals. Saccade onset was defined as the time when acceleration exceeded threshold and the saccade terminated when acceleration dropped below threshold. Fixations were defined as the periods between successive saccades. Saccades occurring within 50 ms of each other were considered to be continuous.

*Comparative map analysis.* As described above, comparative map analysis is based on comparing patterns of gaze within-observer and between-observers as they search a scene. Accordingly, the analysis is performed separately for each scene, and the critical factor is to have many examples of within and between observer variation. In this experiment, a total of 48 scenes were searched by were searched by equal numbers of participants in the novel and repeated conditions: observers in group *A* searched scenes 1-24 repeatedly and searched scenes 25-48 only once, while the other observers in group *B* searched scenes 25-48 repeatedly and scenes 1-24 once. Notably, the design of this experiment meant that observers who generated the fixation populations for repeated search ($F_{self}$ and $F_{group}$) were the same observers who contributed to the fixation populations for novel search ($F_{novel}$). The Learning and Test phases were combined for this comparative map analysis (excluding the first block[8]), yielding a maximum of 7 repeated trials for each observer (N=11 observers per scene). The following experiment conditions correspond to each population: (1) $F_{self}$: one person's repeated searches of a familiar scene, (2) $F_{group}$: other observers' repeated searches of the same familiar scene, (3) $F_{novel}$: different observers' *novel* search of the same scene, (4) $F_{char}$: random scenes searched by the same observer as $F_{self}$, (5) $F_{scene}$: random scenes searched by random novel observers.

[8] The first block is excluded from the Fself and Fgroup distributions because these populations represent the role of experience, and including the first block would provide overlap with Fnovel.

## RESULTS

Behavioral results are summarized in Table 1 and presented in greater detail in Chapter 3. The results of comparative map analysis are shown in Figure 4. Our main finding is that individuals are highly consistent in their pattern of fixations across multiple searches of the same image. A person's own history of where they fixated in a particular scene provides the most accurate prediction of where that individual is likely to look, specifically during the first 3 search fixations in a scene. An identical pattern of results was found when using only the *first* search fixation. Additionally, comparative map analysis replicates previous reports in the literature that scene context (e.g. Neider & Zelinsky, 2006; Torralba et al, 2006) and oculomotor biases (Tatler, 2007; Tatler & Vincent, 2009) play a significant role in guiding the location of gaze. We first report results from the populations based on scene dependent information ($F_{self}$, $F_{group}$, $F_{novel}$), followed by the scene independent control populations.

Two measures of performance are reported. The AUC, or area under the ROC curve, provides a summary of how well each prediction map performed over all thresholds. The performance of each prediction map at a 10% threshold is also reported, which is a single point along the ROC curve which gives a sense of how sensitively- 10% of the image area- each fixation population predicts the observer's fixations. The mean AUC values of all conditions in all three experiments is summarized in Table 2.

**Table 1**

***Summary of behavioral eye movement measures in experiment 1***

| | Repeated Condition | Novel Condition |
|---|---|---|
| Reaction Time | | |
| *M* | 1628 ms | 2401 ms |
| *SE* | *67* | *54* |
| Number of Search Fixations | | |
| *M* | 2.1 | 4.5 |
| *SE* | *0.22* | *0.16* |
| Avg Fixation Duration | | |
| *M* | 176 ms | 212 ms |
| *SE* | *6.6* | *6.9* |
| Avg Saccade Size | | |
| *M* | 7.9 | 7.8 |
| *SE* | *0.18* | *0.18* |

### Role of the person

The role of a person's own search experience was evaluated by using the locations of their own fixations ($F_{self}$) to predict empirical fixations from the same observer on a *separate* search trial of the same image. We found that this population provided the most accurate predictions (mean AUC=0.907) relative to the other scene dependent populations $F_{group}$ ($t(47)=9.04$, $p < 0.001$) and $F_{novel}$ ($t(47)=9.33$, $p < 0.001$), and was significantly higher than all control populations. Furthermore, using a person's own population of fixations to predict where they look (on a separate trial) results in very spatially selective predictions, as evidenced by the steep

$F_{self}$ ROC curve, which indicates high detection rates at low false alarm thresholds. For example, if each fixation distribution was used to select a region representing which 10% of the image was most likely to be fixated, using the $F_{self}$ population yield more accurate predictions than using the $F_{group}$ or $F_{novel}$ populations (mean Hit rates of 0.76, 0.62, and 0.58 at a FA rate of 0.10).

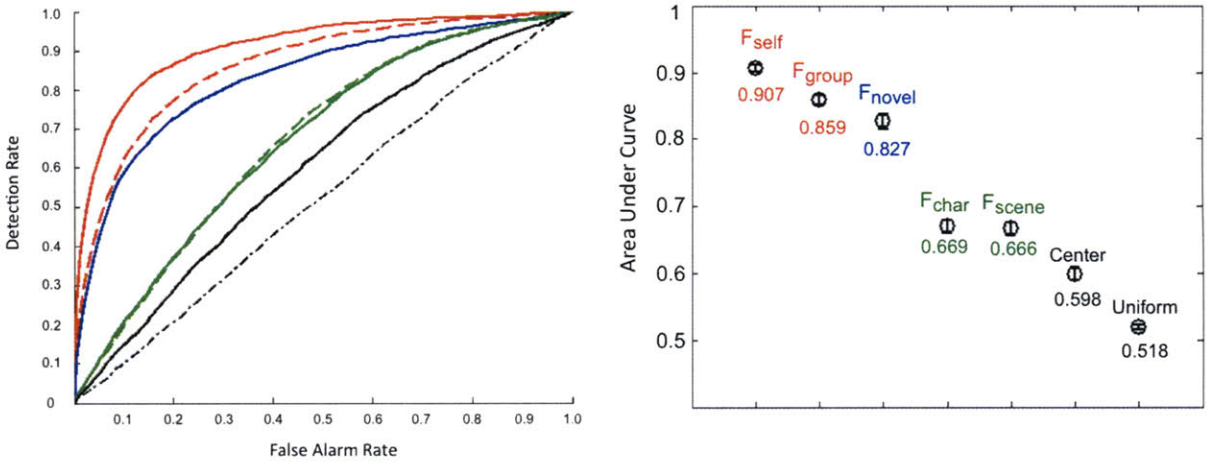



Figure 4. Results of comparative map analysis for Search Experiment 1. The full ROC curve for each fixation distribution is shown on the left (e.g. $F_{novel}$, fixations from novel searchers of a scene, is in blue). The overall performance of each distribution is summarized using the area under the ROC curve (AUC), plotted on the right; error bars represent the standard error of the mean. These data represent the average across scenes.

### Role of the past

The role of past experience was evaluated using the fixation locations from other observers who searched the same scene repeatedly ($F_{group}$). There was a small but significant difference between the prediction accuracy of this group and a group of novel observers (mean AUCs of 0.859 and 0.827, respectively; $t(47)=4.16$, $p < 0.001$). This suggests that sampling from many individuals with past experience may be slightly more informative for predicting where other experienced observers than sampling from the population of novel observers.

### Role of the place

The role of the place is perhaps the most intuitive information that guides where a person looks in a scene: given the context of a scene (e.g. basic-level category, spatial layout), observers wisely direct their eyes to fixate scene regions that are likely to contain the target. The fixation distribution, $F_{novel}$, reflected which scene regions were fixated by novel observers (different participants who searched the scene only once) when looking for a book. As expected, the $F_{novel}$ population provided a significant source of guidance relative to the random scene control $F_{scene}$ (mean AUCs of 0.827 and 0.666, respectively; $t(47)=12.8$, $p < 0.001$). This result, obtained with comparative map analysis, is consistent with other reports in the literature of overall high consistency across observers in search tasks where targets are found in context-consistent locations (e.g. Chapter 1; Torralba et al, 2006).

### Scene Independent Control Populations

Using a random-scene control has become popular in the fixation modeling literature for reasons outlined above in the Comparative Map Analysis section. Here, two control distributions were used: $F_{char}$, a population created by sampling random scene fixations of the *same* observer as those being predicted (also same as $F_{self}$), and $F_{scene}$, a population created by sampling from *different* observers' fixations on random scenes. Although there are purportedly individual differences in eye movement parameters such as saccade length (e.g. Andrews & Coppola, 1999), we believed that using fixations from random scenes to predict the fixations on a **different** scene would be equally well (or poorly) predicted by using one person's fixations ($F_{char}$) as by using different people's fixations ($F_{scene}$). Indeed, these populations performed higher than chance (0.5) at predicting fixations, $F_{char} = 0.669$ and $F_{scene} = 0.666$, and were not significantly different from one another ($t(47)=0.52$). The overlap in these distributions is not surprising given that these populations reflect systematic oculomotor tendencies and regularities in the stimuli set (e.g. photographer bias).

Two model distributions, central gaussian and uniform, were used to compare with the other populations and examine the soundness of the approach of sampling from fixation distributions. Since human fixations have been shown to exhibit a central bias (e.g. Tatler, 2007), we expected a central Gaussian model to predict fixation locations more accurately than a uniform model (chance). Indeed, the central gaussian model was a better predictor of fixations than the uniform distribution (mean AUCs 0.598 and 0.518, $t(47)=8.2$, $p < 0.001$). The fact that the uniform model performance was near the ROC diagonal (AUC 0.5) suggests that the approach of comparative map analysis is reliable.

## INTERPRETATION

These results indicate that people are highly self-consistent in selecting fixation locations across multiple searches of the same scene. How does this finding fit into what is already known about visual search and memory? On one hand, search times are known to reliably improve when a particular spatial configuration predicts the location of a target. On the other hand, individual eye fixations tend to be rapidly executed- about every 250-300 ms (Rayner, 1998)- and although scene familiarity appears to influence the deployment of attention, it is not clear how systematic the relationship is between scene specific location priors and eye guidance. Using comparative map analysis to analyze spatial regularities in observer's search fixations has shown that (1) scene specific memory plays unique a role in guiding attention beyond the influence of general scene context, and beyond that, (2) a person's *own* scene specific search experience influences where they will look when searching a familiar scene.

Is this effect in fact due to a person's *specific* search experience of having fixated certain locations but not others? Perhaps the high degree of within-observer consistency was driven, not by the person's past history of fixations, but by an incidental feature of performing the search task repeatedly (e.g. general familiarity with the scene). For example, if a particular object in a scene effectively attracted fixations across multiple search repetitions, we might expect that one familiar observer's fixations would predict any other familiar observer with equal accuracy. A quick comparison of the red ROC curves in Figure 4 indicates that this is was not the case. The $F_{self}$ population, which was created from a person's own search fixations, was consistently more

accurate than the $F_{group}$ population, sampled from different observer's repeated searches of a scene. To the extent that fixation locations provide a way to operationalize a person's specific scene priors, we interpret this as evidence that the relationship between scene specific memory and eye guidance is systematic and robust.

One concern is that the experimental paradigm itself may have contributed to the observed pattern of results. Given recent findings that guidance by a familiar spatial context is effective but slow (Chapter 3; Kunar et al, 2008; Peterson & Kramer, 2001), the fact that our task involved a central delay[9] may have exaggerated the influence of memory relative to what would be observed during normal conditions of scene viewing. Towards this end, we performed comparative map analysis on from two search experiments, previously published in Hidalgo-Sotelo et al (2005), which were collected with a more traditional search paradigm (no delay or fixation cross overlaid on scenes). If a similar pattern of results is obtained (Role of Person > Role of Past > Role of Place) without a delay introduced between scene onset and search fixations, as in Experiment 1, that would constitute stronger evidence of a unique role of a person's past search experience on attentional guidance.

Furthermore, it is not altogether clear what is driving the high consistency of fixation locations when the same observer searches the same scene repeatedly. Although observers seem to explicitly recognize having seen the repeated scenes (e.g. Brockmole & Henderson, 2006), it is not clear whether observers have conscious access to which scene regions they have searched. In general, individual's memory for specific fixations is thought to be poor (Irwin & Zelinsky, 2002). Our results, since they were found taking the first 3 fixation locations, suggest that the bias from a person's search history is rapidly and unconsciously incorporated into fixation selection mechanisms. What aspect(s) of the task, if any, were critical for eliciting self-consistency in a person's scan patterns over repeated searches? One limitation of our experimental paradigm is that all scenes contained a target. As such, it is not known whether a similar pattern of results would be obtained when a familiar scene did not contain a target.

Performing a comparative map analysis of the following two experiments allowed us to investigate two questions that the data from this experiment could not address: (1) Does within-observer consistency depend on the presence of a target in the familiar scene? and (2) Does this pattern of results change when the familiar scenes are viewed with and without a target (i.e. an inconsistent response) ?

[9] Length of delay duration varied: 640 ms scene presentation in Learning Phase; 200 ms scene presentation followed by 0–1400 ms delay on gray screen in Testing Phase. All delays were enforced on a 2x2 degree fixation cross in the middle of the screen.

## Search Experiment 2

In this experiment (Hidalgo-Sotelo et al, 2005), the task was to look for a person in outdoor urban scenes that became increasingly familiar as participants searched the scenes over many repetitions. Unlike the book search study, this experiment presented scenes with and without a target (person) and observers responded whether a target was present in the scene. Participants learned scene specific location priors, like the previous study, because the target- when present- was always in the same location for a given scene. As depicted in the examples in Figure 5, some scenes were repeatedly presented without a target (e.g. Charles River scene), while different scenes (e.g. neighborhood street scene) always contained a person in a single location. The analyses we previously published addressed how scene specific location priors influenced the overall improvement in reaction time and, specifically, how different stages of visual search (i.e. search initiation, exploration stage, or decision stage) were affected by previous search experience. Here, we use comparative map analysis to address a different question: what spatial regularities exist in the locations fixated by individual observers and how do they compare with the fixation patterns of novel viewers?

This experiment differs from the previous one in that, in this study, observers *only* searched repeated scenes. In other words, the conditions for comparative map analysis were not met with this experiment alone. Fortunately, a control study was run in parallel with this experiment in which an independent set of observers searched the same scenes used in this repeated search study (block 1) followed by many blocks of novel scenes. This eye movement data, first introduced in Chapter 1, allowed us to (a) compare the behavioral results of Hidalgo-Sotelo et al (2005) with a novel scene control, (b) model the human fixations with computational models of search guidance (Chapter 1), and (c) satisfy the criteria of comparative map analysis by providing a population of novel searcher's fixations from which to sample.

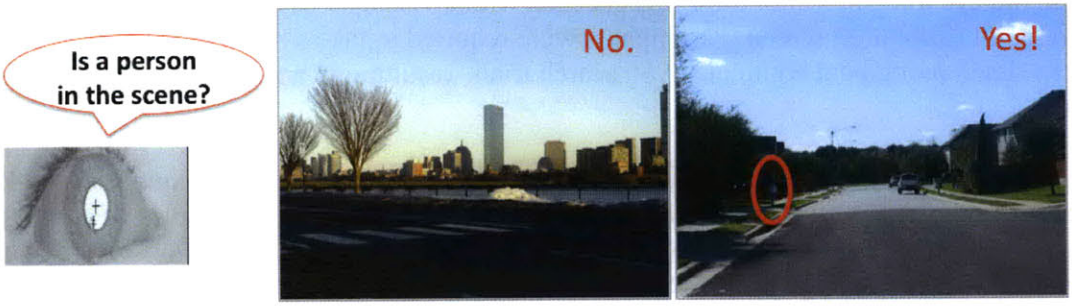



Figure 5. Schematic of the people search task of search experiment 2. Observers searched for a person in the scene. Note that the target person is outlined in red in the figure (not in the experiment).

### METHOD

*Participants.* Twelve observers participated in this repeated search study (Hidalgo-Sotelo et al, 2005). Fourteen observers participated in the novel search study (Chapter 1). All observers were

between the ages of 18-40, with normal acuity gave informed consent, passed an eyetracking calibration test, and were paid $10 for their participation.

*Materials.* A total of 96 color photographs of outdoor scenes comprised the experimental stimuli (primarily street and park scenes in the Boston area). The 96 unique images were collected with the purpose of having 48 scene pairs, in which each photograph taken with a target-person in the scene and without (as shown in Figure 5). All stimuli were presented on a 21" CRT monitor with a resolution of 1024 x 768 px and refresh rate of 60 Hz. The original images were cropped and resized to be presented at a resolution of 800 x 600 px, subtending 23.5 x 17.7 deg of visual angle. Presentation of the stimuli was controlled with Matlab and Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). Eye movements were collected using an ISCAN RK-464 video-based eyetracker with a sampling rate of 240 Hz. Eye movement analysis was performed as described in search experiment 1.

*Procedure.* Participants were instructed to detect the presence (or absence) of a person in a scene and press a corresponding key as soon as a decision had been reached. This search was performed over 20 successive blocks, each block was comprised of 48 scene stimuli (50% target prevalence), and the scenes were presented in randomized order for each block. In the *repeated search study*, the same 48 scenes were searched in each block; these scenes were presented in block 1 of the *novel search study*, but different scenes were searched in the subsequent 19 blocks. As described in the Materials section above, the scene stimuli were originally collected as scene pairs. These pairs were counterbalanced across observers such that an equal number of participants searched the target present and absent versions, but in **this** experiment, *different* participants searched target absent and target present versions of the scene. Framed alternately, there was a 1-to-1 mapping between a scene's identity and the response. Prior to the experiment, participants were required to pass an eyetracking calibration test (as in experiment 1) and performed 10 practice trials to become accustomed to the procedure. Following each search block, the eye tracking calibration was checked with a visual assessment of tracking accuracy on a five point calibration screen. Participants were required to take a small break after every 5 blocks. Each participant completed 960 search trials, resulting in an average experiment duration of 45 minutes.

*Comparative map analysis.* This analysis is performed for each of 96 unique scenes (48 target present, 48 target absent). The critical factor, again, is to have many examples of search patterns generated by the same observer (over multiple searches) and across novel observers. Unlike the previous study, the design of this experiment meant that the twelve observers who generated the fixation populations for repeated search, were *different* than the fourteen observers who generated the novel fixation populations during novel search. Due to stimuli counterbalancing, each scene was searched by six repeated observers and seven novel observers. Notably, there were substantially more blocks in this experiment than in the previous one, 20 blocks and 8 blocks respectively. The results below are derived from all blocks, excluding the first block, in which observers made at least one search fixation. Performing the analysis using only blocks 2-7, as in experiment 1, yields the same the pattern of results.

