UNIVERSITY OF CALIFORNIA

Santa Barbara

A Taxonomy of Contextual Influences on Visual Search

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Psychological & Brain Sciences

by

Kathryn Louise Koehler

Committee in charge:

Professor Miguel P. Eckstein, Chair

Dr. Craig Abbey, Researcher

Professor Barry Giesbrecht

Professor Mary Hegarty

September 2016

The dissertation of Kathryn Louise Koehler is approved.

_____

Craig Abbey

_____

Barry Giesbrecht

_____

Mary Hegarty

_____

Miguel Eckstein, Committee Chair

May 2016

A Taxonomy of Contextual Influences on Visual Search

by

Kathryn Louise Koehler

ACKNOWLEDGEMENTS

I would like to extend gratitude to my advisor, Miguel, and committee, Craig, Mary, and Barry for their thoughtful comments and training over the past years; throughout my degree milestones, but especially through our unstructured interactions in the hallways, during lab meetings, and at conferences, you have all set excellent professional examples and encouraged me to think critically about my and others' contributions to science. I owe special thanks to Miguel for his patience with my sometimes divided focus throughout graduate school, his guidance through all aspects of academic and professional pursuit, and his well-timed words of encouragement along the way.

Many research assistants, lab-mates, friends, and family helped directly with this research and indirectly, by providing support, friendship, laughter, and love throughout my graduate career as well. If you happen to be reading this, you will know who you are—thank you!

VITA OF KATHRYN LOUISE KOEHLER

September 2016

EDUCATION

University of California, Santa Barbara
Ph.D. in Cognitive Psychology **Expected 2016**
Dissertation: A Taxonomy of Contextual Influences on Visual Search

Arizona State University
B.S. in Psychology and B.A. in Physics **2006 - 2010**
Summa Cum Laude
Honors Thesis: Identity Based Categorization of Faces

RESEARCH EXPERIENCE

Graduate Research Assistant
Vision and Image Understanding Lab, UCSB **2010 - 2016**
Advisor: Dr. Miguel Eckstein

Honors Researcher
Arizona State University **January 2010**
Advisor: Dr. Donald Homa

Undergraduate Research Assistant
Arizona State University **2008 - 2009**
Advisor: Dr. Steve Goldinger

TEACHING EXPERIENCE

Teaching Assistant
Department of Psychological and Brain Sciences **2010 - 2016**
Laboratory in Perception (x4), Laboratory in Attention, Statistics (x3), Visual Perception (x2), Intro to Biological Psychology, Research Methods (x2), Intro to Psychology (x3), Mathematical Methods for Physics (Arizona State University)

PUBLICATIONS

Koehler, K. & Eckstein, M.P. (in prep). Temporal and Peripheral Extraction of Contextual Cues from Scenes during Visual Search

Koehler, K. & Eckstein, M.P. (submitted). Beyond Scene Gist: Objects Guide Search more than Scene Background

Koehler, K., Guo, F., Zhang, S., Eckstein, M.P. (2014). What Do Saliency Models Predict? *Journal of Vision. 14(3), 14.*

Do, P., Homa, D., **Koehler, K.** (2014). Identity Categories and Transformational Paths for Face Changes Across the Age Spectrum. Memory and Cognition, 42(2), 340-353.

CONFERENCE PAPERS AND PRESENTATIONS

Eckstein, M.P. & **Koehler, K. L.** (2016). Scene Context Leads to Inattentional Scale Blindness During Search. Poster presented at Vision Sciences Society Annual Meeting (Naples, FL, USA).

Koehler, K. L. & Eckstein, M.P. (2015). Independent Contributions of Multiple Types of Scene Context on Eye Movement Guidance and Visual Search Performance. Talk given at Vision Sciences Society Annual Meeting (St. Petersburg, FL, USA).

Koehler, K. L. & Eckstein, M.P. (2015). Scene Inversion Slows the Rejection of False Positives through Saccade Exploration During Search. Paper archived in proceedings and poster presented at Cognitive Science Society (Pasadena, CA, USA).

Koehler, K. L., Akbas, E., & Eckstein, M.P. (2013). Relating peripheral processing ability to learning in a visual search task. Poster presented at Vision Sciences Society Annual Meeting (Naples, FL, USA).

Akbas, E., **Koehler, K.L.**, & Eckstein, M.P. (2013). Learning of eye movements for human and optimal models during search in complex environments. Poster presented at Vision Sciences Society Annual Meeting (Naples, FL, USA).

Koehler, K. L., Akbas, E., Peterson, M.F., & Eckstein, M.P. (2012). Human versus Bayesian optimal learning of eye movement strategies during visual search. Poster presented at Vision Sciences Society Annual Meeting (Naples, FL, USA).

Koehler, K. L., Guo, F., Zhang, S., & Eckstein, M.P. (2011). Assessing models of visual saliency against judgments from one hundred humans viewing eight hundred real scenes. Poster presented at Vision Sciences Society Annual Meeting (Naples, FL, USA).