Recall the source of each population of search fixations: (1) $F_{self}$: one person's repeated searches of a familiar scene, (2) $F_{group}$: other observers' repeated searches of the same familiar

scene, (3) $F_{novel}$: different observers' novel search of the same scene, (4) $F_{char}$: random scenes searched by the same observer as $F_{self}$, (5) $F_{scene}$: random scenes searched by random novel observers.

## RESULTS

The results of comparative map analysis are shown in Figure 6. Note that nine scenes in the target-present condition were excluded from the analysis because they were too easy to require an overt search, meaning that observers tended to saccade directly to the target person after searching the same scene for a few repetitions. The criteria for scene inclusion was that all observers had to have at least 5 search trials in which there was at least 1 search fixation prior to the observer fixating the target (if present).

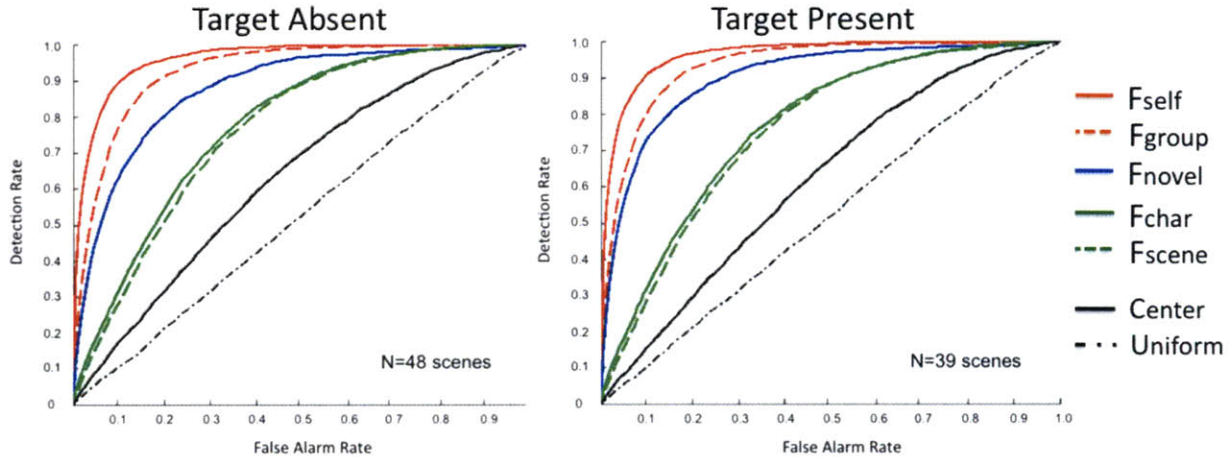



Figure 6. Results of comparative map analysis of the fixations from search experiment 2.

### Role of the person

As in experiment 1, the $F_{self}$ fixation distribution (a measure of within-observer consistency) provided the most useful information for predicting fixation locations on a familiar scene: this was true in target-absent scenes (mean AUC 0.965; $F_{self} > F_{group}$, $t(47)=13.57$, $p < 0.001$; $F_{self} > F_{novel}$, $t(47)=6.79$, $p < 0.001$) and target-present scenes (mean AUC 0.971; $F_{self} > F_{group}$, $t(38)=10.39$, $p < 0.001$; $F_{self} > F_{novel}$, $t(38)=5.64$, $p < 0.001$).

The steepness of the $F_{self}$ ROC curves indicates that this population gave the most spatially precise information (high detection rates at low false alarm thresholds). At a 10% threshold, for example, the accuracy of the $F_{self}$ fixation maps were significantly higher than either the $F_{group}$ or $F_{novel}$ maps (target-absent scenes: mean Hit rates of 0.89, 0.77, and 0.63). Table 2 contains a summary of results for each fixation population by scene condition.

Interestingly, the $F_{self}$ population was as accurate in predicting fixations on the target-absent scenes ($0.965 \pm 0.013$, mean and standard deviation) as in target-present scenes ($0.971 \pm 0.013$). In other words, within-observer consistency was roughly equal in familiar scenes, with or without a target present.

### Role of the past

The $F_{group}$ distribution served as a control for the hypothesis that scene specific memory guides search irrespective of a person's individual search experience. Replicating the pattern of results from experiment 1, the $F_{group}$ distribution was less accurate than $F_{self}$ but was significantly more accurate the $F_{novel}$ population: target-absent scenes ($t(47)=6.79$, $p < 0.001$) and target-present scenes ($t(38)=5.63$, $p < 0.001$). There was not a significant difference between accuracy of the $F_{group}$ population on scenes with or without a target ($t(84)=1.70$).

### Role of the place

The fixation distribution, $F_{novel}$, represented how well category location priors (i.e. scene context in the absence of scene specific experience) guided attention to particular scene regions when searching for a person. As expected, the $F_{novel}$ population provided a significant source of guidance relative to a random scene control: target-absent scenes (mean AUC 0.861; $F_{novel} > F_{scene}$, $t(47)=6.57$, $p < 0.001$) and target-present scenes (mean AUC 0.875; $F_{novel} > F_{scene}$, $t(38)=9.26$, $p < 0.001$). There was no significant difference between accuracy of the $F_{novel}$ population on scenes with or without a target ($t(84)=0.81$).

### Scene Independent Control Populations

Overall, results from the scene independent control populations were similar to those from experiment 1. As evident from Figure 6, fixations from random scenes (green lines) predicted search fixations on a specific scene with lower accuracy than the scene-dependent populations, but higher than would be expected by chance (dotted black line). This was true for target-absent scenes (mean AUCs 0.777 and 0.767, $F_{char}$ and $F_{scene}$) and target-present scenes (mean AUCs 0.741 and 0.729). Unlike experiment 1, there was a trend for an observer-specific guidance effect, independent of scene content: in other words, the $F_{char}$ distribution was slightly more accurate than the $F_{scene}$ distribution on target-absent scenes ($t(47)=2.10$, $p < 0.05$), however the trend did not reach significance on the target present scenes ($t(38)=1.83$). Although we expected these distributions to overlap almost entirely, this trend is consistent with the idea of individual differences in some eye movement parameters (e.g. Andrews & Coppola, 1999; Boot, Beic, & Kramer, 2009); also, the effect is small when compared to the magnitude of guidance by scene dependent information.

Two model distributions, central Gaussian and uniform, are useful as points of reference and for comparing the accuracy of populations of empirical fixations. In both conditions, a central Gaussian model (solid black line in Figure 6) was a better predictor of fixations than a uniform model (dotted black line), but less accurate than the random scene controls (green lines). Interestingly, the central Gaussian model predicted fixations on target-absent scenes, on the whole, more effectively than it predicted fixations on target-present scenes ($t(84)= 3.23$, $p < 0.002$). This suggests that the population of target-absent fixations were, on the whole, more centrally distributed than target-present fixations. Given that observers in this experiment never saw a target in those scenes, it is understandable that observers learned, over time, that the target-absent scenes never contained a target and that their fixations may have been more centrally distributed as a result.

## INTERPRETATION

The results of this people search experiment replicated the main findings of experiment 1 with different scenes, different observers, and, importantly, an experimental paradigm that did not involve a delay prior to making saccades. Performing comparative map analysis of the eye movement data from novel searchers (Chapter 1) and repeated searchers (Hidalgo-Sotelo et al, 2005) allowed us to estimate the influence of each of the following sources on eye guidance (least-to-most influential): oculomotor biases, category location priors, scene specific location priors, and *person specific*, scene specific location priors. Our results showed that (1) pooling across the scene specific experiences of many observers provided significant source of guidance beyond that provided by the scene's context, and (2) a person's own scene specific search experience also plays a unique role in eye guidance. This pattern of results was observed on familiar scenes which **always** contained a target (consistent location across repetitions) and on different but familiar scenes which **never** a target. What does this suggest about the underlying basis of within-observer consistency of fixation locations?

First of all, our data further support the idea that these influences are exerted rapidly. The difficulty of this people search task was easier than the previous book search experiment, meaning that, over time, fewer search fixations were made before responding. Nevertheless, these spatial regularities are evident within the first fixation(s) on the scene. Categorical scene priors have been shown to influence the first search fixation on the scene (e.g. Eckstein et al, 2006; Torralba et al, 2006). These data suggest that even more specific memory representations can be rapidly incorporated into fixation selection mechanisms.

Secondly, the magnitude of guidance by the "Person" ($F_{self}$) and the "Past" ($F_{group}$) was approximately equal in familiar target-present and target-absent scenes. This suggests that the conditions for eliciting consistency in a person's fixations (over repeated searches of a scene) did not require the presence of a target. The identity of the scene itself apparently was sufficient to retrieve memory-based information to help guide attention to the target. Thus far, however, the scene's identity has been perfectly correlated with the response. All the scenes in experiment 1 contained a book-target and, in experiment 2, the familiar target-present scenes were always *different* scenes that the familiar target-absent scenes. Will scene specific experience guide the placement of fixations if observers learn that a given scene may or may not contain a target (consistently-located within the scene)?

Experiment 3 enabled us to investigate whether our main finding (Role of Person > Role of Past > Role of Place) would be replicated with an independent set of observers and, furthermore, whether guidance by scene specific experience was still evident when the same observer searched target-present *and* target-absent versions of a given scene (i.e. scene's identity dissociated from a consistent response).

## Search Experiment 3

In this experiment (Hidalgo-Sotelo et al, 2005), as in the previous one, the task was to look for a person in outdoor urban scenes that became increasingly familiar as participants searched the scenes over many repetitions. Critically, unlike the previous experiment, observers viewed the same scene (e.g. Charles River scene) in versions that indicated an inconsistent response: sometimes the target was present and sometimes it was absent (50% prevalence). This is depicted schematically in Figure 6. Again, the target- when present- was always in the same location for a given scene. Here, we use comparative map analysis to address a different question: what spatial regularities exist in the locations fixated by individual observers and how do they compare with the fixation patterns of novel viewers? As in experiment 2, the novel scenes condition is taken the data in Chapter 1 and allowed us to satisfy the criteria of comparative map analysis by providing a population of novel searcher's fixations from which to sample.

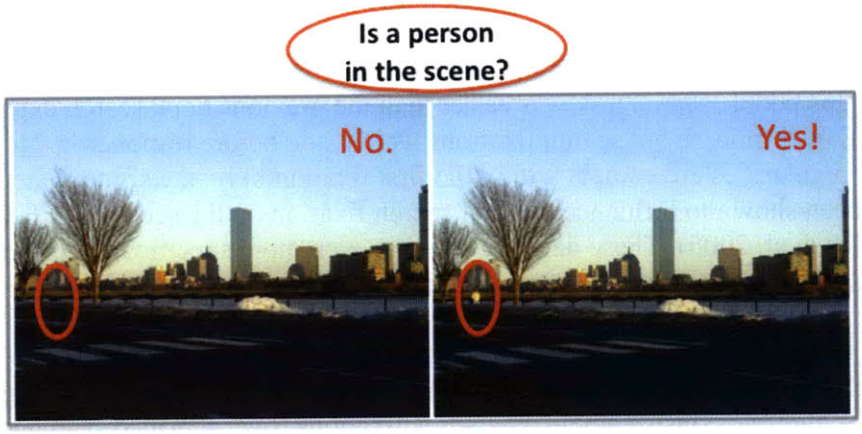



Figure 7. Schematic of the people search task of search experiment 3.

### METHOD

*Participants.* Twelve observers participated in this repeated search study (Hidalgo-Sotelo et al, 2005). None of the same observers participated in experiments 2 and 3. Again, fourteen observers participated in the novel search study (Chapter 1). All observers were between the ages of 18-40, with normal acuity gave informed consent, passed an eyetracking calibration test, and were paid $10 for their participation.

*Materials.* Same as in experiment 2.

*Procedure.* The procedure was identical to that described in experiment 2, with the exception that, in this experiment, the *same* participant searched both target absent and target present versions of the scene. Each scene, therefore, was associated with both types of responses.

*Comparative map analysis.* Same as in experiment 2.

## RESULTS

The results of this people search experiment are shown in Table 2 along with the results from the previous 2 experiments. Overall, the same pattern of results is replicated in this experiment.

### Role of the person

The $F_{self}$ fixation distribution was again the most accurate map for predicting fixation locations on a familiar target-absent scenes (mean AUC =0.953; $F_{self} > F_{group}$, $t(47)=16.73$, $p < 0.001$; $F_{self} > F_{novel}$, $t(47)=13.93$, $p < 0.001$) and familiar target-present scenes (mean AUC =0.954; $F_{self} > F_{group}$, $t(38)=11.56$, $p < 0.001$; $F_{self} > F_{novel}$, $t(38)=7.95$, $p < 0.001$). In both conditions, this distribution created the most spatially precise maps, as evident by the steeply rising $F_{self}$ ROC curves, implying high detection rates over a range of false alarm thresholds.

As in experiment 2, the accuracy of the $F_{self}$ population in predicting fixations on the target-absent scenes not significantly different than in target-present scenes ($t(84)=0.23$). Thus, within-observer consistency was not evidently influenced by dissociating the scene's identity from the response.

### Role of the past

The $F_{group}$ distribution was again less accurate than $F_{self}$ but significantly more accurate the $F_{novel}$ population: target-absent scenes ($t(47)=6.66$, $p < 0.001$) and target-present scenes ($t(38)=2.50$, $p < 0.05$). There was no significant difference between accuracy of the $F_{group}$ population on target-absent and target-present scenes ($t(84)=0.09$).

### Role of the place

The $F_{novel}$ population provided a significant source of guidance relative to a random scene control: target-absent scenes (mean AUC 0.894; $F_{novel} > F_{scene}$, $t(47)=13.13$, $p < 0.001$) and target-present scenes (mean AUC 0.911; $F_{novel} > F_{scene}$, $t(38)=10.21$, $p < 0.001$). There was no significant difference between accuracy of the $F_{novel}$ population on target-absent and target-present scenes ($t(84)=1.83$), however the trend was for the target-present $F_{novel}$ maps to be slightly more predictive than the target-absent $F_{novel}$ maps.

### Scene Independent Control Populations

Overall, the scene independent control populations gave very similar to those from experiment 2. Firstly, the random scene controls were less accurate than scene-dependent maps but higher than would be expected by chance: target-absent scenes (mean AUCs 0.748 and 0.740, $F_{char}$ and $F_{scene}$), target-present scenes (mean AUCs 0.777 and 0.751).

Secondly, we found a small trend for the random scene fixations from one person ($F_{char}$) to be slightly more accurate than pooling over many observers' random scene fixations ($F_{scene}$): on target-absent scenes, the trend did not reach significance ($t(47)=1.83$), but was significant on target-present scenes ($t(38)=3.45$, $p < 0.001$). Since this weak trend was observed in experiment 2 but not experiment 3, the effect of person (independent of scene) is likely small when compared with guidance from scene-dependent information.

As expected, the central Gaussian model was a better predictor of fixations than a uniform model, but less accurate than the random scene controls. In the previous search experiment, we had found that target-absent fixations were more centrally distributed than fixations on target-present scenes. In this experiment, however, the central Gaussian model was an equally good predictor of fixations target-absent fixations or target-present scenes ($t(84)=0.80$). This finding is

entirely understandable given the nature of this task in this experiment, specifically that the same observer searched target-present *and* target-absent versions of familiar scenes.

Table 2

*Summary of Comparative Map Analysis Results*

| | **Experiment 1** | **Experiment 2:** | | **Experiment 3:** | |
|---|---|---|---|---|---|
| | *Target Present* | *Target Absent* | *Target Present* | *Target Absent* | *Target Present* |
| $F_{self}$ | 0.907 | 0.965 | 0.971 | 0.953 | 0.954 |
| $F_{group}$ | 0.859 | 0.933 | 0.941 | 0.922 | 0.923 |
| $F_{novel}$ | 0.827 | 0.861 | 0.875 | 0.894 | 0.911 |
| $F_{char}$ | 0.669 | 0.777 | 0.741 | 0.748 | 0.777 |
| $F_{scene}$ | 0.666 | 0.767 | 0.729 | 0.740 | 0.751 |
| *Center* | 0.589 | 0.623 | 0.600 | 0.608 | 0.599 |

## INTERPRETATION

This experimental paradigm, similar to experiment 2, did not enforce a delay before observers moved their eyes to search. Nevertheless, the results replicated our main finding- namely that gaze is guided by scene context, past experience with a specific scene, **and** person specific, scene specific experience. The critical experimental manipulation was that participants searched both target-present and target-absent versions of a scene, therefore a scene's identity did not predict the trial's response. Consistent with the results of experiment 2, the $F_{self}$ maps (person specific, scene specific fixations) were as accurate for target-present scenes as the $F_{self}$ maps for target-absent scenes; the same was true for the $F_{group}$ population (pooling observers' scene specific fixations). This implies that mapping between a scene's identity and trial response was not the underlying reason for the observed pattern of results.

## General Discussion

Spatial regularities in eye fixations were assessed using a novel approach, *comparative map analysis*, which is based on comparing fixation distributions that reflect categorical and scene specific location priors, across and within observers. Comparative map analysis produced a highly consistent pattern of findings in three search experiments. Overall, these results are interpreted as evidence that scene specific experience biases attentional selection during the first 3 search fixations in a scene.

In experiment 1, observers performed a difficult visual search for a book in novel and familiar indoor scenes that always contained a target; notably, the target's location was consistent across scene repetitions (e.g. Chun & Jiang, 1999). We showed that **systematic individual differences** in fixation patterns persist across multiple searches of a scene. This finding was replicated and extended in experiment 2, in which observers searched for a person in outdoor scenes with 50% target prevalence: half of the repeated scenes always contained a target (directly comparable with experiment 1) and the other half were different scenes that never contained a target. To exclude the possibility that a consistent response mapping (between a scene and the response on all trials with that scene) contributed to these results, in experiment 3 the same scene (near identical versions of the same place) could be either target present or absent on different trials. Again, systematic individual differences in fixation patterns were found in both target absent and target present repeated scenes. Over all three experiments, comparative map analysis provides converging evidence that (1) scene specific experience, pooled across individuals, affects gaze deployment beyond what would be expected from the scene's context alone and, (2) a person's own search experience influences where they look in familiar scenes.

This main idea of this chapter is that, in a specific sense, the past repeats itself: a person's experience, as indexed by which scene locations were fixated, influences how they search familiar scenes. Although the presence of idiosyncratic gaze patterns has previously been reported (e.g. Noton & Stark, 1971), this is the first time that systematic individual differences in fixation patterns have been shown in a naturalistic search task. What is the nature of the information that underlies within-observer consistency? Is it behaviorally relevant or an incidental consequence of scene exposure? In this general discussion, I address how this finding contributes to a better understanding of the relationship between eye movements and memory.

In what respect did the task contribute to the pattern of results we observed? The fact that our instructions gave observers a specific goal- to look for a book, for example- suggests a different pattern of viewing may have been evoked if observers had been asked to memorize the scene or to simply look around (Buswell, 1935; Yarbus, 1967; Castelhano, Mack, Henderson, 2009). Visual search paradigms have made it possible to study and model attentional processes (e.g. Wolfe, 1994) by providing a setting that controls for certain stimulus and goal driven factors. Knowing the location of gaze reveals the information selected for foveal processing during execution of the search task. Oculomotor responses arise from a variety of interacting factors, however, many of which are not task dependent. The approach in this paper has been to compare gaze patterns between populations of observers who vary in the degree of scene and memory dependent information guiding eye movements. All three experiments in this chapter tested a

condition in which a scene's identity was wholly consistent with the location of the target (if present). Comparative map analysis allowed us to evaluate how scene specific experience influenced fixation selection, as compared to guidance by scene context information alone. Still, it is not clear whether similar results would have been obtained outside of the context of a visual search task.

Two lines of evidence suggest that one of the chapter's main findings- that a person's own history with a scene biases attentional selection- is not directly driven by our task's demands. The first argument appeals to the fact that the $F_{self}$ and $F_{group}$ fixation distributions were derived from observers under precisely the same task and memory demands. Indeed, what distinguished these distributions was that $F_{self}$ represented only the fixations of a *one* observer in the group, while $F_{group}$ was sampled from *all* observers. No purely bottom-up or goal-driven account would have predicted a systematic individual differences between observers viewing the same scene and performing the same task. Yet in all three experiments, using a person's own search fixations provided more spatially precise and accurate information than using fixations of the other observers. The second argument is based on the target absent condition of experiment 2, in which some scenes were repeatedly searched but *never* contained a target. This condition, more reminiscent of a memory search (Wolfe, Klempen, & Dahlen, 2000) than a visual search, still produced more within-observer regularities than would have been expected from task-matched controls. To some extent, then, these observations suggest that similar findings might be obtained using other tasks.

In the tradition of ecological psychology (e.g. Gibson, 1979), our findings raise questions about the behavioral significance of self-similar fixation patterns over repeated scene exposures. One possibility is that within-observer consistency may promote good search performance (e.g. fast overall reaction time). A widely recognized feature of human memory is that reinstating the encoding context is beneficial for retrieval (Jacoby & Craik, 1979; Tulving & Thomson, 1973). Eye fixations, since they reflect which regions of the scene have been attended, may facilitate memory retrieval when deployed to previously attended scene regions. Indeed, Noton and Stark (1971) proposed a similar explanation for the self-consistent pattern and sequence of fixations ("scanpath") made by an observer when an image was viewed for the second (or third, etc) time. The existence of idiosyncratic fixation patterns, they posited, was evidence that sensory *and* oculomotor traces were encoded in the representation of a familiar image (Noton & Stark, 1971). Modern embodied cognition accounts also suggest that a person's own movements play a unique role in perceptual and cognitive performance (e.g. Knoblich & Flach, 2001). When imagining a previously viewed stimulus, for example, observers tend to make reenact eye movements made in the initial viewing (Brandt & Stark, 1997; Laeng & Teodorescu, 2002; Spivey & Geng, 2001). But are the eye movements in fact playing an important role in long-term memory retrieval?

Recent findings from eyetracking of recognition memory experiments provide some clues about the role of intra-observer fixation similarity on scene recognition. Holm & Mantyla (2007) used a remember/know paradigm to evaluate whether recognition performance was associated with how similarly observers re-fixated the same locations during study and test phases. Indeed, they found that recollection ("remember" responses) were related to a high degree of study-test consistency. Recently, Underwood and colleagues (2009) investigated the roles of domain knowledge and visual saliency on fixation consistency in scene recognition. Students of

engineering and history performed a recognition memory test with pictures of machinery, civil war artifacts, and neutral scenes. They confirmed that observers fixated similar places during study and test phases and that, interestingly, the effect was stronger for individuals who were experts in the domain related to the picture's content. Domain knowledge also had an interaction on recognition accuracy (Underwood, Foulsham & Humphrey, 2009). In each of these cases, however, the causal role between memory performance and study-test fixation similarity is suggestive but not conclusive.

Using visual search to investigate the functional connection between gaze and long-term memory has the advantage of allowing different types of memory to be tested. In the real world, people learn location priors without being explicitly instructed to encode the context and without knowing precisely how objects and contexts covary. These learning conditions may promote a dissociation between how information is used and what information is available for conscious report. Ryan, Althoff, Whitlow and Cohen (2000), for example, used the eye movements of amnesic patients and controls to assess the memory for scenes and spatial relations. After an initial scene presentation, participants viewed versions of the scenes that had been edited (objects were added, removed, or switched location). Control participants fixated the manipulated scene regions frequently, even when they were unable to explicitly report the change. Amnesic patients, interestingly, did not show this effect. Classical contextual cuing also shows that observers search repeated displays faster, despite being unable to recognize which displays had been repeated (Chun & Jiang, 1998, 2003). With natural scenes as backgrounds, observers do readily distinguish repeated from novel scenes (Brockmole & Henderson, 2006). Nevertheless, the role of explicit recognition processes on eye guidance remains poorly understood.

This chapter has shown that a person's experience biases the spatial distribution of search fixations. Our approach enabled us to probe long-term memory representations of specific scenes as they might be accessed during a ecologically relevant task. In this way, it allowed us to address issues related to how episodes of experience contribute to our behavior (e.g. visually searching a familiar environment), specifically how our eye movements and attention move around a scene. In the next chapter, I investigate temporal factors pertaining to the affect search of familiar scenes.

## Concluding Remarks

Comparative map analysis allowed us to evaluate how scene specific experience influenced fixation selection, as compared to guidance by scene context information alone. The results of these experiments support the idea that a familiar scene somehow triggers the retrieval of past search experience and rapidly biases fixation selection. This effect can be considered to be independent of the task, since all participants succeeded in the learning task (i.e. improved their reaction time over repeated searches of the scene). Since the same pattern of results is evident when predicting on the first search fixation, it is likely that this effect is implicitly activated by searching a familiar environment.

# References


Andrews, T. J., & Coppola, D. M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. *Vision Research*, 39, 2947-2953.

Araujo, C., Kowler, E., & Pavel, M. (2001). Eye movements during visual search: The cost of choosing the optimal path. *Vision Research*, 41, 3613-3625.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14, 143-177.

Boot, W.R., Becic, E., & Kramer, A.F. (2009). Stable Individual Difference in Search Strategy?: The Effect of Task Demands and Motivational Factors on Scanning Strategy in Visual Search. *Journal of Vision*, 9, 1-16.

Boyce, S.J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception & Performance*, 15, 556-566.

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

Brandt, S.A., & Stark, L.W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, 9, 27-38.

Brockmole, J. R., & Henderson, J. M. (2006). Using real-world scenes as contextual cues for search. *Visual Cognition*, 13, 99-108.

Bruce, N. D. & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9, 1-14.

Buswell, G. T. (1935). *How people look at pictures.* Oxford, UK: Oxford University Press.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception & Performance*, 33, 753-763.

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, 9(3), 1-15.

Castelhano, M.S., & Rayner, K. (2008). Eye movements during reading, visual search, scene perception: an overview. In K.Rayner, D. Shem, X. Bai, & G. Yan (Eds). *Cognitive and Cultural Influences on Eye Movements*. (pp.3-33). Tianjin People's Press/Psychology Press.

Chun, M.M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36,* 28-71

Chun, M. M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, 10, 360-365.

Chun, M. M., & Jiang, Y. (2003). Implicit, long-term spatial contextual memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 29, 224-234.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, 15, 559-564.

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting and Bayesian priors. *Psychological Science*, 17, 973-980.

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

Findlay, J.M. & Gilchrist, I.D. (2005). Eye guidance and visual search. In *Cognitive Processes in Eye Guidance*. Underwood G Oxford: Oxford University Press. 259-281.

Gibson, J. (1979). The ecological approach to visual perception. Boston: Houghton Mifflin.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. New York: John Wiley.

Gronau, N., Neta, M., & Bar, M. (2008). Integrated Contextual Representation for Objects' Identities and Their Locations. *Journal of Cognitive Neuroscience*, 20, 371-388.

Einhaüser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, 8, 1-26.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movement during visual search in real-world scenes. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movement research: Insights into mind and brain* (pp. 537-562). Oxford, UK: Elsevier.

Henderson, J. M., Weeks, P. A., Jr., & Hollingworth, A. (1999). Effects of semantic consistency on eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception & Performance*, 25, 210-228.

Hidalgo-Sotelo, B., Oliva, A., & Torralba, A. (2005). Human Learning of Contextual Object Priors: Where does the time go? *Proceedings of the IEEE Computer Society Conference on CVPR* (pp. 510-516).

Hollingworth, A. (2009). Two forms of scene memory guide visual search: Memory for scene context and memory for the binding of target object to scene location. *Visual Cognition*, 17, 273-291.

Holm, L., & Mantyla, T. (2007). Memory for scenes: Refixations reflect retrieval. *Memory & Cognition*, 35, 1664-1674.

Hwang, A.D., Higgins, E.C., Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision,* 9, 1-18.

Irwin, D. E., & Zelinsky, G. J. (2002). Eye movements and scene perception: Memory for things observed. *Perception & Psychophysics*, 64, 882-895.

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

Jacoby, L.L., & Craik, F.I. (1979). Effects of elaboration of processing at encoding and retrieval: Trace distinctiveness and recovery of initial context. *Levels of processing in human memory*. Hillsdale, NJ: Erlbaum.

Knoblich, G., & Flach, R. (2001). Predicting the effects of actions: Interactions of perception and action. *Psychological Science*, 12, 467-472.0

Kunar, M. A., Flusberg, S. J., & Wolfe, J. M. (2008). Time to Guide: Evidence for Delayed Attentional Guidance in Contextual Cueing. *Visual Cognition*, 16, 804-825.

Laeng, B. & Teodorescu, D. (2002). Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cognitive Psychology*, 26, 207-231.

Mannan, S., Ruddock, K.H., & Wooding, D.S. (1997). Fixation patterns made during brief examination of two-dimensional images. *Perception*, 26, 1059-1072.

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

Noton, D., & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science*, 171, 308-311.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, 11, 520-527.

Palmer, S.E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, 3, 519-526.

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

Parkhurst, D.J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision*, 16, 125-154.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10, 437-442.

Peterson, M. S., & Kramer, A. F. (2001). Attentional guidance of the eyes by contextual information and abrupt onsets. *Perception & Psychophysics*, 63, 1239-1249.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124, 371-422.

Reinagel, P., & Zador, A.M. (1999). Natural scene statistics at the centre of gaze. *Network*, 10, 341-350.

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

Ryan, J. D., Althoff, R. R., Whitlow, S., & Cohen, N. J. (2000). Amnesia is a deficit in relational memory. *Psychological Science*, 11, 454-461.

Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science*, *8,* 374-378.

Spivey, M. & Geng, J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological Research*, 65, 235-241.

Tatler, B.W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 1-17.

Tatler, B.W., Baddeley, R.J., & Gilchrist, I.D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45, 643-659.

Tatler, B.W., & Vincent, B.T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17, 1029-1054.

Torralba, A., Oliva, A., Castelhano, M., & 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.

Tseng, P.-H., Carmi, R., Cameron I.G.M., Munoz, D.P. & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes, *Journal of Vision,* 9, 1-167.

Tseng, Y. & Li, C.R. (2004). Oculomotor correlates of context-guided learning in visual search. *Perception & Psychophysics*, 66, 1363-1378.

Tulving, E., & Thomson, D.M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80, 352–373.

Underwood, G., Foulsham, T., & Humphrey, K. (2009). Saliency and scan patterns in the inspection of real-world scenes: Eye movements during encoding and recognition. *Visual Cognition*, 17, 812-834.

Wolfe, J.M. (1994). Guided search 2.0. A revised model of visual search. *Psychonomic Bulletin and Review*, 1, 202-228.

Wolfe, J. M., Klempen, N., & Dahlen, K. (2000). Postattentive vision. *Journal of Experimental Psychology: Human Perception & Performance*, 26, 693-716.

Yarbus, A. (1967). Eye movements and vision. New York: Plenum Press.

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

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 1-20.

CHAPTER 4

# Time Course of Guidance by Scene Specific Location Priors

In revision:
Hidalgo-Sotelo, B. & Oliva, A. Memory retrieval enhances eye guidance in specific real world scenes. In *Attention, Perception, & Psychophysics.*

# Introduction

Imagine walking along a street you believe you have never visited until, suddenly, you realize that *have* been on that street before… the realization prompts a host of other knowledge, for example remembering the location of a café on the next block. In real environments, scene context provides an associative basis for stitching together memories of past search experiences (Bar, Aminoff & Schacter 2008). Compared to searching *without* assistance of memory, using scene location priors can lead to a speedier outcome. Storing scene context information, however, is practical only if it can be retrieved in a timely fashion and input to attentional processes. How rapidly can specific scene context be retrieved and used to guide visual search?

The main topic of this chapter relates to memory based guidance over time, primarily studying how experience-based search priors guide selective attention in the orienting stages of a visual search. What cognitive processes are called into action when searching a familiar environment? Chapter 1 showed how local visual features of the environment can attract attention, for example fixating regions that are visually similar to the target. In searches with ecologically relevance, *long term memory* is an important source of orienting information, specifically, memory of the scene background or context.

Scene context provides an effective way of predicting where interesting or important events are likely to occur. Empirical support for the view that memory plays an active role in orienting attention comes mainly from visual search studies using the contextual cuing paradigm (Chun & Jiang, 1998, 1999). Additionally, semantic associations between objects has also been shown to influence the deployment of attention in speeded search tasks using real world objects (Moores, Laiti, & Chelazzi, 2003). In the remainder of the introduction, I provide more detailed evidence that memory can influence attentional mechanisms directly and then discuss issues related to timing of memory based guidance.

With this background, I motivate my investigation of how the temporal profiles of scene recognition and attentional guidance interact to guide attention in a familiar scene. I introduce an experimental paradigm- the *delayed-search* approach- which is similar the response deadline procedure (e.g. Reed, 1973) but adapted for oculomotor responses. The two experiments in this chapter use this paradigm to test the hypothesis that scene specific memory becomes increasingly available to guide search over time.

## *Spatial Orienting using Long Term Memory*

Beyond our everyday intuitions that memory provides direction in our lives, evidence that attentional mechanisms are directly influenced by memory comes from electrophysiological measures and functional imaging studies, as well as from contextual cuing studies with eye tracking. One interesting study from Summerfield and colleagues (2006) directly compared the behavioral and neural profiles of memory-based attention guidance with traditional spatial-cue based guidance. Memory based spatial cues were manipulated in their study, as in this chapter's experiments, by repeatedly presenting complex scenes with target objects consistently located across repeated presentations of the scenes. Event related fMRI results showed that both memory-guided and visually-guided attentional orienting activated a common network of brain

areas (Summerfield, Lepsien, Gitelman, Mesulam, & Nobre, 2006), including the parietal-frontal network which is consistently implicated in spatial orienting studies using perceptual cues (Corbetta & Shulman, 2002; Kastner and Ungerleider, 2000; Yantis et al, 2002).

Chun and Jiang (1998, 1999) initially proposed that the contextual cuing phenomenon, in which reaction times become reliably faster for targets in repeated versus novel displays after a few repetitions, occurs because an association is implicitly learned between the target location and surrounding context. When an incoming image matches one of the learned representations, search time is decreased because memory helps allocate attention to the target location. More recent experiments have shown that contextual cuing, in many ways, does not behave like typical attentional guidance (Kunar, Michod, & Wolfe, 2005) and, in part, comes from facilitating response selection (Kunar, Flusberg, Horowitz, & Wolfe, 2007). Current electrophysiological evidence, however, suggests that the *allocation* of attention is indeed modulated by contextual cuing. Johnson and colleagues recently recorded event-related potentials (ERPs) during contextual cuing and showed that repeated displays were associated with an increased amplitude beginning 175 ms after stimulus onset in the N2pc component (Johnson, Woodman, Braun, & Luck, 2007), a well-validated electrophysiological marker of the focusing of attention (Luck, Girelli, McDermott, & Ford, 1997). Furthermore, Chapter 2 of this dissertation showed that eye movement patterns were influenced by a person's past history of deploying attention in the scene. These results strongly suggest that long term memory for particular contexts can provide spatial cues (scene-specific location priors) that effectively orient spatial attention.

*Recognition and Attentional Guidance*

Given the close ties between attention and memory, what is the nature of their interaction during a particular episode of search? Visual search of a familiar environment involves two basic processes: *recognition*, in which the current context is matched to a stored representation in memory, and *guidance*, in which the remembered context is used to direct attention to particular regions of the environment. As noted above, the question of **how** scene-specific context guides attention is not well understood, but it does appear to act on brain regions involved in spatial attention (Johnson et al, 2007; Summerfield et al, 2006). Intriguingly, the hippocampus and adjacent medial temporal lobe (MTL) structures, known to be critical for encoding and retrieving information from declarative memory (Eichenbaum, 2004; Squire, Knowlton, & Musen, 1993; Squire & Zola-Morgan, 1991), are also the same regions implicated in scene processing and spatial cognition (Aguirre, Detre, Alsop, & D'Esposito, 1996; Burgess, Maguire, & O'Keefe, 2002; Epstein & Kanwisher, 1999; Summerfield et al, 2006). Recollecting encounters with scenes seems to require hippocampus-dependent processing (Tulving & Schacter, 1990; Eichenbaum, 1997; Gabrieli, 1998). Individuals with MTL damage do not seem to develop contextual cuing (Chun & Phelps, 1999) or show differences in differences in eye movements between repeated and novel scenes (Ryan, Althoff, Whitlow, Cohen, 2000). Thus, even when conscious awareness of the context is not required, using contextual information involves the participation of memory-related brain structures.