HONORS AND AWARDS

New Venture Competition, University of California, Santa Barbara
1st Place in Category **May 2015**

Department of Psychological and Brain Sciences, University of California, Santa Barbara
Contribution to Excellence in Teaching Award **February 2015**

Vision Sciences Society
Student Travel Award **February 2015**

ABSTRACT

A Taxonomy of Contextual Influences on Visual Search

by

Kathryn Louise Koehler

Although the facilitation of visual search by contextual information is well established, there is little understanding of the independent contributions of different types of contextual cues in scenes. Additional work in quantifying the time course of the influence of various contextual cues has not been performed, nor have researchers investigated how cue information is extractable across the visual field. Here we manipulated three types of spatial contextual information: object co-occurrence, multiple object configurations, and scene gist. We measured the spatial informativeness of each cue to target localization and isolated the benefits of each contextual cue to target detectability, its impact on decision bias, and the guidance of eye movements. To assess how cues are combined, we compare observed sensitivity during detection with multiple cues to a theoretical optimal combination of the various cues. We also utilize a novel paradigm where scene viewing time was contingent upon the number of fixations within a scene made by observers. To assess observers' ability to extract cue information across the visual field, we observed their performance at detecting cues in scenes shown exclusively in the visual periphery.

We find that object-based information guides eye movements and facilitates perceptual judgments more than background information. Despite its relatively weaker influence on search behavior, background information is shown to be most easily extracted across the visual field and likely to support the parsing of multiple object configurations in scenes. Multiple object

configuration information specifically is implicated in the guidance of initial search, providing coarse guidance that is later localized further by co-occurring object information. The degree of guidance and facilitation of each contextual cue can be related to its contribution to reducing the spatial uncertainty about the target location as measured by human explicit judgments about likely target locations. Comparison of target detectability across various cue conditions suggests that the performance improvements with multiple cues are consistent with an optimal integration of independent cues.

In addition to exploring influences of spatial cues on visual search task performance, we were also interested in assessing a non-spatial cue's effect on target detection performance and eye movement guidance. We manipulated the scale of a target object relative to its surroundings and found that observers were significantly worse at detecting mis-scaled targets relative to normally sized targets. Unsurprisingly, this non-spatial cue did not have as dramatic an effect on eye movement guidance as the three spatial cues. However, this emphasizes the importance of considering non-spatial scene information as a possible contextual influencer of human behavior.

Overall, our results improve the understanding of the interplay of distinct contextual scene components and their contributions to search guidance, a landmark behavior that differentiates human and biological vision from basic machine vision. The results are also useful in informing the type of information that might improve computer-based object detection and scene understanding.

TABLE OF CONTENTS

LIST OF FIGURES

I: Introduction

Vision is one of our primary senses for interacting with and understanding our world and we actively *see* without awareness of the complex neural architecture that supports our visual perception. We frequently search our visual environment; whether we are looking for the vitamin we dropped on the kitchen floor or the television remote in an unfamiliar living room, we have many cognitive mechanisms trained and ready to perform such tasks (Eckstein, 2011; Wolfe, 1994). Making sense of the outside world is a complicated process, but we are not without help. We rarely navigate a visual environment that lacks perceptual cues to help us complete typical visual tasks. Familiar backdrops, arrangements of objects, specific items, or simple cues surround us and guide our visual processing. It is uncommon that we encounter a particular visual situation in which we do not have any precedent of awareness, any expectations on at least a basic level, any "context". There is a rich history in understanding visual search using controlled, artificial displays (Eckstein, 2011; Hoffman, 1979; Koopman, 1956(a, b); Koopman, 1957; Treisman & Gelade, 1980; Williams, 1966; Wolfe, 1998), but less work has utilized realistic stimuli complete with contextual cues as they might occur naturally (Brockmole, Castelhano, & Henderson, 2006; Monica S. Castelhano & Heaven, 2010; Miguel P. Eckstein, Drescher, & Shimozaki, 2006; Neider & Zelinsky, 2006; Torralba, Oliva, Castelhano, & Henderson, 2006), where context can exert its usual influence on our perception.

It is well known that relationships between real-world objects and scenes can improve detection and recognition of objects in contextually rich environments. This has been termed a context effect in the case of real scenes (Biederman, 1972), or a contextual cueing effect with artificial displays (Chun & Jiang, 1998). Work that has investigated this effect typically pertains to tasks where an observer identifies an object in a contextually cued or uncued location (Biederman, Mezzanotte, & Rabinowitz, 1982; Chun, 2000; Neider & Zelinsky, 2006), or searches for a target

object in images where the consistency of the object with the scene containing it is manipulated (De Graef, Christiaens, & d'Ydewalle, 1990; Henderson, Weeks Jr, & Hollingworth, 1999). Although studies have established the importance of contextual influences on visual search and object recognition, a number of different image properties and object relations have been operationalized as contextual cues (Divvala, Hoiem, Hays, Efros, & Hebert, 2009). Precise definitions of various types of context and their separate contributions in facilitating search are unknown. Furthermore, additional characteristics of the nature of contextual influence have remained unexplored.

The overall goal of this work was to further the understanding of the role of real-world context in guiding eye movements and facilitating perceptual performance in visual search. There are many open questions to be addressed. What exactly constitutes scene context given the multiplicity of definitions and manipulations in the literature? Can we partition scene context into separate contextual cues and how do these combine when present in an image? What are the temporal dynamics of contextual influences of scenes and do different sources of information (contextual cues) operate at different temporal scales? And finally how do different contextual cues interact with the foveated nature of the human visual system thereby influencing how accessible the sources of contextual information are at different retinal eccentricities. For reference, Table 1 below gives an overview of the methods used in the experimental work included in this dissertation.

Specifically, we assessed the individual contributions of different types of spatial contextual cues in real scenes, the temporal dynamics of these influences, and their extractability across the visual field. Importantly, each cue was manipulated in such a way that we were able to understand their interactions and combined contributions to visual search performance. In line with existing work in the field, we focused on spatial cues that provide information about the spatial location of a target object. However, in the real world, not all cues are spatial. Thus we also investigated the

influence of a non-spatial cue, the relative size of objects within common scenes, to determine whether contextual information can also constrain visual processing during search to objects of requisite expected properties.

In Chapter II, we begin by reviewing existing work in the field of scene context. Namely, we establish a general definition of context in real scenes, explore current manipulations of different contextual cues, and review work that can provide foundational insight for our novel work in assessing the temporal dynamics of scene context influence and its extractability in the periphery.

In Chapter III, we introduce our three spatial contextual cues and assess the extent to which they are spatially informative to observers. We asked observers to specify whether scenes containing a single cue were more informative of a possible target's location than scenes without the cue and collected their expectations about target locations in the individually and fully cued scenes. In most cases, our cues were confirmed to be spatially informative, and we later relate the extent of their informativeness to measures of visual search performance and eye movement guidance as measured by the proximity of observers fixations to the target region.

Having established the informativeness of the cues, in Chapter IV we then manipulate scenes to contain all combinations of no cues, one cue, two cues, or all three cues and assess their impact on measures of target detection performance and eye movement guidance. We consistently demonstrate that object information more than background category information facilitates visual search. Furthermore, in order to determine how cues are combined to impact behavior, we compare our results to the derived theoretical optimal combination of statistically independent cue information on observer index of detectability (d') to demonstrate that observers are not sub- or super-optimally utilizing the independent cue information. We also relate the behavioral benefits of the cue information to their spatial informativeness from Chapter III. We measured the

informativeness of observer expectations of target location for each cue and showed that instances with higher informativeness of target location expectations were correlated with instances of greater search guidance.

Chapter V addresses the nature of the temporal influences of each cue as scene exploration unfolds. We sought to determine if cues were utilized differentially throughout the first few fixations of a scene, such that one cue might exert influence over behavior earlier than the others. Using the same stimuli with independent cue manipulations, observers performed a search task that terminated contingent on the number of fixations they had made within the scene. When assessing eye movement guidance, there was no interaction between cue and fixation allowance on target present trials, but there was a significant interaction on target absent trials. Further analysis relating the singly cued behavior to the fully cued behavior suggests that multiple object configuration information is initially utilized to guide eye movements to general highly probable target regions. Object co-occurrence information then further constrains search to a more proximal target region. We found evidence of this occurring within as few as three fixations, and consistent with chapter IV, background category information is relatively less influential than object based information. We also re-visit the combination of each cue type and again find evidence that observers optimally and linearly combine the information provided by each cue across fixations.

Chapter VI addresses how the various cues are extracted across the visual field. We assessed observers' abilities to detect cue presence in images displayed at various distances from central fixation. To assess the interaction of contextual information we explored a condition in which observers were required to detect each cue when it was the only cue available in a scene or when it was present along with the other two cues. We found that background category information was detected most robustly across the visual field, followed by object co-occurrence information, and

multiple object configuration information. Tying this together with results from previous chapters, even though background information is extracted most robustly across the visual field, recall that it directly influences visual search to a much lesser extent than object-based information. Crucially, multiple object configuration information was the only cue whose detectability was dependent on the presence of other cues, such that it was much easier to detect when the other cues were present. This indicates the probable role of background information: to facilitate our ability to make sense of and utilize the structure provided by objects in a scene.

Chapter VII extends our consideration of contextual influences into the non-spatial domain. We explore the influence of the relative sizes of objects on target detection by manipulating target objects to be mis-scaled relative to other objects in the scene. We show that observers heavily rely on size expectations and often fail to detect targets that are larger than would be expected. This result shows that contextual information provides more than just spatial expectations related to search. Context can influence our expectations concerning the properties of objects, in this case size, and constrain visual search to candidates that align to those expectations. We also demonstrate the lack of this effect in a state-of-the-art object detector.

Finally, in Chapter VIII we discuss how the combined results provide an overall discussion of contextual cues and their interplay. We also consider the implications these results have in understanding the neural mechanisms that underlie scene perception. Finally, we recommend future directions to improve future cue manipulations, including ways to assess the quality of human estimations and intuitions about contextual information, and the relation between background information content and scene gist.