Does memory of a specific context always guide attention? Studies have investigated how and when searching memory might be a more efficient strategy than searching visually (Kunar, Flusberg, Wolfe, 2008a; Oliva, Wolfe, & Arsenio, 2004; Wolfe, Klempen, Dahlen, 2000). One

key result suggests that memory guides attention when the context allows you to learn that targets can appear in *some* locations but not others, thereby reducing the effective set size (Kunar et al, 2008a). In contextual cuing, however, repeated displays do indeed cue a single location but context does not immediately guide attention to the target, or at least not on every occasion. To this effect, Chun and Jiang (1998, Experiment 4) report steep search slopes on repeated displays, approximately 27 ms/item, even after extensive repetition (see also Kunar, Flusberg, & Wolfe, 2006; Kunar et al, 2007). What does this imply about the relationship between recognition and guidance? Importantly, the contextual cuing effect reflects the average behavior across a block of trials, making it difficult to know whether recognition occurs on *every* trial.

Eye movements provide one means of probing the strength and timing of recognition's effect on guidance. Peterson and Kramer (2001) recorded eye movements while participants performed a contextual cuing-style search for a small rotated T among distractor Ls, a task which has a history of eliciting serial deployments of attention (Wolfe Cave, Franzel, 1989). Once a repeated display is matched to a memory representation (recognition), the authors hypothesized, the eyes then should be biased to move towards the target (guidance). The proximity of each fixation to the target was taken as a measure of the precision of the guidance system. Interestingly, recognition of the context was rarely immediate, given that participant's **first** search fixation landed on the target on only 15% of trials in the last epoch of learning (repetitions 12-16). On the remainder of trials, the first fixation was not biased towards the target *yet still* fewer fixations were ultimately required to locate the target than for novel displays (Peterson & Kramer, 2001). Indeed, if recognition of the repeated context generally occurred *after* search had begun, that would account for at least part of why search slopes remain steep even after extensive training (e.g. Chun & Jiang, 1998). These results are directly relevant to the proposal in this chapter, as they support the idea that retrieving scene-specific information facilitates visual search but that successful retrieval is more likely over longer time intervals.

To date, there is little direct evidence that *time* increases the likelihood that specific real world scenes will be recognized and that those spatial cues will help guide attention to the target. This is the hypothesis that I propose to investigate in this chapter. The experiments in this chapter use eye tracking for two reasons: (1) as a dependent measure of the difference between repeated and novel conditions, but also (2) to manipulate an independent variable, specifically, the time elapsed before overt attention is deployed in a scene. **By manipulating the time between the *scene context presentation* and the *oculomotor response*, I propose to study the effect of memory retrieval on the guidance of attention during search.** Similar approaches have a tradition of being used to investigate temporal properties of memory and attentional processing.

*Studying the Time Course of Internal Processes*

Decisions, whether of the laboratory variety ("Was the gabor tilted to the left or right?") or of the everyday variety ("Do I know the person waving across the street?"), are accompanied by requisite tradeoffs between speed and accuracy. The pressure to make speedier decisions often causes a corresponding increase in error rates, which recovers as more time is allowed before making a response. Speed and accuracy tradeoffs have long been studied in the memory literature with a method called the response-signal, or response-deadline, procedure (Corbett & Wickelgren, 1978; Dosher, 1976; Reed, 1973). This procedure involves presenting a test item to

participants, followed by a delay, and ending with a signal to make a recognition judgment (E.g. Was this word previously studied?). The delay before the signal is manipulated to control the amount of time available for retrieval (Reed, 1973).

One question has involved studying automatic versus controlled processes in recognition memory (Toth, 1996). The argument is that, at the shorter response deadline, recognition will depend mainly on faster, more automatic familiarity processes. With a longer response deadline, the recognition judgment will depend more on slower, more effortful recollection processes. The response-signal procedure is a way to separate the two components of recognition (Toth, 1996). Similarly, this procedure has been used to study interference effects in memory retrieval (Dosher, 1981), episodic versus semantic memory associations (Dosher, 1984), and models of decision making (Ruthruff, 1996). Measuring the time-accuracy functions of visual search tasks has also been useful for refining models of attentional processing (Carrasco & McElree, 2001; Dosher, Han, Lu, 2004; McElree & Carassco, 1999).

*Overview*

What is the influence of memory retrieval on eye guidance during search of specific, familiar environments? The experiments in this chapter study how overt visual attention is guided by memory at different points in time by using a *delayed-search* approach. This procedure required that participants hold their gaze at central fixation on scene (or a blank background) for a variable amount of time before overtly localizing a target embedded in the scene. Using covert attention to search during the delay period was not effective because the search targets were small and peripherally located. Specifically, the role of memory retrieval in search was investigated by manipulating (1) the familiarity of the scene, and (2) the duration of the interval between presentation of the scene and initiation of overt search. The main proposal of this chapter is that enhancing the retrieval of learned scene-target associations improves attentional guidance in cluttered, real world scenes.

Experiment 1 begins to build the case that manipulating the time before an initial saccade can be an effective way to control the retrieval of scene specific location priors. Since there have been relatively few studies of how people learn where real world objects are located in scenes (Hidalgo-Sotelo, Oliva, & Torralba, 2005; Summerfield et al, 2006), this experiment explored how retrieval influences the *acquisition* (i.e. learning curve over repeated searches) and the *expression* of acquired scene specific knowledge. Interestingly, steeper learning curves and overall lower reaction times resulted when observers were delayed for an extended interval before making an initial saccade. Experiment 2 investigated the time course of retrieving scene specific knowledge in greater detail, focusing on how varying amounts of retrieval-time affected search performance. The results again supported the idea that **longer time intervals** increased the effectiveness of using memory to help guide attention to the target. Taken together, the experiments in this chapter suggest that remembering scene specific associations requires some time, but that doing so improves attentional guidance to a consistently-located object in familiar scenes.

# Experiment 1

The purpose of this experiment was to determine whether delaying overt search could effectively control the degree to which scene specific location priors were retrieved. Participants were instructed to find a pedestrian in novel and repeated outdoor scenes in which the location of the target was consistent within a particular scene. The unique aspect of this task was that on each trial, participants were required to fixate centrally on the scene until a signal was given that eye movements could be initiated. Importantly, two delay durations were tested: a control group fixated centrally for 300 ms prior to making a saccade- a duration selected to approximate the typical central fixation duration in visual search tasks (Findlay, 1997); another group of observers, the extended-delay group, fixated centrally for 1300 ms before initiating a saccade.

Typically, search performance improves on repeated scenes as observers learn the association between a particular scene and the location of the target. This study investigated how the delay-duration manipulation influenced the *learning* of this association and suggests that enhanced memory retrieval was the key reason. Consistent with earlier findings (Kunar et al, 2008b; Summerfield et al, 2006), observers who were delayed for a *extended* time interval demonstrated better search guidance relative to a shorter time interval. In fact, the extended-delay group seemed to learn the scene specific location priors more rapidly than the control group. Was the performance of the control group due to poorer learning of the visual context or poorer search guidance? In the final block of search, the delay durations of each group were switched. Interestingly, we found evidence that the control group could eliminate a fixation and perform more efficient paths when their delay duration was lengthened. We conclude that achieving stronger context retrieval before initiating overt search may enhance the efficacy of attentional guidance in very familiar environments.

Figure 1. Example of repeated scene stimuli in experiment 1 (top) with eye movements (bottom) from one observer's search of each scene. Central and search fixations are shown in red; the fixation on the target is green.

## METHOD

*Participants.* Eighteen observers, age range 19-33, participated in the study. Observers were randomly assigned to one of two groups (control group or extended-delay group). All participants were tested for normal visual acuity (one individual required soft contact lens, others did not require visual correction) and were paid $15 for their participation. Informed consent was obtained for all participants.

*Apparatus.* Eye movements were recorded at 240 Hz using an ISCAN RK-464 video-based eyetracker. Observers sat at 75 cm from the display monitor, 65 cm from the eyetracking camera, with their head centered and stabilized in a headrest. The position of the right eye was tracked and viewing conditions were binocular. Stimuli were presented on a 19" LCD monitor with a resolution of 1280 by 1024 pixels and a refresh rate of 70 Hz. Presentation of the stimuli was controlled with Matlab and Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

*Stimuli.* The stimuli were color photographs of outdoor urban environments (see Figure 1). The images subtended visual angles of 22.7° (horizontal) by 17.0° (vertical) on the screen.

A total of 220 distinct images were shown to each observer. These images belonged to one of three categories: norming scenes, novel scenes, and repeated scenes. Norming stimuli were 38 scenes presented in the initial search block to all observers, regardless of group, with an 800 ms delay preceding search; the purpose of having this norming block was to assess whether the observers comprising each group were comparable in their search performance. Novel stimuli consisted of 144 scenes that were presented once for each observer and distributed randomly across experimental blocks. Repeated stimuli consisted of 22 scenes that were presented a total of 9 times for each observer. All observers viewed the same set of repeated scene stimuli. The search target for all stimuli was a **person** in the scene and all scenes contained a target (100% prevalence). Within a particular scene, the search target was always at the same location within that scene.

Critically, the repeated scenes were chosen such that the targets were not detectable in the periphery while observers were centrally fixating. If observers in the extended-delay group could make greater use of peripheral cues than the control group, then covert search of the periphery would contribute to faster search time for even the first exposure of the scene. In order to reduce the utility of covert search, the target objects in the repeated scenes were small (mean size of 0.4° by 0.8°), camouflaged, and located at a average eccentricity of 8°. Under these conditions, searching the perceptual information in the periphery would not favor either group regardless of delay duration.

*Procedure.* First, the eyetracker was calibrated for each observer. Tracking accuracy was then checked using a routine in which observers fixated a dot that appeared at each of 9 locations evenly distributed across the screen at 10° of eccentricity from the center. In this routine, observers pressed a key to indicate that their gaze was on the dot and eye position was recorded for 0.5 sec. The estimated fixation position had to be within 0.75° of visual angle for all 9 points, otherwise the experiment halted and the observer was re-calibrated. Observers performed this testing routine every 50 trials (roughly every 6 min) throughout the experiment. Additionally, two scheduled breaks were interspersed in the experiment in which observers were required to

take at least a 2 minute break. We found that this helped to alleviate eye fatigue and facilitated tracking accuracy. The total experiment duration was approximately 60 minutes.

Observers performed a visual search task in which their goal was to move their eyes to the person in the scene and, once found, to maintain fixation on the target-person. Experiment instructions explicitly indicated that the observer's eye position would cue the beginning and end of each trial. Figure 2 depicts the trial events. Observers fixated a central cross on a gray screen to begin a trial (600 ms), followed by the onset of the search scene with a fixation cross overlaid on the scene (size 2°). This cross indicated that the observer was to **hold** central fixation (300 ms for the control group; 1300 ms for the extended-delay group), until the cross turned red (100 ms) and a simultaneous auditory cue signaled that the observer should begin an overt visual search. The search scene remained on screen until the observer fixated the target or for a maximum of 10 seconds. Fixation on the target was determined to have occurred when the observer's gaze fixated the target region for 800 ms.

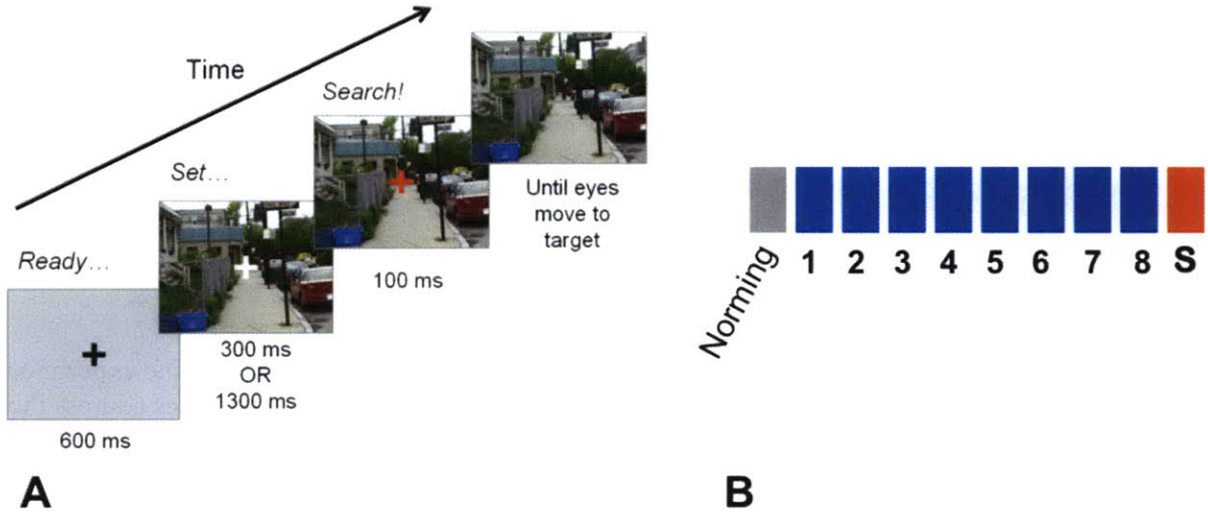



Figure 2. (A) **Trial sequence**. Observers fixated a central cross on a gray screen to begin the trial (600 ms). The search scene was presented with an overlaid fixation cross to indicate that the observer was to hold central fixation (Control group: 300 ms; Extended-delay group: 1300 ms), until the cross briefly turned red (100 ms) and a simultaneous auditory cue signaled that the observer should begin overt visual search. The search scene remained on screen for a maximum of 10 s or when the observer fixated the target for 800 ms. (B) **Experiment structure**. A total of 10 blocks of search were completed. *Norming block*: a block with unique scenes searched by both groups with 800 ms delay enforced at central fixation. *Learning Phase*: 8 blocks of search with a delay determined by condition. *Switch Phase*: a block in which the delay duration was lengthened to 1300 ms for the control group, and shortened to 300 ms for the extended-delay group.

Each observer performed 10 blocks of search trials, with each block containing 38 trials. In the initial, norming search block, all observers searched norming stimuli, presented in randomized order and preceded by an enforced delay of 800 ms. Since the critical manipulation- central fixation delay- assumed that observers were sampled randomly from the population, this

norming block provided a baseline measure of observer's search performance before experimental manipulation. The stimuli shown in this block were never repeated in the experiment.

The Learning phase consisted of 8 blocks of search, each of which contained 22 repeated scenes and 16 novel scenes randomly interleaved in each block. The Switch phase consisted of one final block, in which the delay-time was either lengthened (1300 ms for the Short-delay group) or shortened (300 ms for the Long-delay group). With this last manipulation, we probed the robustness of the learned scene representations by varying the amount of scene exposure (delay-time) between the Learning phase and the Switch phase.

*Eye Movement Analysis.* Eye position was recorded throughout each trial and was used as a measure of search performance, as well as a way to manipulate the independent variable of central fixation duration. Eye movement analyses were performed on smoothed eye position data, averaging the raw data over a moving window of 8 data points (33 ms). The beginning and end positions of saccades were detected using an algorithm implementing an acceleration criterion (Araujo, Kowler, Pavel, 2001). Specifically, the velocity was calculated for two overlapping 17 ms intervals; the onset of the second interval was 4.17 ms after the onset of the first. The acceleration threshold was set at a velocity change of 6 °/s between the two intervals. Saccade onset was defined as the time when acceleration exceeded the threshold, and saccade termination was defined as the time when acceleration dropped below the threshold. Fixations were defined as the periods between successive saccades.

## RESULTS

One observer did not show the classical task improvement over the blocks of the experiment (see slope analysis of search reaction time below). Since the proposed hypotheses address the development and expression of memory in visual search performance, the analyses reported below exclude this observer's data.

*Accuracy.* A search was scored as correct when an observer located the target-person in the scene before the maximum search time, 10 sec, elapsed. Observers failed to find the target in 5.9% and 5.8% of trials for the control and extended-delay groups, respectively. Trials in which a saccade was initiated before the delay duration elapsed or in which the eye movement signal was lost for at least 5 successive data points were removed. These criteria resulted in the removal of 6.7% and 8.1% of trials for the control and extended-delay groups, respectively. Overall, the number of trials in the analysis below did not significantly differ across condition, $t(15) = 0.09$.

Data analyses are reported on the following measures: search reaction time, number of search fixations, and scan path efficiency. *Search reaction time*, or search RT, is defined as the time elapsed from the offset of the start-search cue until the beginning of the final fixation on the target (fixed at 800 ms). This measure corresponds to the efficiency of the search phase of the task. Search RT directly arises from the number of fixations made until the observer fixated the target, and the duration of fixations and the interleaved saccades. *Number of search fixations* was defined by the number of discrete fixations on the scene up until, but not including, the final

fixation on the target. *Scan path efficiency* was defined as the ratio of the cumulative length of all saccades made in a trial divided by the most direct path to the target from the center of the scene (Henderson, Weeks, & Hollingworth, 1999). Consequently, obtaining an efficiency value of 1 would indicate that the observer's eyes saccaded directly to the target.

**Norming Block Performance.**

The norming block of scenes was measured for the purpose of assessing search performance among observers randomly assigned to each group. The delay duration for all observers was 800 ms. Search performance did not significantly differ between groups in mean search reaction time ($M$ = 627 ms, 639 ms for the control and extended-delay groups respectively, $t(15) = 0.22$), number of search fixations ($M$ = 2.16, 2.21 fixations, $t(15) = 0.33$), cumulative saccade distance ($M$ = 18.8°, 18.8°, $t(15) = 0.04$), or scan path efficiency ($M$ = 2.31, 2.33, $t(15) = 0.17$). This result indicates that the observers were comparable in search performance before the experimental manipulation.

**Learning Phase Results**

First, we consider search performance during the Learning phase (blocks 1-8): When scenes become familiar over repetition, search performance improves as observers learn the location of the target in each scene (e.g. Chun & Jiang, 1998). Can this improvement be modulated by the strength of memory retrieval, modeled by a delay at central fixation before initiating eye movements?

*Search Reaction Time.*

On the first block of search, mean search RT did not differ between groups (See Figure 3; $t(15) = 0.11$), indicating that the manipulation of delay duration did not provide one group with an initial advantage in search speed. Mean search RT decreased significantly across search blocks for both groups: one-way repeated measures ANOVA for the control group, $F(7,56) = 16.1$, $p < 0.0001$ and for the extended-delay group $F(7,49) = 34.9$, $p < 0.0001$. There was also a significant interaction between group (control or extended-delay) across search blocks, with observers in the extended-delay group improving search RT significantly more than the control group: $F(7,105) = 2.26$, $p < 0.05$. Table 1 shows a summary of search performance measures for each group, early and late in the learning process.