Chapter	*n*	Task	Stimuli	Stim Duration	IV	DV
III (Definition Verification)	360	1) Select the object/image that provides the most information about where you expect the [TARGET] to be located 2) Click on the image where you would expect the [TARGET] to be located	Images with only one type of cue information present in scene, but target object absent. Note: Background category was removed by replacing the background with uniform mid-level grey.	∞	Type of (single) contextual cue tested; 3 levels (O, M, B)	1) Proportion of times the selected image/object was the image containing the cue/was the co-occurring object (depending on condition) 2) Distribution of selected target locations, quantified later as informativeness and mode
III (O Follow-up)	21	Click on the object that you would expect to be spatially closest to the [TARGET]	Images of scenes with all contextual cues present, but target object absent	∞	N/A	Proportion of times the selected object was the co-occurring object
III (Expected T Loc)	60	Click on the image where you would expect the [TARGET] to be located	Images of scenes with all contextual cues present, but target object absent	∞	N/A	Mode of location selections
IV (Independent Manipulations)	120	Was the [TARGET] present in the image?	Images with no cues, one cue, two cues, or all three cues present. 50% chance that the target object was present.	1500 ms	Amount of contextual information; 8 levels (None, O, M, B, OM, OB, MB, OMB)	Target detection performance (d', bias), distance of closest fixation to target or expected target location, time to foveate target region

V (Saccade contingent)	300	Was the [TARGET] present in the image?	Images with no cue, one cue, or all three cues	Variable (see IV)	Amount of contextual information (5 levels: None, O, M, B, OMB) x number of fixations allowed (3 levels: one, two, or three)	Target detection performance (d', bias), distance of closest fixation to target or expected target location
VI (Peripheral Extraction)	360	Determine if cue information was present in image, exact instructions dependent upon cue type tested (see Figure 20)	Images with a single cue paired with images with no cue (i.e., None/O, None/M, or None/B if no other cues present condition) or images with all cues paired with images with a single cue missing (i.e., OMB/MB, OMB/OM, OMB/OB if all other cues present condition)	500 ms	Type of (single) contextual cue tested (3 levels: O, M, B) x presence of other cues (2 levels: no other cues present, all other cues present)	Cue detection performance (d', PC)
VII (Mis-scaled targets)	60	Was the [TARGET] present in the image?	Images with a normal or mis-scaled target, or zoomed-in control images	1000 ms	Target and context scale (3 levels: normal target and context, big target and normal context, big context and big target)	Target detection performance (hit rate), distance of closest fixation to target region

Table 1: Summary of methods used in experimental work throughout the various chapters.

II: Background

Introducing Contextual Information and the Context Effect

Successful search for objects in cluttered scenes is challenging. Computers still cannot attain the competence with which humans and non-human animals perform visual tasks because species have evolved visual systems that optimize search (Eckstein, 2011; Hayhoe & Ballard, 2005; Tatler, Hayhoe, Land, & Ballard, 2011; Wolfe, Alvarez, Rosenholtz, Kuzmova, & Sherman, 2011). Species ranging from pigeons (Wasserman, Teng, & Castro, 2014) and bees (Eckstein et al., 2013) to monkeys (Maunsell & Cook, 2002) and humans (Chun, 2000; Chun & Jiang, 1998; Luck, Hillyard, Mouloua, & Hawkins, 1996) can exploit statistical regularities of the visual environment to facilitate visual search. For example, artificial cues (e.g., boxes, arrows) that are predictive of a target location will lead to detection accuracy improvement (Carrasco, 2011; Dosher & Lu, 2000; S. S. Shimozaki, Eckstein, & Abbey, 2003; Smith, 2000), shorter response times (Posner, Snyder, & Davidson, 1980), and frequent eye movement fixations toward predictive cues and spatial locations likely to contain the target (Droll, Abbey, & Eckstein, 2009; Peterson & Kramer, 2001; Walthew & Gilchrist, 2006). In the real world, statistical regularities arise because visual search for a target object does not typically occur amongst unfamiliar objects scattered in random locations, but rather, scenes associated with targets consistently present specific visual properties, objects and spatial relationships among the objects. If a friend asks you to run into his kitchen to stir the contents of a frying pan, even if you have never been in his/her kitchen, you will have plenty of familiar context clues to guide your visual search.