A linear regression was used to estimate the search improvement of each observer during the Learning phase: search block (1-8) was used as the regressor for mean RT on repeated scenes. As described above, one observer was excluded from this analysis because no evidence of learning was observed over the duration of the experiment; this observer's slope estimate was more than two standard deviations outside of the mean of all observers and the mean of the condition assigned (extended-delay group). For the rest of observers, an independent t-test between group slope averages showed that learning occurred more quickly (i.e. steeper slope) in the extended-delay group than the controls ($t(15) = 1.96$, $p < 0.05$).

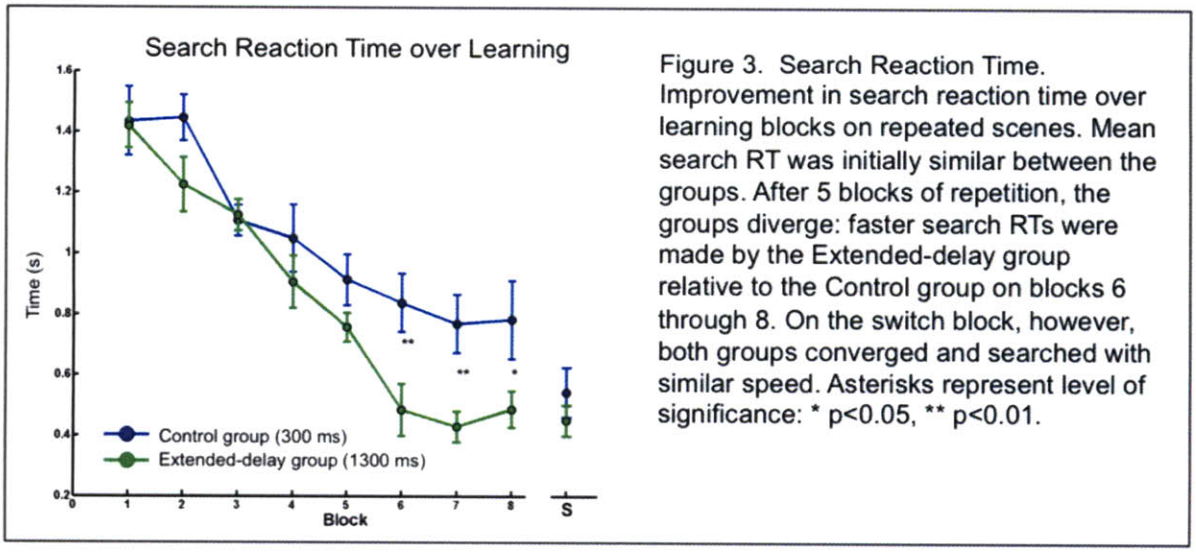



Figure 3. Search Reaction Time. Improvement in search reaction time over learning blocks on repeated scenes. Mean search RT was initially similar between the groups. After 5 blocks of repetition, the groups diverge: faster search RTs were made by the Extended-delay group relative to the Control group on blocks 6 through 8. On the switch block, however, both groups converged and searched with similar speed. Asterisks represent level of significance: * p<0.05, ** p<0.01.

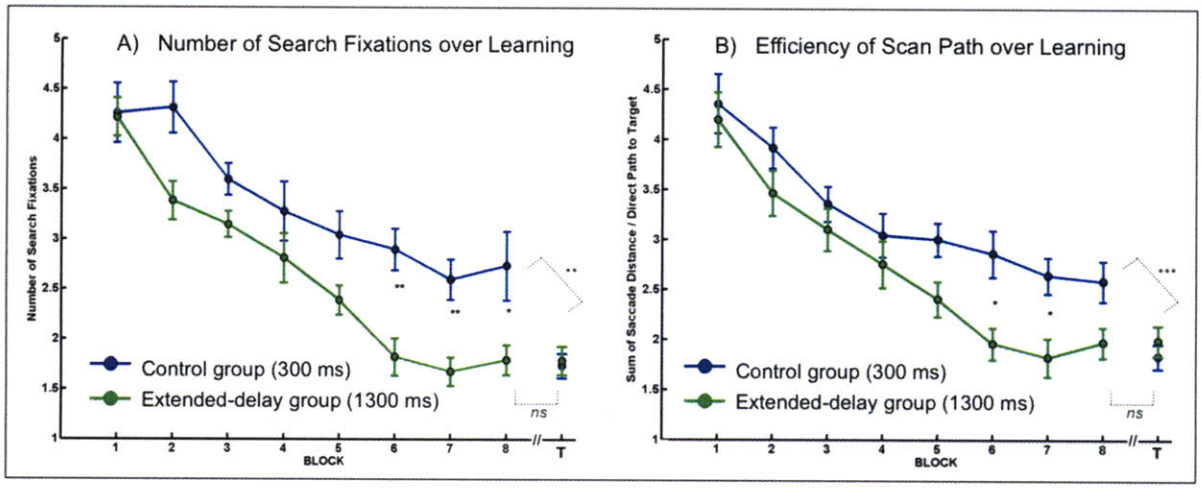



Figure 4. Eye movement measures.

A) Number of fixations on each trial, from the beginning of search until (and excluding) the final fixation on the target. Under this definition, the optimal number of fixations on a trial is 1, where the only fixation is at central fixation and the subsequent fixation falls on the target. On blocks 6 through 8 of the Learning Phase, the Control group made significantly more fixations per trial than the Extended-delay group. On the Switch Block, however, the Control group was given an *additional* 1000 ms of delay time and made significantly fewer fixations. In contrast, a *reduction* of 1000 ms of delay time did not significantly change search performance for the "Extended-delay" group.

B) Scan path efficiency was computed from the sum of saccade lengths normalized by the most direct path-to-target. Under this definition, the most efficient scan path has a value of 1, indicating that the total saccadic distance was equivalent to the optimal path to the target. Results showed a pattern in which, again, both groups initially performed search with similar scan path efficiency, but in late learning blocks (blocks 6 and 7) the Extended-delay group exhibited more efficient scan paths than the Control group. Asterisks represent level of significance: * p<0.05, ** p<0.01, *** p<0.001.

*Number of Search Fixations.*

The average number of search fixations is shown in Figure 4a. Search fixations include those occurring **after** the signal to begin overt search but excluding the final target fixation. The pattern of results is very similar to the results of search RT.

In the first block, the number of search fixations did not significantly differ between the control and extended-delay groups ($t(15) = 0.11$, see Figure 4a). Again, this suggests that the additional delay time (1000 ms) between the extended and control groups did not allow covert search mechanisms to locate the target with fewer fixations. The mean number of search fixations decreased significantly across search blocks for both groups: $F(7,56) = 14.6$, $p < 0.0001$ for controls and $F(7,49) = 34.9$, $p < 0.0001$ for the extended-delay group. Additionally, the interaction of group X search blocks was significant, with the extended-delay group exhibiting a greater decrease in search fixations over the Learning phase ($F(7,105) = 2.23$, $p < 0.05$).

*Efficiency of Scan Path.*

This measure reflects how directly gaze was deployed towards the target in a search trial. The efficiency of the scan path is computed from the sum of saccade lengths normalized by the shortest distance between the target and central fixation. Thus, the most direct scan path has an efficiency value of 1. Similar to the pattern of results above, the mean scan path efficiency was not significantly different between groups on the first block ($t(15) = 0.41$, see Figure 4b). Scan paths became significantly more efficient across blocks: $F(7,56) = 16.6$, $p < 0.0001$ for the control group and $F(7,49) = 30.1$, $p < 0.0001$ for the extended-delay group. However, the interaction between group by block was not significant: $F(7,105) = 1.51$.

*Novel Scene Results*

Comparing search performance with the novel scenes is informative for addressing the possibility that a generalized improvement over time, rather than scene-specific memory, could underlie part of the decrease in search RT over blocks. Indeed, the delay-manipulation was not predicted to have any significant effect on novel search: without scene specific location priors, the duration of the delay was not predicted to have a differential impact on search RT.

There was no evidence that the two groups performed search of novel scenes differently: mean search RT for novel scenes by control group was 389 ms and the extended-delay group was 377 ms, with no significant difference between the two groups ($t(15) = 0.11$). There was a significant overall main effect of scene familiarity condition- novel vs familiar- $F(1,32) = 168$, $p < 0.0001$. Unfortunately, this shows that the novel scenes were overall searched faster than the repeated scenes, and indicates a possible floor effect in search performances. To pursue the question further, a subset of 27 novel scenes were extracted (from total of 144 novel scenes) to compare with the repeated condition, on the selection basis of having 2 search fixations or more, on average, before the target was found. These 27 scenes were selected without regard to whether the scenes were searched by the control or extended-delay group, and were presented an approximately equal number of times in the earlier search blocks (1-3) as in the later ones (blocks 6-8).

Search RTs ($M$=747 ms) and number of search fixations ($M$=2.71) were comparable to repeated scenes on these 27 novel scenes. One key question was whether these difficult novel scenes tended to be searched more efficiently at the *end* of the experiment (blocks 6-8) relative to the beginning (blocks 1-3). This was not the case, as the results did not show a difference in search RT between the early vs late search blocks, analyzed by each group alone (control group: $t(26)$=0.06, extended-delay group: $t(26)$=0.07) or by both groups combined ($t(26)$=0.35). Likewise, the number of fixations and scan path efficiency was similar in early and late search blocks in the control and extended-delay group. Table 2 shows a summary of search performance on the 27 difficult novel scenes, in early (1-3) and late (6-8) search blocks.

*Summary*

In the first block of the learning phase, both controls and the extended-delay group exhibited equivalent search performance. As observers learned the predictive relationship between a specific scene and target location (scene specific location priors), both groups located the target with fewer fixations and more efficient scan paths. This showed that our experimental paradigm of introducing a delay between scene presentation and gaze deployment did not hinder the classic effect of context improving search RT. More importantly, the length of the delay- 300 ms or 1300 ms- modulated the magnitude of overall improvement in search RT and corresponding eye movement measures. In late blocks of learning, the extended-delay group performed markedly more efficient searches of the familiar scenes relative to the control group: scene learning occurred at a faster rate and converged to an overall lower search RT.

A main effect of scene type (repeated versus novel) did not allow a comprehensive comparison between delay-duration and scene familiarity. However, a subset of the novel scenes were analyzed to address the possibility that the extended-delay group learned a more general strategy that allowed them to search novel as well as repeated scenes more efficiently. Comparing trials in which novel scenes were presented early versus late in the experiment did not show a difference in search performance for either group. Overall, these results suggest that searching familiar scenes was more efficient when observers were given a longer time interval before deploying eye movements.

**Switch Phase Results**

Next, we address the Switch phase (block 9): How did search performance respond to changing the time interval preceding overt search? For the control group, does an *additional* 1000 ms of delay (from 300 ms to 1300 ms) facilitate a more efficient visual search through the familiar scenes? Likewise, for the extended-delay group, does *reducing* the delay by 1000 ms (from 1300 ms to 300 ms) impair the ability to perform an efficient search of familiar scenes? The results of the switch phase are shown in the last block of Figures 3 and 4, as well as in Figure 5.

*Search Reaction Time*

A measure of learning transfer was calculated from the difference in search RT on the familiar scenes between the end of the Learning phase (block 8) and the Switch phase (block 9); results are shown in Figure 3b. When the delay was lengthened for the control group, search RT dropped significantly ($M$ = 782 ms, 543 ms on blocks 8 and 9, Δ-Transfer: $t(8) = 2.26$, $p < 0.05$). In contrast, search RT was not significantly different when the delay was reduced for the

extended-delay group ($M$ = 487 ms, 450 ms on blocks 8 and 9, Δ-Transfer: $t(7) = 0.53$). Table 2 summarizes search eye movement measures at key points in the Learning and Switch phases.

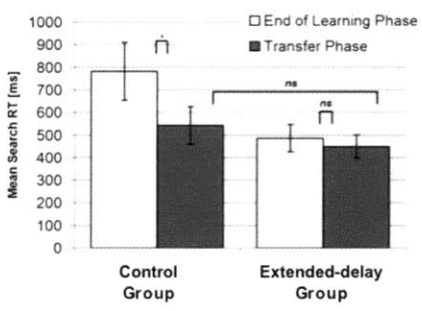



Figure 5. Effect of switching delay duration. Changing the delay duration between the end of the Learning Phase (Block 8) and the Switch Phase (Block 9). The effect depended on the delay-duration during the learning phase: For the control group, enforcing an *additional* 1000 ms of delay time (from 300 ms to 1300 ms) resulted in a significantly faster mean search RT on the Switch block. For the extended-delay group, a *reduction* of 1000 ms of delay time (from 1300 ms to 300ms) did not significantly affect search performance.

*Number of Search Fixations*

Comparing performance between the end of the Learning phase and the Switch phases showed that increasing the delay duration for the control group (from 300 ms to 1300 ms) had the effect of significantly reducing the number of search fixations ($M$ = 2.73, 1.72 fixations on blocks 8 and 9, $t(8) = 3.94$, $p < 0.01$). In contrast, decreasing the delay by 1000 ms for the extended-delay group had no significant effect of on the number of search fixations ($M$ = 1.79, 1.79 fixations on blocks 8 and 9, $t(7) = 0.03$).

*Efficiency of Scan Path*

The control group showed a dramatic improvement in scan path efficiency between the end of the Learning phase and Switch phase ($M$ = 2.58, 1.83 on blocks 8 and 9, $t(8) = 4.54$, $p < 0.001$). On the other hand, the scan path efficiency of the extended-delay group was not impaired by reducing the delay by 1000 ms ($M$ = 1.97, 1.99 on blocks 8 and 9, $t(7) = 0.15$).

*Summary*

Critically, the switch block showed that lengthening the delay by 1000 ms allowed observers in the control group to find targets in familiar scenes as quickly as the extended-delay group. A significant improvement in all eye movement measures was found between the end of the Learning phase and the Switch phase for the control group. This result suggests that the duration of the time interval prior to deploying eye movements was important for allowing observers retrieve scene specific location priors and deploy gaze most effectively. Interestingly, the extended-delay group was not significantly impaired in search performance when the delay was shortened to 300 ms. Potential reasons for this asymmetry in the results are discussed further below.

Table 1

*Summary of search eye movement measures on Repeated Scenes at key points in the Learning and Switch phases*

| | **Early Learning**: Block 1 | | **Late Learning**: Block 8 | | **Switch**: Block 9 | |
|---|---|---|---|---|---|---|
| | Control Group | Extended Delay Group | Control Group | Extended Delay Group | Control Group | Extended Delay Group |
| | *300 ms* | *1300 ms* | *300 ms* | *1300 ms* | ***1300 ms*** | ***300 ms*** |
| Search RT (ms) | | | | | | |
| *M* | 1433 | 1419 | 782 | 487 | 543 | 450 |
| *SE* | *113* | *74* | *128* | *60* | *82* | *51* |
| Number of Fixations | | | | | | |
| *M* | 4.26 | 4.22 | 2.73 | 1.79 | 1.74 | 1.79 |
| *SE* | *0.30* | *0.19* | *0.34* | *0.15* | *0.12* | *0.14* |
| Scan Path Efficiency | | | | | | |
| *M* | 4.36 | 4.20 | 2.58 | 1.97 | 1.83 | 1.99 |
| *SE* | *0.03* | *0.27* | *0.20* | *0.15* | *0.12* | *0.15* |

Table 2

*Summary of search performance on difficult novel scenes (n=27)*

| | **Search RT (ms)** | | **N Fixations** | | **Scan Path Efficiency** | |
|---|---|---|---|---|---|---|
| | Early | Late | Early | Late | Early | Late |
| *Both Groups* | 776 | 748 | 2.78 | 2.63 | 2.48 | 2.50 |
| *Controls Only* | 710 | 715 | 2.72 | 2.60 | 2.37 | 2.44 |
| *Extended-delay Only* | 821 | 811 | 2.83 | 2.77 | 2.60 | 2.57 |

## INTERPRETATION

The present study showed that memory retrieval improved attentional guidance toward a target location in familiar, real world scenes and suggests that *time* played a key role in controlling how recognition guided attention. The pattern of RT improvement depended on the duration of an enforced delay prior to the initial saccade, either a typical (300 ms) or extended (1300 ms) central fixation on the scene before the initial saccade. Although both groups searched with similar mean RTs in the early blocks of learning, the extended-delay group exhibited significantly faster search times in later search blocks and overall steeper slopes (RT X block). Observers in the extended-delay group made, on average, fewer fixations en route to the target and overall more efficient scan paths in familiar scenes than controls. Did the difference in performance arise from a more robust *acquisition* or more efficient *expression* of the learned association by the extended-delay group?

On a final block, the delay duration was switched between groups: observers in the control group were forced to fixate centrally for 1300 ms instead of 300 ms, and observers in the extended-delay group were likewise cued to search a second earlier than before. Suddenly, the control group searched as quickly as the extended-delay group's faster search performance, making fewer fixations and more efficient scan paths than when those *same observers* had been cued with a shorter delay (i.e. block 8). This supports the idea that a longer time interval is a critical factor allow promoting better expression of learned associations. Interestingly, however, the "extended-delay" group was relatively unaffected by shortening the duration of the central fixation: few search fixations and efficient scan paths were executed by these observers when given only 300 ms before making a saccade. This result, not consistent with slow- time course account of memory-based guidance, suggests that this group also benefited from a more robust acquisition of the scene-target association during the learning phase.

One possible explanation for why the extended-delay group acquired the learned association more robustly than controls is related to the steeper slopes in the learning phase. After 5 blocks, the average search time of the extended-delay group matched the *fastest* search time of the control group. What aspects of the delay manipulation could explain the difference in search slope? For both groups, repetition reinforced a *single* location in which the target was consistently located across multiple searches of a given scene. But the "extended-delay" introduced time- 1 sec difference between a typical and "extended" delay- in which observers could retrieve information, such as relevant locations associated with that scene. It is possible that retrieval over this time interval may represent an upper bound on the ability of familiar scene recognition to guide attention.

It should be noted again that response selection did not contribute to the observed pattern of results in this experiment. Not only did search performance differences only emerge in *late* learning blocks, the main dependent measure, search RT, includes the search time until but excluding the final target fixation. Furthermore, a **longer** time interval was associated with better retrieval but, in this experiment, the procedure involved keeping the scene present during the course of the delay. How critical was the *presence of the visual scene* during the delay interval for retrieving location priors and biasing attention toward the cued location? Since the experiment did not adequately control for search difficulty across familiarity conditions- the overall novel search RT regrettably approached floor performance- it would be a stronger test of the hypothesis if the same, identical scenes were searched in both conditions. An even stronger test would be to use within-subject measures for the amount of search facilitation from memory over different time intervals. Experiment 2 was designed to address these concerns and further investigate the time course of memory retrieval and search guidance.

## Search Experiment 2

The purpose of this experiment was to further investigate how long-term memory influences visual search as a function of the amount of time to retrieve scene-specific location priors. Observers were given the task of finding a book in indoor scenes (e.g. kitchens, bedrooms) that were either searched once (Novel condition) or searched repeatedly (Familiar condition). As in experiment 1, the target book's location was unchanged in repeated presentations of the same scene, allowing observers to learn an association between a specific scene and the location of a book in that scene. In accord with previous findings (Kunar et al, 2008b; Peterson & Kramer, 2001; Summerfield et al, 2006), the results of experiment 1 showed that extending the time interval longer than a typical central fixation duration (approx. 300 ms, Henderson & Hollingworth, 1998) made it more likely that scene specific associations were retrieved and used to the eyes to the target on a given trial.

The present experiment provided a stronger test of this idea by using a within-observer manipulation of the critical condition, *retrieval time* (SOA manipulation). Furthermore, search difficulty was matched across the other critical condition, *scene familiarity* (novel vs familiar scenes), by counterbalancing scenes across observers. Finally, the delayed-search procedure in this experiment enforced the variable time-interval on a blank gray screen, after a 200 ms scene preview, which further points to memory retrieval as the key factor being influenced by time. Whereas experiment 1 was concerned with the learning process, this experiment focuses on the role of time on previously established scene-target associations. The procedure is therefore divided into a Learning phase, in which specific scene location priors were learned, and a Test phase, in which the main hypotheses were tested.[10]

The purpose of an initial Learning phase was for observers to learn scene specific location priors for the repeated scenes. A subsequent Test phase then presented each scene briefly (200 ms) followed by a variable SOA (ranging from 0 to 1.4 seconds). In both parts of the experiment, observers were instructed to visually search the scenes and fixate the book in each scene. The main difference was that, after a 200 ms scene preview, in the Test phase each trial was followed by a variable delay on a gray-screen before the scene was visible again. The main hypothesis, positing a direct relationship between memory-based guidance and time, predicted an interaction between the influence of SOA and scene familiarity on overt search. Longer delays, it was hypothesized, would predict better search performance on familiar, but not novel, scenes.

### METHOD

*Participants.* Twenty observers, ages 18-34, with normal acuity gave informed consent, passed an eyetracking calibration test, and were paid $15/hr for their participation.

*Apparatus.* Eye movements were collected using an ISCAN RK-464 video-based eyetracker with a sampling rate of 240 Hz. The stimuli were high resolution color photographs presented on

[10] Note that this experiment has been described in Chapter 2, Search Experiment 1. Comparative map analysis was performed on the eye movement data that was collected from the learning and test phases, collapsing across SOA manipulation. Here, I report the results of the time manipulation on search performance (as opposed to fixation location, per se).

a 15" LCD monitor with a resolution of 1280 x 1024 px and refresh rate of 60 Hz. Presentation of the stimuli was controlled with Matlab and Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

*Stimuli.* The stimuli were high resolution color photographs of indoor environments- kitchens, living rooms, bedrooms, dining rooms- obtained by digitizing interior design books, and downloading from LabelMe and Flickr® databases. The original images contained a target-book in the scene and were cropped and resized to be presented at a resolution of 1280 x 1024 px, subtending visual angles of 22.7° (horizontal) by 17.0° (vertical) on the screen. The target prevalence in the stimuli set was 100%: all scenes contained a target and, importantly, the target location never changed in a particular scene. To make the task challenging, book targets were small (from 1 to 2°) and spatially distributed across the image periphery.

A total of 48 critical scenes were searched by observers in the Test phase, with scene familiarity counterbalanced across equal numbers of observers. In the Learning phase, other novel scenes were used as "filler" images interspersed with the repeated scenes and were viewed by all participants; these novel scenes were not directly relevant to the questions of interest.

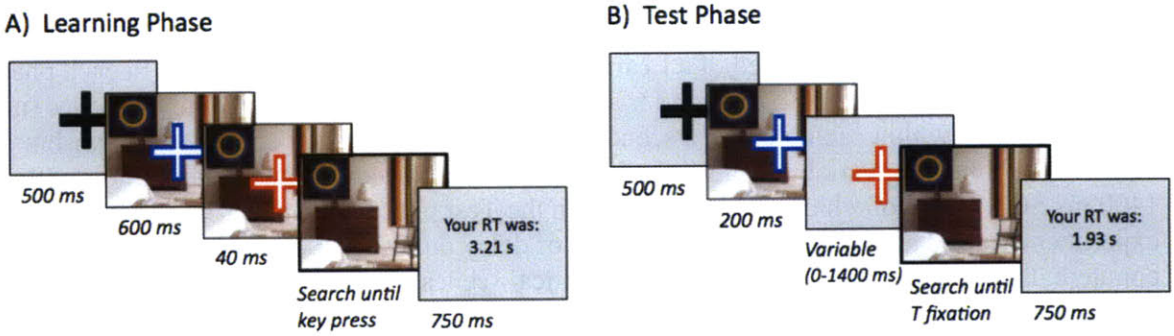
*Design.* The experiment consisted of a Learning phase followed by a Test phase. In each phase, observers performed 4 search blocks and trials of repeated scenes (n=24) and novel scenes (n=8) were intermixed within each block (32 scenes). The main hypothesis proposes a relationship between the following variables: scene familiarity (familiar or novel), and delay-duration (0-1400 ms SOA), both of which are manipulated in the Test phase. This experiment, unlike experiment 1, used a within-subject comparison of delay durations. Scene familiarity was counterbalanced across subjects: a total of 48 critical scenes were searched by were searched by equal numbers of participants: observers in group *A* searched scenes 1-24 repeatedly and searched scenes 25-48 only once, while the observers in group *B* searched scenes 25-48 repeatedly and scenes 1-24 once. In each block of the Test phase, 8 scenes (6 repeated, 2 novel) were tested at each delay-duration (0, 200, 600, 1400 ms).

*Procedure.* Observers were instructed, at the beginning of each phase, to find the book in each scene as quickly as possible. The purpose of the Learning phase was for participants to learn the location of a book in scenes that became familiar because they were searched once in each block. The purpose of the Test phase was to manipulate the amount of time between the *scene onset* (observers fixating centrally) and the *initial search fixation* on each scene; participants searched following a variable SOA (200, 400, 800, or 1600 ms) on a novel or familiar scene. Each phase was comprised of 4 search blocks: 24 repeated scenes and 8 novel scenes were randomly intermixed in each block (32 trials per block). Scenes were counterbalanced across observers with respect to the novel or repeated conditions.

The trial sequence was designed to be similar in Learning and Test phases, in order to habituate participants to the procedure of holding their gaze on a fixation cross. As shown in Figure 5, participants fixated a central fixation cross for 500 ms to begin the trial (gaze contingent). Next, the scene was presented with a blue fixation cross superimposed and participants were required to fixate the central cross for the duration of this interval (600 ms or 200 ms, Learning and Test phase respectively) without making a saccade, otherwise trial terminated. In the Learning phase, the fixation cross then briefly turned red (40 ms) and disappeared, signaling participants to

actively explored the scene to find the book. In the Test phase, the initial scene presentation (200 ms) was followed by a variable delay on a gray screen, giving an overall SOA (delay plus the initial presentation time) of 200 ms, 400 ms, 800 ms, or 1600 ms; the *same scene* was then presented again and participants moved their eyes to find the target. Participants had a maximum of 8 s to respond via key press (Learning phase) or by fixating the target for 750 ms (Test phase). Feedback was given after each trial (reaction time displayed for 750 ms) to encourage observers to search speedily throughout the experiment. Short mandatory breaks were enforced in order to avoid eye fatigue. The entire experiment lasted approximately 50 min. Eyetracker calibration and assessment was as described in experiment 1.

Figure 5. Trial sequence for each phase of Experiment 2. In the Learning Phase, participants learned the location of books in the repeated scenes. In the Test Phase, a scene (novel or repeated) was briefly shown, followed by a variable SOA, then participants searched until the target was fixated. The size of the fixation cross is exaggerated below for illustration: all fixation crosses were 2 x 2 degrees of visual angle.




## RESULTS

There are many measures of search performance that could be reported with these eye movement data, many of which are highly correlated (e.g. reaction time and number of fixations). Here, we report a set of measures that represent information about the time course of early and middle search processes, and a coarse spatial analysis of early eye movements. These measures include: *Reaction time*, defined as the entire time during which the scene was visible, from the offset of the variable-duration red fixation cross until the end of the trial; *Initial saccade latency*, which represents the duration of the initial, central fixation on the scene before making the first eye movement; *Average fixation duration*, which represents the average duration of **search** fixations in the scene- i.e. all fixations except the initial, central fixation and the fixations on the target; *Number of search fixations*; and *First fixation error*, the distance from the first search fixation to the center of a bounding-box around the target (in degrees of visual angle). We also report the proportion of trials (per condition) in which the following search behavior occurs: *initial saccade toward same-side as target*, *initial saccade lands on the target*, *second fixation lands on the target*, and the *third fixation lands on the target*. Our main hypothesis predicted that that longer delay-durations (SOAs) would enhance search performance in familiar scenes, owing to more effective retrieval of scene specific location priors, relative to searching novel scenes.

*Accuracy.* A search was scored as correct when an observer located the target-book in the scene before the maximum search time, 10 sec, elapsed. Observers failed to find the target in 4.6% of trials in the learning phase and 3.6% of trials in test phase. Trials in which a saccade was initiated before the delay duration elapsed or in which the eye movement signal was lost for at least 5 successive data points were removed. These criteria resulted in the removal of 2.1% and 1.8% of trials for the learning and test phases, respectively. Finally, trials in which the observer initiated a saccade before the entire SOA had elapsed were removed; this removed 8.8% and 4.2% of trials in the learning and test phases, respectively. Since the proposed hypotheses address the effect of SOA, the number of removed trials from each duration was compared (range: 3.4–5.1% of trials), and it was found that the number of trials did not significantly differ across level of SOA, $F(3,84) = 0.36$.

**Learning Phase Results**

The learning phase consisted of the first four search blocks, during which observers searched scenes in the repeated condition in each block, with the intent of training an association between specific scenes and the target's location in those scenes. The main hypothesis did not make predictions about the nature or speed of the learning process over the course of the learning blocks, only that learning would occur. To show a trajectory of improvement, therefore, the learning phase data are presented as a comparison between the average of blocks 1-2 ("early") and blocks 3-4 ("later") for each measure.

Reaction times on repeated scenes fell from 2222 ms to 1611 ms in early and late blocks, respectively ($t(21)=8.6$, $p < 0.0001$) and, correspondingly, the number of search fixations decreased from 4.5 to 3.1 fixations per trial ($t(21)=7.3$, $p < 0.0001$) over the learning phase. The initial saccade latency decreased from 276 ms to 226 ms ($t(21)=5.1$, $p < 0.0001$) and the average fixation duration fell from 214 ms to 188 ms ($t(21)=5.4$, $p < 0.0001$), but there was no difference in the first fixation error between early and late blocks (early *M*, 11.04 deg, late *M* 11.07 deg, $t(21)=0.16$).

Reaction times on novel "filler" scenes were faster overall (owing to less difficult stimuli) than repeated scenes and fell from 1501 to 1196 ms ($t(21)=6.6$, $p < 0.0001$), along with fewer search fixations ($t(21)=4.4$, $p < 0.001$). The initial fixation duration decreased from 253 ms to 210 ms ($t(21)=4.2$, $p < 0.001$), but the average fixation decrease remained unchanged ($t(21)=0.22$) across early and late search blocks. These novel scenes do not have direct relevance to the experimental hypothesis, as delay duration was not manipulated in this phase, but the results are reported here and may serve as a potential point of comparison against results in the Test phase.

**Test Phase Results**

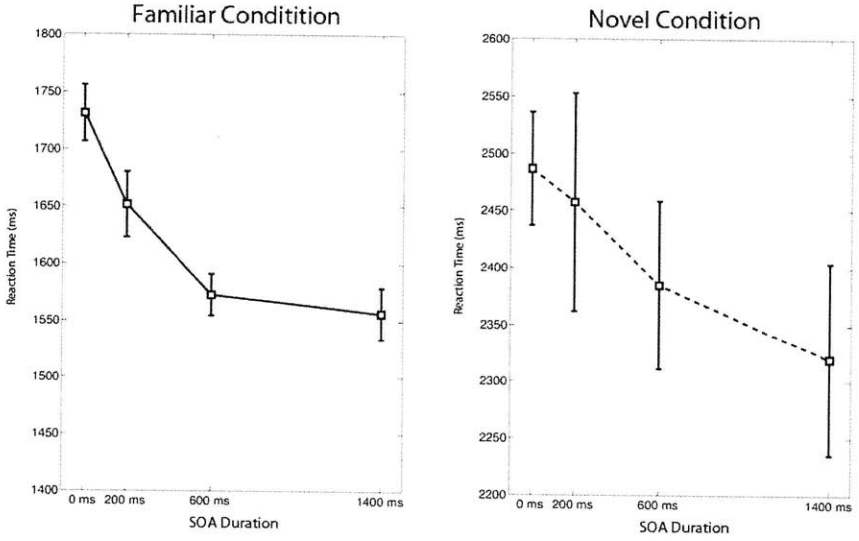
The test phase consisted of the last four search blocks, during which observers saw a brief preview of the scene and, critically, were delayed for a variable time interval (meanwhile fixating a blank gray screen). The main hypothesis, following up on the results of experiment 1, predicts that the longer delays will be associated with better search performance. Specifically, the delay-duration (SOA manipulation) was expected to have a disproportionate impact on search trials with repeated scenes than novel scenes, indicating the retrieval of specific scene associations.

*Reaction Time*

Our first main result, the mean reaction time (RT) as a function of SOA, is shown in Figure 6 below. As expected, there was a large main effect of scene familiarity ($F(1,42) = 8.5$, $p < 0.0001$), with novel scenes being searched considerably slower than familiar scenes. For familiar scenes, a one-way repeated measures ANOVA showed a significant effect of SOA, $F(3,63) = 8.5$, $p < 0.0001$, but not for the novel scenes ($F(3,63) = 0.70$). A mixed design ANOVA of the RTs from both conditions showed no significant interaction between the two variables ($F(3,126) = 0.10$). The trend in both conditions was for search to be speedier when it was preceded by a longer delay. Eye movement data allow us to isolate different contributions to the overall RT, making it possible to determine which SOA-driven effects were common to both familiar and novel scenes, and which were unique to each condition.

Overall reaction time reflects the entire time that the scene was visible and, behaviorally, includes the initial saccade duration, scan time (search fixations + saccades), and the gaze duration on the target. Since gaze duration in the Test phase was pre-determined by the procedure (gaze contingent trial termination by fixating the target for 750 ms), this measure is largely uninteresting, and the remainder of the results focus on the influence of scene familiarity and SOA on early and overt search processes.

Figure 6. Reaction time as a function of delay (SOA). Error bars represent within-observer standard error.




*Initial Saccade Latency*

The relationship between SOA and initial saccade latency is shown in Figure 7 (left). There was a significant main effect of SOA ($F(3,63) = 74.2$, $p < 0.0001$), and small effect of familiarity ($F(1,42) = 2.2\text{x}10^5$, $p < 0.0001$), likely driven by the value at 600 ms SOA. There was no interaction ($F(3,126) = 0.10$). Overall, longer SOAs in both novel and familiar scenes were associated with faster initial saccade latencies, pointing to a common basis underlying the RT results.

*Average Fixation Duration*

The average duration of fixations in a trial (excluding initial saccade latency and gaze duration on the target), is shown in Figure 7 (right). There was a main effect of scene familiarity ($F(1,42) = 2.1\text{x}10^6$, $p < 0.0001$), with fixations on novel scenes being about 30 ms slower than on familiar scenes. For novel scenes, there was a significant effect of SOA, $F(3,63) = 3.2$, $p < 0.03$, but not for the familiar scenes ($F(3,63) = 1.0$). A significant interaction between the two variables ($F(3,126) = 2.7$, $p < 0.05$) suggests that SOA duration influenced the conditions differently: in novel scenes, longer SOAs were associated with faster fixations, but not in familiar scenes, potentially due to a floor effect (*M* 180 ms).

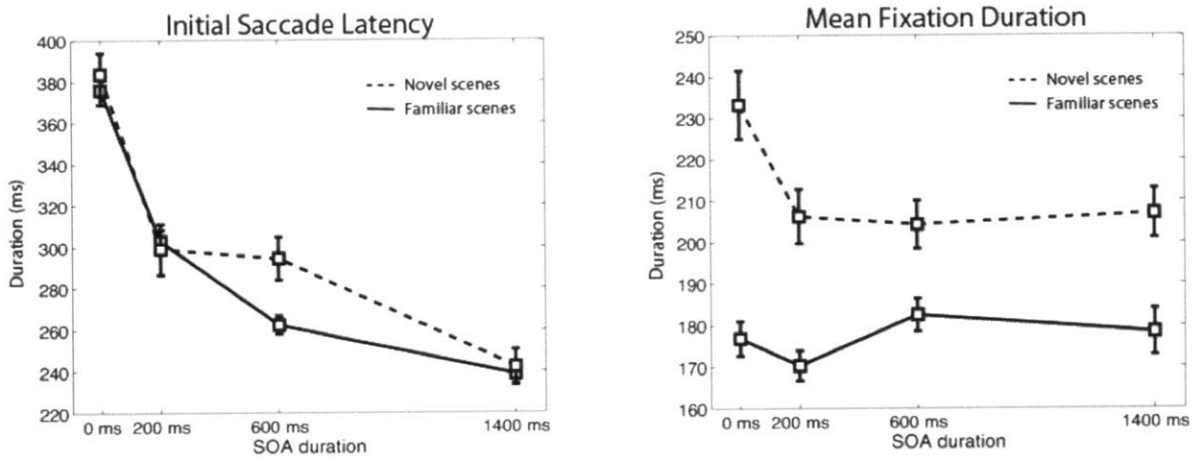



Figure 7. Influence of SOA on fixation durations. Error bars represent within-observer standard error

*Number of search fixations*

Although there is a slight trend for the number of search fixations to decrease in familiar scenes- from 2.21 (0 ms SOA) to 2.05 (1400 ms SOA)- there was not a significant effect of SOA for familiar ($F(3,63) = 1.3$), or for novel scenes ($F(3,63) = 0.51$), and no significant interaction ($F(3,126) = 0.47$).