Influences on visual search behavior are often categorized into bottom-up, feature-based factors and top-down knowledge- and context-based factors. Bottom-up influences arise from features inherent to a stimulus, such as luminance, intensity, orientation, or color (Itti & Koch, 2000;

Koch & Ullman, 1985; Parkhurst, Law, & Niebur, 2002), and can exert effects in the absence of familiarity with a scene or its contents. Top-down factors come into play when we can use prior experience and external knowledge to guide our behavior. Returning to our earlier example of a vitamin on the kitchen floor, if we know the basic features of the vitamin we dropped, we can use this information to facilitate locating the vitamin (Burgess, 1985; Miguel P Eckstein, Beutter, Pham, Shimozaki, & Stone, 2007; Malcolm & Henderson, 2009; Rao, Zelinsky, Hayhoe, & Ballard, 2002; G. J. Zelinsky, 2008). Similarly for locating a television remote in a friend's living room, we know that television remotes are generally on coffee tables, coffee tables are usually in front of couches, and we can easily identify the location of those things to define a small—relative to the entire visual field—region of space to search for a television remote in an unfamiliar living room. In real-world search tasks, we often employ our pre-existing knowledge about scenes and targets. As such, the incorporation of top-down information into models of human eye-movements has been shown to be much more important than bottom-up information for correctly predicting human fixations during a variety of visual search tasks (Birmingham, Bischof, & Kingstone, 2009; Chen & Zelinsky, 2006; Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009; G. Zelinsky, Zhang, Yu, Chen, & Samaras, 2005; Koehler, 2015).

It should have come as no surprise that top-down information would be highly influential on our visual search patterns. Researchers concerned with object recognition (as opposed to visual search per se), had long advocated for studying object processing in tasks that better reflect real world environments, and were among the first to introduce the notion of "context" as an important top-down factor. The original inspiration for research about the effects of context arose from Biederman's (1972) observation that sparse, unrelated elements (whether real or artificial) on a blank display do not accurately reflect the real world scenes we view while performing daily tasks. A crucial element that is missing from such bare displays is that of semantic or contextual relations,

both among objects within the display and to the background that encompasses the displayed scene. Biederman went on to demonstrate the context effect, where observers were more accurate at identifying an object in a scene that provided contextual information than a scene that provided no context (in this case, a scene that was divided into parts and rearranged). Additional early work replicated the context effect (Boyce & Pollatsek, 1992a; see Aude Oliva & Torralba, 2007a for a review; Palmer, 1975b) and debated whether it was due to rapid acquisition of the categorical schema of a scene (Biederman, Mezzanotte, & Rabinowitz, 1982; Biederman, 1981) or to semantic priming of objects as they are fixated (De Graef et al., 1990). A corroboration of the context effect is also evident in early work that showed a shorter latency of eye movements on objects that are semantically consistent with a scene than those that are not and a tendency for eye movements to be generated in the direction of contextually expected locations of objects during target search (Miguel P Eckstein, Drescher, & Shimozaki, 2006; Henderson, Weeks Jr, et al., 1999; Loftus & Mackworth, 1978; Neider & Zelinsky, 2006; Palmer, 1975; Torralba et al., 2006).

Work evaluating the contextual cueing effect in artificial images (Chun & Jiang, 1998) was useful for quantifying the aspects of images that produce context effects, despite its departure from focusing on real-world scenes. The contextual cueing effect was established using highly controlled stimuli to ensure precise operationalization of context and its influences on visual processing, reminiscent of other research of top-down factors of attentional allocation. The contextual cueing paradigm originally demonstrated that global image properties (the spatial layout of artificial objects in a search display) facilitate attentional guidance. These results tied contextual cues to visual search behavior in artificial scenes, paving the way for researchers to understand how natural contextual information can also influence visual search.

We therefore refer to contextual cues as they occur in natural scenes as scene context. Scene context can be broadly thought of as the portions of a scene with which observers have familiarity or pre-existing knowledge and that can used to aid their perception of the scene. Scene context is useful to the extent that it provides information to better make sense of a scene or better perform a task such as object recognition or visual search.

Types of Contextual Cues and their Relative Influences

Contextual cues can range from the identification of the category or basic semantic descriptions of real scenes (Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Aude Oliva, 2005, p. 200; Torralba et al., 2006), to semantically related objects within a real scene (Hwang, Wang, & Pomplun, 2011; Moores, Laiti, & Chelazzi, 2003; Wu, Wick, & Pomplun, 2014), spatially related objects in a scene (Bar, 2004; Chun & Jiang, 1999; Mack & Eckstein, 2011), the spatial relations of objects to a scene background (Castelhano & Heaven, 2011; Jiang & Wagner, 2004), and the semantic relation of objects with their backgrounds (Henderson, Weeks Jr., & Hollingworth, 1999; Hollingworth, 1998; Loftus & Mackworth, 1978). More recently, the effects of scene context on visual search have been explored, demonstrating that scene-based expectations can guide eye movements to expected target locations (Monica S. Castelhano & Henderson, 2007; Miguel P Eckstein et al., 2006; Mack & Eckstein, 2011; Neider & Zelinsky, 2006; Torralba et al., 2006; Wu, Wang, & Pomplun, 2014). Beginning in the study of artificial images, contextual cues were usually dichotomized into local and global forms of context (Brockmole, Castelhano, & Henderson, 2006). Local cues are structural and spatial regularities immediately surrounding a visual target and have been shown to be the important factor in facilitating target localization (Olson & Chun, 2002) whereas global cues are comprised of elements in the overall display and have been shown to also improve observers' performance at target detection across repeated display epochs (Jiang & Wagner, 2004).

Drawing on this distinction, much work has characterized the global information extracted from real scenes as the "gist" of a scene. Written definitions of scene gist often appeal to our intuition, referring to our natural, quick impressions of a scene and its contents. Harkening to a real world example, Oliva (2005) exemplifies scene gist as our sense of what is playing on each station as we surf through the various channels on our television, even when we see each channel for less than a second. In practice, any information that is thought to be or has been empirically shown to be extracted within about 300 ms of scene onset is often referred to as "scene gist'. Therefore, gist has been experimentally characterized many ways, and frequent use of the term has equivocated vastly different treatments of gist across the literature.

Table 2 summarizes a sample of papers that explicitly defined or manipulated scene gist or that are consistently referred to in the literature as manipulations of scene gist. Gist has been most commonly operationalized as the category of a scene along some dimension (Larson & Loschky, 2009; Schyns & Oliva, 1994; S. J. Thorpe, Gegenfurtner, Fabre-Thorpe, & Bülthoff, 2001), but has also been defined as the perceptual content of a scene, which could refer to object and background information, or even high-level relations between those things (e.g., a father helping a little boy in a cubicle) (Fei-Fei, Iyer, Koch, & Perona, 2007) or a description of the main event or focal foreground objects in a scene (e.g., girl sitting in bed; Mary C. Potter, 1976). Crucially, such studies posit the perceptual aspects of scene gist, whereas other work has attempted to manipulate scene gist, thus mapping the perception of gist to certain properties of a scene itself. For example, Castelhano and Heaven (2011) define gist to be knowledge of whether a particular object belongs in a scene, therefore likening scene gist to semantic information about that scene and manipulating whether an object is typically present in a scene in order to manipulate "gist" for an object. Wu, Wang, and Pomplun (2014) eliminate scene gist information by removing the background content of a scene and displaying only the foreground objects of a scene on a grey background. Work in modeling

scene gist has typically been concerned with successful machine categorization of scenes and emphasized that local, object-based information is not necessary for determining scene gist. Instead, models of gist successfully categorize scenes and reliably predict spatial properties about scenes (e.g., openness, ruggedness) using holistic properties of scenes and global scene statistics (Oliva & Torralba, 2001, 2006).

Citation	**Referred to as gist by authors or by others?**	**Definition/manipulation**
(Larson, Freeman, Ringer, & Loschky, 2014; Larson & Loschky, 2009; L. Loschky et al., 2015; L. C. Loschky & Larson, 2010)	Authors	Basic-level category
(Schyns & Oliva, 1994)	Others	
(M. S. Castelhano & Henderson, 2010)	Authors	
(Rousselet, Joubert, & Fabre-Thorpe, 2005)	Authors	

(Fei-Fei et al., 2007)	Authors	"Perceptual content" of an image: low-level properties, objects, scene-object relations, and high-level descriptors of scene-object relations
(Mary C. Potter, 1976)	Others	Description of the main event or focal foreground objects in a scene
(S. J. Thorpe et al., 2001)	Others	Detection of whether a scene contains an animal
(Monica S. Castelhano & Heaven, 2011; Monica S. Castelhano & Henderson, 2008)	Authors	Knowledge of whether a particular object belongs in a scene
(Groen, Ghebreab, Prins, Lamme, & Scholte, 2013)	Authors	Scene naturalness categorization
(Wu, Wang, et al., 2014)	Authors	The background of a scene (i.e., everything except foreground objects)

Table 2: Sample of the variability in treatments of the term “scene gist” in the literature

The global vs. local distinction is not the only dichotomy sometimes adopted when breaking down scene features. Some researchers distinguish “scene context” itself from the structural or movable object elements in a scene (Pereira & Castelhano, 2014), whereas others treat objects as a form of scene context, akin to local contextual information (Brockmole et al., 2006). The confusability of these distinctions are especially prominent in the literature using the flash-preview moving-window paradigm, where an observer is shown a brief preview of a scene (from which they can presumably extract global information), and then explores the scene with a restricted circular view of the scene centered on their eye-tracked gaze location, restricting them to utilize local information during search (e.g., Castelhano & Henderson, 2007; Võ & Henderson, 2011). It is unclear what information exactly is being extracted during the flash-preview and moving-window segments of search. The state of the literature calls for a more structured treatment of various scene elements

as types of contextual information. Our work therefore avoids relying on vague terms to describe specific properties in scenes. Instead, we argue that scene context is comprised of many different features of a scene, related to both background and object properties. Specifically, we explore information provided by an object that spatially co-occurs with a visual search target, multiple other objects that provide spatial structure of the scene based on their typical arrangements, and the background content of the scene that portrays the superordinate category membership of the scene.