*Proportion of Trials with Different Search Outcomes*

On any given trial in which the target was found (>96% of trials in Test phase), the outcome will be one of the following: the initial saccade could land on the target, the 2nd fixation could land on the target, the 3rd fixation could land on the target, etc (the maximum number of search fixations in a trial was 18). Table 4 shows the proportion of trials in each condition exhibiting these search outcomes: target localization on the first, second, or third fixation of the trial. The main hypothesis predicted that longer SOAs would facilitate retrieval of scene location priors, increasing the likelihood of directing attention to the target. This was assessed by evaluating how often the **initial saccade** was directed to the target (e.g. Peterson & Kramer, 2001), shown in Figure 8.

There was a significant main effect of scene familiarity ($F(1,42) = 26.5$, $p < 0.0001$), a main effect of SOA for familiar scenes ($F(3,63) = 9.6$, $p < 0.0001$) but not novel scenes ($F(3,63) = 0.98$). Importantly, there was a significant interaction between the two conditions ($F(3,126) =$

3.6, $p < 0.01$). Longer time intervals seem to make it more likely that the first saccade in a familiar scene will land on the target region.

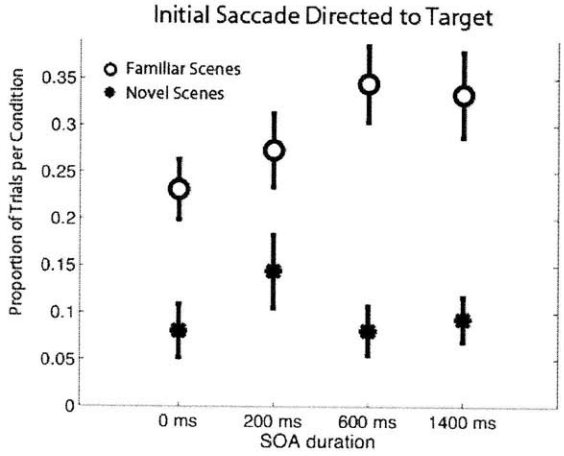



Figure 8. Proportion of trials in which the first fixation landed on the target

When the target was *not* localized on the initial fixation, what proportion of the time was the target localized on the **subsequent** (second) fixation? On average, this was a more common outcome when the scenes were familiar (39%) versus novel (14%), collapsed across SOAs. The effect of scene familiarity was significant ($F(1,42) = 38.1$, $p < 0.0001$), however there was no effect of SOA in familiar scenes ($F(3,63) = 0.46$) or in novel scenes ($F(3,63) = 0.14$), and no interaction ($F(3,126) = 0.46$).

Furthermore, when the target was not localized on the first or second fixations, what proportion of the time was the target localized on the third fixation? Again, this outcome was more common when the scenes were familiar (36%) versus novel (20%), averaged across SOAs. The effect of scene familiarity was significant ($F(1,42) = 27.8$, $p < 0.0001$), but there was no effect of SOA duration (familiar: $F(3,63) = 0.06$; novel: $F(3,63) = 1.2$), and no interaction ($F(3,126) = 0.52$).

*Spatial Characteristics of First Search Fixation*

Since none of the previous measures have addressed how SOA influences the *placement* of search fixations, we report two measures that evaluate spatial characteristics of the initial eye movement away from the center of the screen: first fixation error (in degrees of visual angle) and the proportion of trials in which the initial saccade is toward the same-side as the target.

Excluding trials in which the initial saccade landed on the target (see Table 4), the first fixation error was not significantly influenced by SOA in familiar scenes ($F(3,63) = 0.93$) or novel scenes ($F(3,63) = 1.1$). We also found that the error was smaller for familiar scenes than novel scenes ($F(1,42) = 2.5 \times 10^6$, $p < 0.0001$). Values are reported below in Table 3.

Since scene familiarity seemed to influence the placement of the first fixation, though not in an SOA-dependent manner, we analyzed a more spatially coarse measure of early search guidance: the proportion of trials with an initial saccade toward the same side as the target, regardless of whether the fixation landed on the target. Interestingly, these results indicated that longer SOAs increased the likelihood of making a saccade toward the target-side on familiar scenes ($F(3,63) =$

3.7, $p < 0.01$) but not novel scenes ($F(3,63) = 1.3$), and the interaction was marginally significant ($F(3,126) = 2.4$, $p = 0.07$). Values are reported below in Table 4.

Table 3

*Summary of eye movement measures*

| | **Familiar Condition** | | | | | **Novel Condition** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *0 ms* | *200 ms* | *600 ms* | *1400 ms* | *Time Difference* | *0 ms* | *200 ms* | *600 ms* | *1400 ms* | *Time Difference* |
| Reaction Time (ms) *M* | 1732 | 1652 | 1573 | 1556 | 175 ms | 2487 | 2458 | 2385 | 2320 | 167 ms |
| Num. Seach Fixations *M* | 2.21 | 2.11 | 1.99 | 2.05 | 0.16 fixations | 4.37 | 4.88 | 4.51 | 4.43 | 0.04 fixations |
| Init. Sac. Latency (ms) *M* | 376 | 303 | 262 | 239 | 137 ms | 383 | 299 | 295 | 242 | 142 ms |
| Fixation Duration (ms) *M* | 177 | 170 | 183 | 178 | -1 ms | 233 | 206 | 204 | 207 | 26 ms |
| First Fixation Error *M* | 10.8° | 10.8° | 10.7° | 10.4° | 0.4° | 11.7° | 11.8° | 11.6° | 11.4° | 0.3° |

Table 4

*Spatial Characteristics of Search Behavior for different delay durations*

| | Proportion of Trials: **Familiar Scenes** | | | | Proportion of Trials: **Novel Scenes** | | | |
|---|---|---|---|---|---|---|---|---|
| | *0 ms* | *200 ms* | *600 ms* | *1400 ms* | *0 ms* | *200 ms* | *600 ms* | *1400 ms* |
| **Initial Fixation on Target** (n search fix. = 0) | 0.23 | 0.27 | 0.35 | 0.34 | 0.08 | 0.14 | 0.08 | 0.09 |
| **2nd Fixation on Target** (n search fix. = 1) | 0.27 | 0.27 | 0.21 | 0.24 | 0.12 | 0.11 | 0.14 | 0.13 |
| **3rd Fixation on Target** (n search fix. = 2) | 0.17 | 0.14 | 0.15 | 0.13 | 0.19 | 0.12 | 0.17 | 0.13 |
| *SUM* | ***0.67*** | ***0.68*** | ***0.71*** | ***0.71*** | ***0.39*** | ***0.37*** | ***0.39*** | ***0.35*** |
| **Initial Saccade goes to same-side as Target** | 0.68 | 0.73 | 0.71 | 0.77 | 0.56 | 0.59 | 0.49 | 0.51 |

*SUMMARY*

The results, overall, were consistent with the notion that longer SOAs improve search performance by enhancing retrieval of scene specific location priors. We also observed some surprising results which potentially arose from the unique experimental paradigm. Four main results were reported:

(1) Reaction time in both familiar and novel conditions decreased as a function of SOA. While familiar scenes were searched much more quickly than novel ones, the magnitude of decrease was approximately the same in each condition (Figure 6, Table 3).

(2) The decrease in RT was, in part, based on a common trend for the initial saccade latency to decrease with longer SOAs in novel and familiar scenes; scene familiarity had no effect on the speed of launching the initial saccade. Surprisingly, scene familiarity had a significant effect on average fixation duration; in familiar scenes, fixation dwell-time was fast (about 180 ms) and unaffected by SOA, whereas in novel scenes, fixations were slower but became faster with longer SOAs (Figure 7, Table 3).

(3) The probability of making a saccade directly to the target increased as a function of longer SOA durations when the scene was familiar but not in novel scenes. This serves as the primary support for the experimental hypothesis (Figure 8, Table 4).

(4) A spatial characterization of the first search fixation suggested that scene familiarity exerts, at most, a coarse influence in early saccade guidance. When the initial saccade failed to land on the target, there was no evidence that guidance was more accurate with longer SOAs (Table 3). When looking at all trials, however, there is evidence that longer SOAs on familiar scenes were associated with greater probability of directing the initial saccade to the same-side as the target, relative to novel scenes (Table 4).

**INTERPRETATION**

The main hypothesis predicted that search performance would be systematically facilitated by longer delay durations when searching familiar scenes. Critically, comparing search behavior between the novel and familiar conditions allowed us to assess the influence of the variable delay (SOA manipulation) independently from the influence of familiarity. Reaction time, as expected, became faster with longer SOAs, although the reasons underlying the decrease were not expected. Most of the RT decrease resulted from an unexpected impact of SOA on initial saccade latency, regardless of scene familiarity; in retrospect, it is not surprising that observers would be faster at initiating an eye movement after holding central fixation on a blank gray screen for a longer time. Given this clear impact of the SOA manipulation, however, it is perhaps surprising that scene familiarity had such little effect. In contrast, search fixations in familiar scenes were consistently of shorter duration than novel scenes. This pattern of results- fewer, consistently fast fixations in familiar scenes, along with a greater number of fixations in novel scenes that became faster with longer SOAs- accounts for the pattern of RT results.

Interestingly, our study largely replicates the findings of the Peterson and Kramer study (2001) that also investigated recognition and guidance in contextual cuing (letter array stimuli). We found a significant interaction between delay and scene familiarity on the proportion of trials in which the initial saccade landed on the target. This is interpreted as evidence that longer time

intervals increase the likelihood of retrieving specific scene priors from long term memory. If the initial saccade was **not** directed to the target, however, there was no tendency for the first fixation to be more accurate after longer delays. Unlike Peterson and Kramer (2001), the first fixation error was slightly smaller on familiar scenes than novel ones. This is perhaps not surprising, given the nature of our real world scene stimuli in which attentional guidance can exploit a rich variety of visual and semantic scene features.

## General Discussion

Using scene specific search priors involves a combination of recognition and attentional guidance processes (Peterson & Kramer, 2001). The experiments in this chapter investigated temporal characteristics of how these learned priors guide search in familiar real world scenes. Whereas the contextual cuing phenomenon seems to be based on predominantly implicit recognition processes, it less clear (1) how long-term memory representations of real world contexts support attentional guidance and (2) how conscious awareness is involved. To the first issue, if observers' search speed was indeed improved by retrieving learned scene-target associations, what was the mechanism used to improve guidance?

This question has been investigated in the repeated search experiments of Wolfe and colleagues (Wolfe et al, 2000) and their classic finding that attentional guidance does not improve after hundreds of searches through the same, unchanging display. Despite observer's clear memory of the search array, it is interesting that novel displays were searched as efficiently as the well-rehearsed displays. At face value, this seems to suggest that visual attention may not recruit information from long-term memory. Kunar, Flusberg, and Wolfe (2008a) recently studied this issue in more detail by matching response-types in the memory and visual search conditions. Critically, the search was also modified such that targets only appeared in a *subset* of the possible locations (experiments 5, 6). Under these conditions, observers indeed learned to guide attention to specific locations and away from irrelevant items in the display. Within the relevant subset of locations, however, search remained inefficient (Kunar et al, 2008a). Memory improved search performance, the authors conclude, by *reducing the effective set size* rather than changing the nature of the search. Our experiments, similarly, allowed observers to learn that a scene's identity predicted a *single* target location and that other scene regions could effectively be ignored. Correspondingly, participants located targets in familiar scenes with faster reaction times and fewer fixations than novel scenes (experiments 1, 2). In this respect, our findings are consistent with the proposal that learning improved search efficiency by guiding attention to a subset of the scene.

This still leaves open the question: how do the recognition processes triggered by a familiar context ultimately bias attention toward a subset of informative scene regions? This chapter showed that longer time intervals between scene presentation and the initial saccade were associated with more direct scan paths (experiment 1) and a higher probability of making saccades directly to the target or the same-side of the scene as the target (experiment 2). These results are consistent with a recent study from Kunar, Flusberg, and Wolfe (2008b) in which the authors used two approaches to lengthen the time before a response in a contextual cuing task. In one approach, increasing the display complexity resulted in a greater effect of context-based

guidance (Kunar et al, 2006b, experiment 1); notably, the average RTs for the difficult conditions in that study were between 1-1.5 seconds, which are comparable to the search times in this chapter's experiments. A second approach was to introduce a delay between the onset of an initial display and the final search stimulus; with or without place-holders visible during the delay (where items would later appear). Results again showed that the additional time allowed participants to implement memory-based guidance, provided that there was something in the place-holder locations to guide attention to (Kunar et al, 2006b, experiment 2). In a behavioral study that accompanied the fMRI experiment, Summerfield and colleagues (2006) manipulated SOA (100 ms, 500 ms, or 900 ms) along with memory versus visual orienting. Similarly, they found a strong effect of SOA- the fastest response times occurred after a 900 ms SOA and slowest response times after a 100 ms SOA- however there was little evidence that the delay was more effective in the memory-guided condition than the visually-guided condition (Summerfield et al, 2006). Overall, this work suggests that the conditions in which specific contexts guide attention tend to involve prolonging the time before a response.

Potentially, the time needed for memory to guide attention arises from the time course of recognition processes. Memory retrieval may be considered a bottleneck limiting what information from long-term memory gets passed along to attentional guidance mechanisms. In fact, converging cognitive neuroscience evidence (for reviews see Buckner & Wheeler, 2001; Rugg & Yonelinas, 2003; Yonelinas, Otten, Shaw, Rugg, 2005) points to at least two temporally distinct components operating in recognition memory: a rapidly available familiarity process and a slower recollective process (Mandler, 1980). Subjectively, these two processes can be experienced, for example, when seeing a photograph from a past vacation and having an immediate sense of *familiarity* before fully recollecting the exact identity of the place and its contextual associations. Recordings of event-related brain potentials (ERPs) during memory tasks find temporally, topographically, and functionally distinct correlates of familiarity and recall during retrieval (Rugg & Henson, 2002). Although the term "familiar condition" has been used throughout this chapter, this was not intended to imply that scene specific priors rely primarily on "familiarity" as opposed to "recollective" processes. In fact, it seems reasonable that scene specific location priors, being a type of contextual association, would involve a slow time course mediating this form of retrieval.

A key caveat to the above proposal was illustrated by the results in the final block of experiment 1. Observers in the extended-delay group performed 8 search blocks with a 1300 ms delay at central fixation; on the final switch block, the delay was shortened to only 300 ms before starting overt search. Surprisingly, these observers had fast search RTs and made equally efficient scan paths, despite the shorter time interval. What does this imply about the retrieval-time account of memory based search guidance? This finding does not preclude the idea that the time course of retrieval processes *often* limits the extent of attentional guidance in familiar contexts. However, it does point to the existence of multiple types of learning and memory mechanisms, some of which vary in the degree of transfer across temporal and spatial transformations. Automaticity is one classic example, in which a specific, overlearned situation rapidly triggers a well rehearsed motor plan with low-demands on cognitive resources but little flexibility (Logan, 1992). An interesting distinction in the attentional cuing literature is that unconscious cues can drive extremely rapid attentional shifts (Posner, 1980; Yantis & Jonides,

1984) while volitional shifts of attention are considerably slower (Muller, Teder-Salejarvi, Hillyard, 1998; Wolfe, Alvarez, & Horowitz, 2000).

To be fair, it is likely that both rapidly and slowly developing contributions from memory factor into the results of the current studies. The bulk of evidence nonetheless indicates that scene specific location priors require longer time intervals to influence attentional mechanisms. In the results of experiment 2, for example, we observed that the earliest available measure of search performance- the initial saccade latency- was indistinguishable between scene familiarity conditions; only slightly later in the trial, search fixations tended to be approximately 30 ms shorter in familiar scenes. More importantly, delay duration was positively correlated with the proportion of initial eye movements directed to the target in familiar, but not novel, scenes. These data, along with recent ERP (Johnson et al, 2007) and eye tracking (Peterson & Kramer, 2001) results, suggest that long-term memory retrieval helps search performance by increasing the likelihood that one of the first few shifts of attention will be to the target.

This mechanism describes how scene specific location priors, as studied in contextual cuing work and the present chapter's experiments, in effect bias visual processing on any given trial. This is supported by the eye movement data of Peterson & Kramer (2001), in which recognition of the repeated contexts sometimes occurred immediately- indicated by an initial saccade directed to the target- but, more often, occurred after an overt search had begun. The main result of chapter 2, furthermore, demonstrated that person-specific, scene specific search experience systematically influenced fixation locations on a given (repeated) search trial. Electrophysiological measures can provide reliable information about the timing associated with particular types of cognitive processing, such as attentional selection. However since waveforms are averaged across many trials, they do not generally inform about individual trials. A recent ERP study of contextual cuing found that brain responses on repeated displays gave a higher amplitude in the early portion of the N2pc waveform, and a greater proportion of fast RTs, relative to novel displays (Johnson et al, 2007). The authors argue that the increased amplitude in the early part of the waveform resulted from averaging over more trials in which attention was quickly shifted to the target, similar to increasing the early portion of the RT probability distribution (Johnson et al, 2007). If so, this would support the idea that searching a familiar context increases the likelihood that early shifts of attention land on the target. The experiments in this chapter extend this argument by showing that memory for specific scenes effectively orients visual attention and, furthermore, that retrieval becomes increasingly more likely to guide search over time.

At present, the nature of recognition processes in visual search of cluttered, real world scenes are not sufficiently understood to characterize memory influences as wholly implicit or explicit, or based on exclusively one type of retrieval. One possibility is that a degree of explicit memory encoding and retrieval comes "for free" when the search display is a real world scene. When real world scenes were used as backgrounds in contextual cuing (Brockmole, Castelhano, & Henderson, 2006; Brockmole & Henderson, 2006a, 2006b), observers explicitly recognized repeated scenes from novel ones (Brockmole & Henderson, 2006a). More importantly, when observers searched *upright* real world scenes, the scene-target associations were learned twice as rapidly as when the same scenes were inverted. Consistent with this finding, a semantically

meaningful scene may serve as a richly informative basis upon which specific associations are readily learned (e.g. scene location priors).