Important alongside the effort of improving the clarity of contextual cue definitions is quantifying their independent and combined influences on visual search behavior. There is some debate as to what aspect of contextual information is most informative, further confounded by the fact that comparisons across studies of contextual influences of the same name often do not refer to the same scene properties (e.g., Table 1). Studies have shown that target feature knowledge is more beneficial than scene knowledge during visual search (Monica S. Castelhano & Heaven, 2010). However, this evidence does not explicitly state the role of related objects, nor does it preclude the ability to extract and utilize scene gist, as demonstrated by a number of models (Munneke, Brentari, & Peelen, 2013; Oliva, Torralba, Castelhano, & Henderson, 2003; Oliva & Torralba, 2001; Torralba et al., 2006; Torralba & Oliva, 2003). Some studies have investigated the contributions of both scene gist (defined as the background elements of a scene) and spatial dependencies of objects on the guidance of eye movements between semantically related objects during a memorization task, (Hwang et al., 2011; Wu, Wang, et al., 2014; Wu, Wick, et al., 2014), but not with the express purpose of comparing their independent contributions to visual search behavior. One study (Pereira & Castelhano, 2014) has looked at the contributions of background elements and objects to search guidance. However, they evaluated the presence of objects irrespective of their typical spatial associations with the target and did not evaluate the relative configuration of objects as a potential

contextual cue. It is likely that there are additional types of context that guide visual search, and that understanding the differences between them will be useful in fully understanding the context effect. Existing work has explored quantifying a variety of statistical relations between objects and scene environments (Greene, 2013) but did not assess their relation to behavior or experimentally manipulate the content of scenes. Furthermore, most studies focus on a single type of contextual information (i.e., contextual cue), failing to assess the interactions of multiple types.

Time course of extraction of scene context

With a defined partitioning of various contextual cues, it is also interesting to understand how and when they are used by the visual system to guide search. We are interested in characterizing the temporal dynamics of the contributions of different contextual cues to behavior, therefore literature on the ability of humans to rapidly extract information about scene and object information could inform us about how early during search various contextual cues exert their influence.

The majority of existing efforts have explored the time-course of the extraction of the basic-level category of a scene on very short (within a single fixation) timescales. Larson, Freeman, Ringer, and Loschky (2014) investigated the spatiotemporal unfolding of basic-level category extraction during an observers' first fixation within a scene over the course of about 350 ms. Their results speak to the spatial location of covert attention during short time periods, demonstrating that central vision is more integral to category recognition than peripheral vision in the first 100 ms of scene viewing due to the typical "zoom-out" of covert attention from central to peripheral visual field areas during fixation. Further exploration of the specifically temporal extraction and utilization of scene categories or more complex verbal descriptions of scene content at longer time scales than ~300 ms has likely not occurred because it is well established that complex scene information is

processed prior to the initiation of an eye movement in a scene, within as little as 20 ms of viewing the scene (Antes, Penland, & Metzger, 1981; Fei-Fei et al., 2007; Metzger & Antes, 1983; M. C. Potter, 1975; S. Thorpe, Fize, Marlot, & others, 1996).

However, it is important to understand more than just the time-course of extraction, and to also consider how the extraction of scene properties impacts human behavior over multiple fixations during scene viewing. For example, even though humans are remarkably quick at extracting category information and there is evidence that this information facilitates object recognition (Biederman, 1972), scene information may not be recruited or available to guide eye movement behavior until later during scene viewing.

The types of scene context considered thus far have come from studies specifically exploring "scene gist", which typically overlook or do not specifically address the role of objects within scenes. Objects are also useful providers of scene context. Humans and monkeys have been shown to both detect and remember above chance objects in a rapid serial visual presentation (RSVP) task after object image exposures as low as 14 ms (Keysers, Xiao, Földiák, & Perrett, 2001). Activation in high-level, object oriented visual areas has been observed at 40 ms after simple object image exposure (Grill-Spector, Kushnir, Hendler, & Malach, 2000) and neural signatures corresponding to object detection in complex natural scenes have been detected as early as 150 ms (S. Thorpe et al., 1996). Neural correlates of object recognition are thought to arise later during processing, likely between 135-500 ms after exposure (Johnson & Olshausen, 2003). How scene context affects object perception has been extensively studied (Bar, 2004), but to our knowledge, the influence of the contextual information provided by objects over time has not been investigated. A study by Spotorno, Malcolm, and Tatler (2014) investigated how target template specificity and the consistency of an object with the contextual information provided by other scene elements affected eye movements across three temporal epochs of visual search, initiation, scanning, and verification.

They found that the visual system employs contextual information early during the initiation of visual search. Our work builds upon theirs by decomposing scene context into multiple contextual cues and assessing their influence at a finer scale during search initiation and scanning (the first few eye movements of visual search) using a restricted fixation allowance paradigm. We expect that not all contextual information is treated equally during scanning, and that a fixation-by-fixation analysis will reveal differential influences of object- and scene-based cues on eye movements as scene exploration unfolds, further elucidating the early influence of contextual information on visual search. Although some scene properties can be extracted and objects can be recognized in similarly short time periods, the extraction of scene properties contained in background elements that facilitate determining scene category membership is likely not influenced by an observers' fixation location within a scene, whereas objects may not be recognizable in the distant periphery or if outside the uncrowded window (Pelli & Tillman, 2008). Based on this prediction, one may expect to find evidence that scene background information guides eye movements early on until further localization to relevant objects occurs.