## Concluding Remarks

Given the ubiquitous contextual associations in daily life, it is prudent to study how memory retrieval and visual attention interact in familiar environments. The main idea of this chapter is that **time** increases the likelihood that specific real world scenes will be recognized and that those spatial cues will help guide attention to the target. We introduced a *delayed-search* paradigm to manipulate the interval between scene context presentation and oculomotor responses. Overall measures of search performance indicated that longer delays were associated with more rapid attentional deployment to the target. Scene specific search priors, specifically, influenced the duration and placement of early attentional shifts in familiar scenes. In summary, this work is interpreted as evidence for temporally graded memory retrieval that governs use of context-based guidance.

## References

Aguirre, G.K., Detre, J.A., Alsop, D.C., & D'Esposito, M. (1996). The parahippocampus subserves topographical learning in man. *Cerebral Cortex*, 6, 823-829.

Araujo, C., Kowler, E., & Pavel, M. (2001). Eye movements during visual search: The cost of choosing the optimal path. *Vision Research*, 41, 3613-3625.

Bar M., Aminoff E., & Schacter D.L. (2008). Scenes unseen: the parahippocampal cortex intrinsically subserved contextual associations, not scenes or places per se. *Journal of Neuroscience*, 28, 8539-8544.

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

Brockmole J.M., Castelhano, M.S., & Henderson, J.M. (2006). Contextual cueing in naturalistic scenes: Global and local contexts. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 32(4), 699-706.

Brockmole, J.R., & Henderson, J.M. (2006a). Using real-world scenes as contextual cues for search. *Visual Cognition*, 13, 99-108.

Brockmole, J.R., & Henderson, J.M. (2006b). Recognition and attention guidance during contextual cuing in real-world scenes: Evidence from eye movements. *The Quarterly Journal of Experimental Psychology*, 59, 1177-1187.

Buckner R.L. & Wheeler M.E. (2001). The Cognitive Neuroscience of Remembering. *Nature Reviews Neuroscience*, 2, 624-634.

Burgess, N., Maguire, E.A., & O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron* 35, 625-641.

Carrasco, M., & McElree, B. (2001). Covert attention accelerates the rate of visual information processing. *Proceedings of the National Academy of Sciences*, 98, 5363-5367.

Chun, M.M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36,* 28-71

Chun, M.M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, 10, 360-365.

Chun, M.M., & Phelps, E.A. (1999). Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nature neuroscience*, 2(9), 844-847.

Corbetta, M., & Shulman, G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3, 201-215.

Corbett, A.T., & Wickelgren, W.A. (1978). Semantic memory retrieval: Analysis by speed accuracy tradeoff functions. *The Quarterly Journal of Experimental Psychology*, 30, 1-15.

Dosher, B.A. (1976). The retrieval of sentences from memory: A speed-accuracy study. *Cognitive Psychology*, 8, 291–310.

Dosher, B.A. (1981). The effect of delay and interference: A speed-accuracy study. *Cognitive Psychology*, 13, 551–582.

Dosher, B.A., Han, S., & Lu, Z.-L. (2004). Parallel processing in visual search asymmetry. *Journal of Experimental Psychology: Human Perception & Performance*, 30(1), 3-27.

Eichenbaum, H. (1997). Declarative memory: insights from cognitive neurobiology. *Annual Reviews in Psychology*, 48, 547-572.

Eichenbaum, H. (2004). Hippocampus: cognitive processes and neural representations that underlie declarative memory. *Neuron* 44, 109-120.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392, 598-601.

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

Gabrieli, J.D. (1998). Cognitive neuroscience of human memory. *Annual Reviews in Psychology*, 49, 87-115.

Henderson, J.M., & Hollingsworth, A. (1998). *Eye movements during scene viewing: An overview*, in: Eye Guidance While Reading and While Watching Dynamic Scenes, Underwood, G. (Ed.), pp. 269–293. Elsevier Science, Amsterdam.

Henderson, J.M., Weeks, P.A., Jr., & Hollingworth, A. (1999). Effects of semantic consistency on eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception & Performance*, 25, 210-228.

Hidalgo-Sotelo, B., Oliva, A., & Torralba, A. (2005). Human Learning of Contextual Object Priors: Where does the time go? *Proceedings of the IEEE Computer Society Conference on CVPR* (pp. 510-516).

Johnson, J.S., Woodman, G.F., Braun, E., & Luck, S.J. (2007). Implicit memory influences the allocation of attention in visual cortex. *Psychonomic Bulletin & Review*, 14, 834-839.

Kastner, S., & Ungerleider, L.G. (2000). Mechanisms of visual attention in the human cortex. *Annual Reviews of Neuroscience*, 23, 315-341.

Kunar M.A., Flusberg S.F., Horowitz T. S., & Wolfe J.M., (2007). Does contextual cuing guide the deployment of attention. *Journal of Experimental Psychology: Human Perception and Performance*, 33(4), 816-828.

Kunar, M.A., Flusberg, S.J., & Wolfe, J.M. (2006). Contextual cuing by global features. *Perception & Psychophysics*, 68, 1204-1216.

Kunar, M.A., Flusberg, S.J., & Wolfe, J.M. (2008a). The role of memory and restricted context in repeated visual search. *Perception and Psychophysics*, 70, 314-328.

Kunar, M.A., Flusberg, S.J., & Wolfe, J.M. (2008b). Time to Guide: Evidence for Delayed Attentional Guidance in Contextual Cueing. *Visual Cognition*, 16, 804-825.

Kunar, M.A., Michod, K.O., & Wolfe, J.M. (2005). When we use the context in contextual cuing: Evidence from multiple target locations. Journal of Vision, 5(8), 412a.

Logan, G.D. (1992). Attention and preattention in theories of automaticity. *American Journal of Psychology*, 105, 317-339.

Luck, S.J., Girelli, M., McDermott, M.T., & Ford, M.A. (1997). Bridging the gap between monkey neurophysiology and human perception: An ambiguity resolution theory of visual selective attention. *Cognitive Psychology*, 33, 64-87.

Mandler G., (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, 87, 252-271.

McElree, B., & Carrasco, M. (1999). Temporal dynamics of visual search: A speed-accuracy analysis of feature and conjunction searches. *Journal of Experimental Psychology: Human Perception & Performance*, 25, 1517-1539.

Moores, E., Laiti, L. & Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. *Nature neuroscience*, 6, 182–189.

Muller, M. M., Teder-Salejarvi, W., & Hillyard, S. A. (1998). The time course of cortical facilitation during cued shifts of spatial attention. *Nature Neuroscience*, 1, 631-634.

Oliva A., Wolfe J.M., & Arsenio H.C. (2004). Panoramic Search: The Interaction of Memory and Vision in Search Through a Familiar Scene. *Journal of Experimental Psychology: Human Perception & Performance*, 30(6), 1132-1146.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10, 437-442.

Peterson, M. S., & Kramer, A. F. (2001). Attentional guidance of the eyes by contextual information and abrupt onsets. *Perception & Psychophysics*, 63, 1239-1249.

Posner, M.I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, 32, 3-25.

Reed, A.V., (1973). Speed-accuracy trade-off in recognition memory. *Science*, 181, 574-578.

Rugg, M.D. & Henson R.N.A (2002). Episodic memory retrieval: an (event-related) functional neuroimaging perspective. In: A.E. Parker, E.L. Wilding, & E.L. Bussey (Eds.), *The cognitive neuroscience of memory encoding and retrieval.* Hove, UK: Psychology Press.

Rugg, M.D. & Yonelinas A.P. (2003). Human recognition memory: a cognitive neuroscience perspective. *Trends in Cognitive Science*, 7, 313-319.

Ruthruff, E. (1996). A test of the deadline-model for speed-accuracy tradeoffs. *Perception & Psychophysics*, 58(1), 56-64.

Ryan, J. D., Althoff, R. R., Whitlow, S., & Cohen, N. J. (2000). Amnesia is a deficit in relational memory. *Psychological Science,* 11, 454-461.

Squire, L. R., Knowlton, B., & Musen, G. (1993). The structure and organization of memory. *Annual Reviews in Psychology*, 44, 453-495.

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science*, 253, 1380-1386.

Summerfield, J.J., Lepsien, J., Gitelman, D.R., Mesulam, M.M., & Nobre, A.C. (2006). Orienting attention based on long-term memory experience. *Neuron*, 49, 905-916.

Toth, J.P. (1996). Conceptual automaticity in recognition memory: Levels-of-processing effects on familiarity. *Canadian Journal of Experimental Psychology*, 50, 123-138.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science*, 247, 301-306.

Wolfe, J.M., Alvarez, G.A., & Horowitz, T.S. (2000). Attention is fast but volition is slow. *Nature*, 406(6797), 691.

Wolfe, J.M., Klempen, N., & Dahlen, K. (2000). Postattentive vision. *Journal of Experimental Psychology: Human Perception & Performance*, 26, 693-716.

Wolfe, J.M., Cave, K.R., & Franzel, S.L. (1989). Guided Search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception & Performance*, 15(3), 419-433.

Yantis, S., Schwarzbach, J., Serences, J.T., Carlson, R.L., Steinmetz, M.A., Pekar, J.J., & Courtney, S.M. (2002). Transient neural activity in human parietal cortex during spatial attention shifts. *Nature Neuroscience*, 5, 995-1002.

Yantis, S., & Jonides, J. (1984). Abrupt visual onsets and selective attention: Evidence from visual search. *Journal of Experimental Psychology: Human Perception & Performance*, 10, 601-621.

Yonelinas A.P., Otten L.J., Shaw K.N., & Rugg M.D., (2005). Separating the Brain Regions Involved in Recollection and Familiarity in Recognition Memory, *The Journal of Neuroscience*, 25, 3002-3008.

CHAPTER 5

# Conclusion

The work presented in this thesis provides one approach toward investigating how eye guidance in familiar scenes is distinct from general contextual guidance. Intuitively, it seems reasonable that repeatedly searching an environment will produce an attentional bias toward regions that have been informative in the past. However, studying the influence of scene specific experience is complicated by the fact that, on one hand, fixation selection varies significantly across observers but, on the other hand, even novel real world scenes often constrain which scene regions are likely to contain a target. Three chapters investigated the role of memory-based eye guidance when observers look for an object (e.g. pedestrian, book) in novel and familiar real world scenes.

The behavioral experiment presented in Chapter 2 was important because, firstly, it provided a benchmark of inter-observer fixation agreement that was used to evaluate model performance. Secondly, the eye movements collected in this study comprised the $F_{novel}$ population in comparative map analysis (Chapter 3, experiments 2 and 3). One of the main results of Chapter 2 indicated that a computational model of general contextual location priors (Torralba, Oliva, Castelhano, & Henderson, 2006) outperformed models of target-features (Dalal & Triggs, 2005) and visual salience. It is important to point out that this result was dependent on the nature of the search task and stimuli: all 456 target-present scenes contained a person in the original photograph, therefore targets were overwhelmingly located in context-consistent regions (e.g. on sidewalks). Consequently, even on target absent images, observers fixated scene regions where a pedestrian would typically be located. If the experiment had included a proportion of scenes with a target-person in *unexpected* locations (e.g. on rooftops), other sources of guidance such salience or target features may have emerged as a better predictor of eye movement than scene context.

Combining all three sources of guidance provided the most accurate model-based predictions of eye movements in Chapter 2. Still, using other observer's fixations to build a prediction map generated the most spatially precise predictions. The signal detection metric used to evaluate prediction accuracy, the ROC area (Green & Swets, 1966), is becoming an increasingly popular approach for comparing model predictions against human eye movement data (e.g. Renninger, Verghese, & Coughlan, 2007; Tatler, Baddeley, & Gilchrist, 2005). Rather than choosing a particular threshold (e.g. 10% of image area), the the ROC area summarizes performance across all possible thresholds (0-100%). For this reason, Chapter 3 also used the ROC area to evaluate how well a familiar observer's eye movements were predicted by fixations representing guidance by the same person, the same past history, or the same scene.

These spatial patterns in eye movements were investigated in Chapter 3 using comparative map analysis. Conceptually, this approach is similar to the cross-image control (Chapter 1), in which empirical fixations sampled from random scenes were used as a baseline for comparing inter-observer agreement. Tatler and Vincent (2009) also applied a similar logic in their recent study of oculomotor biases and whether they predict fixation selection independently of a scene's

content. Extending this logic further, comparative map analysis identifies several populations of fixations that share underlying information. The most specific information, $\mathbf{F_{self}}$, represents the regions that were fixated by a single person's repeated searches of one specific scene. Other scene dependent fixation populations include $\mathbf{F_{group}}$ (from other observer's repeated search of the scene) and $\mathbf{F_{novel}}$ (from observers who searched the scene only once). This chapter described the logic and implementation of comparative map analysis and, critically, demonstrated that the analysis of three experiments produced a similar pattern of results.

Comparative map analysis revealed, surprisingly, that fixation selection was systematically influenced by person-specific, scene-specific experience, beyond the guidance provided by (general) scene context alone. In three different search experiments, an observer's *own* fixations on a familiar scene provided significantly more accurate predictions of fixation locations than same-group and same-scene controls. This is interesting because it raises questions about whether oculomotor guidance from a person's specific experience in a scene has any behavioral significance. Does gaze deployment in familiar scenes boost memory by reinstating the "context" (fixation locations) of a previous search? Or does this effect arise as an incidental consequence of repeatedly viewing the same scene? If this was true, it might imply that repeated scene viewings with a different task (e.g. free viewing) would provide as much information for predicting scene fixations as we observed in these visual search experiments.

In 1971, Noton and Stark reported a similar finding in their study of pattern perception and memory: upon repeated exposure, an observer's first few eye movements reinstated a sequential "scanpath" specific to that person's viewing of a particular pattern. Controversially, they posited that this was indicative of how people remember and recognize patterns universally. Subsequent studies, however, failed to find a strong connection between sequential reinstatement of eye movements and memory. The authors themselves noted that people did not generate scan paths *obligatorily* when presented with a familiar pattern, despite the fact that participants easily recognized the familiar patterns (Noton & Stark, 1971). Correspondingly, chapter 3's finding of systematic individual differences in fixation selection may not, in fact, be indicative of a functional relationship between memory and eye movements.

To have a better understanding of this phenomenon, it would be helpful to examine patterns across observers and scenes. Are certain places searched more consistently than others? Scenes vary in the quantity of surfaces likely to contain the target and the spatial distribution of those surfaces. Looking for a book in an office, for example, is likely to require searching a greater proportion of the scene than searching a bathroom. Do constraints of general scene context affect how strongly observers are biased by their history of fixation selection in the scene? Scenes with less-constrained search area (i.e. more office-like than bathroom-like), in the least, may promote more variability in the number of different regions fixated by different observers. Under these conditions, comparative map analysis is more likely to reveal a difference between the $F_{group}$ and $F_{self}$ populations. This consideration should be taken into account when selecting scene stimuli for future experiments. On the whole, identifying properties of the scene and task that promote within-observer consistency in fixation selection remains an open question.

Finally, chapter 4 investigated the time course of using scene specific location priors to guide search in real world scenes. Recent work has suggested that with additional time, observers may

be able to effectively reduce the set size in a familiar scene by directing attention to regions that have been informative in the past (Kunar, Flusberg, & Wolfe, 2008). In our experiments, we used eye tracking to accomplish three things: (1) experimentally manipulate the duration of scene exposure before initiating overt search, (2) serve as a way for observers to respond when they located the object (trials terminated with a prolonged target fixation), and (3) provide a dependent measure which scene locations were foveated during the exploration phase of search. Overall, the two experiments in this chapter provide evidence that scene specific location priors were used more effectively when observers had more time for retrieval before making eye movements in familiar scenes. A greater proportion of initial saccades landed directly on the target when observers were given a longer delay on a familiar scene (experiment 2).

Altogether, the experimental results described in this thesis describe a unique role of scene specific search experience on the deployment of eye movements. There are numerous reasons to be cautious, however, about how broadly to interpret these results. One limitation of this work is that it does not explore the case in which a specific scene is associated with *multiple locations* in which targets are likely to appear. Learning that a single location is likely to contain a target- as was the case in the experiments in this thesis- is applicable only within a limited number of settings. In the real world, observers learn that objects can be in one of several places within a familiar environment. How does spatial uncertainty about a target's location influence how scene specific location priors are used? If a particular scene is associated with two locations, would we still find evidence for scene-specific guidance over repeated searches? Would retrieval time influence how attention was deployed to multiple informative locations?

As addressed earlier in the conclusion, these studied tested learning in which an object was located in contextually-consistent locations within a scene. Objects in the world, however, are sometimes found in unexpected locations. It would be interesting to know how general scene context interacts with scene specific learning. Potentially, objects in contextually-inconsistent location may be treated as distinctive in memory. If so, we might expect that scene specific location priors would be learned more quickly for scenes in which targets were located in unusual places than when targets were in contextually appropriate places. Furthermore, how would the distinctiveness of this association influence the amount of retrieval time needed to use memory to guide attention in the scene?

In this body of work, I have demonstrated a few approaches to studying how memory for specific, familiar scenes can influence the deployment of attention in visual search. Future work is needed to explore how real world conditions influence the spatial and temporal characteristics that have been proposed here. The value in pursuing such efforts is that we stand to enhance our understanding of how familiar environments promote (or discourage) efficient deployments of visual attention.

# References

Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 2, 886-893.

Green, D.M., & Swets, J.A. (1966). Signal detection theory and psychophysics. New York: John Wiley.

Kunar, M. A., Flusberg, S. J., & Wolfe, J. M. (2008). Time to Guide: Evidence for Delayed Attentional Guidance in Contextual Cueing. *Visual Cognition*, 16, 804-825.

Noton, D., & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science*, 171, 308-311.

Renninger, L.W., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision*, 7(3):6, 1-17

Tatler, B.W., Baddeley, R.J., & Gilchrist, I.D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45(5), 643-659.

Tatler, B.W., & Vincent, B.T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17, 1029-1054.

Torralba, A., Oliva, A., Castelhano, M., & 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.