Extracting scene context across the visual field

Given evidence suggesting that some contextual influences occur prior to the initiation of a saccade, it must be the case that some type of contextual information is extractable in the visual periphery. Therefore, we are interested in evaluating how contextual influences on eye movements and performance vary across the visual field. Previous studies evaluating how explicit extraction of scene and object information varies with retinal eccentricity are relevant to our current goals. As we have shown, a large body of work has explored the extraction of categorical and semantically descriptive scene information and object identity, and these things happen over very short timeframes upon fixation. This gives some insight into the expected importance of peripheral visual

fields in extracting contextual information. For example, given that scene identification according to a short semantic description of its content precedes the execution of an eye movement (Potter, 1975), it has generally been assumed that some descriptive scene information is recognizable via the recruitment of peripheral vision.

This assumption arose from literature exploring object processing. Researchers were interested in understanding whether the extraction of various scene properties influenced object recognition and identification or vice versa. The logic of one of the dominant original paradigms was to modify either the semantic or syntactic consistency of objects with the scene context and determine whether and when a preferential processing of those violations took place. Initial evidence suggested that observers are able to direct initial saccades within a scene toward informative scene regions (Antes, 1974; Mackworth & Morandi, 1967) and semantically inconsistent objects (Loftus & Mackworth, 1978), therefore it has been reasoned that observers are able to extract scene properties prior to the first saccade and interpret an objects' consistency with the scene in the visual periphery. Some conflicting results suggest that objects are only semantically interpreted near to and within the fovea (Henderson & Hollingworth, 1998). Furthermore, when visual processing is constrained such that observers are only able to obtain peripheral visual information during a short preview of a scene, their eye movements are no longer directed toward semantically or syntactically inconsistent objects sooner than consistent objects, suggesting that this information is either not extracted in the visual periphery or not utilized during subsequent scene exploration (Võ & Henderson, 2011).

Regardless of one's conclusions regarding the influence of object-scene consistency information on human behavior, it has consistently been confirmed that categorical and basic semantic descriptions of a scene can be extracted within as little as 50 ms of viewing a scene (Antes et al., 1981; Fei-Fei et al., 2007; Metzger & Antes, 1983), and thus it is a reasonable assumption that

scene property extraction must rely at least partially on the peripheral visual field. What properties specifically are being extracted is still an open question.

Some work has investigated perception as a of whole scene images entirely in the periphery, with efforts to directly explore the extractability of scene properties in the periphery (see Strasburger, Rentschler, & Jüttner, 2011 for a review). It has been shown that observers with access to peripheral information perform better at basic-level scene categorization than those with access to central information in a scene viewing paradigm where observers have access either to only peripheral information (a simulated retinal scotoma condition) or to only foveal information (a window-like view of a scene centered on an observers' fixation point with peripheral information masked) on both standard computer monitors (Larson & Loschky, 2009) and large monitors that present stimuli 180 degrees horizontally and more in alignment with real-world scene viewing (L. Loschky et al., 2015). Consistent with this result, classification of scenes shown fully within peripheral regions of the visual field is robust (Calvo, Nummenmaa, & Hyönä, 2008; Li, VanRullen, Koch, & Perona, 2002) even 70 degrees into the visual periphery (Boucart, Moroni, Thibaut, Szaffarczyk, & Greene, 2013). Studies of patients with macular degeneration (impaired foveal vision) have also demonstrated that scene information extracted in the periphery can aid object categorization given foveal impairments (Boucart, Moroni, Szaffarczyk, & Tran, 2013).

Part of the debate surrounding whether objects that are inconsistent with scene contexts attract attention is the assumption that objects must be recognized, at least to the point of identifying their unlikely scene membership, in the periphery. However, when it comes to directly exploring object detection and recognition in the periphery, much work has focused on the perception of letters and artificial objects at various eccentricities with the purpose of understanding visual crowding (see Levi, 2008 for a review). Concerning real, complex objects, researchers have demonstrated that foveal processing is beneficial, but not necessary for object

encoding (Henderson, Mcclure, Pierce, & Schrock, 1997) and that participants are able to recognize objects at a superordinate and sometimes basic category level within a single fixation (Fei-Fei et al., 2007), suggesting at least the plausibility of object identification and recognition in the periphery. In fact, there is direct evidence that observers are able to detect the presence of animals as far as 60 degrees into the periphery at 70% accuracy (S. J. Thorpe et al., 2001).

Object identification in the periphery is complicated by the effects of visual crowding, such that adjacent contours and shapes can impair an observers' ability to resolve object details that would otherwise be discernible in isolation (Levi, 2008). Crucially, crowding only impairs the identification and not the detection of objects (Whitney & Levi, 2011). This suggests that, regardless of an observers' inability to identify objects in their periphery, they may still be able to extract certain information, such as spatial distributions and approximate locations of fuzzy forms. Indeed, it has been demonstrated that the adherence of a scene to proper spatial relations among its features improves recognition of a given feature (Bar, Ullman, & others, 1996). However, how able are we to extract the spatial arrangement of objects in the periphery? The spatial arrangement of objects, even without their constituent recognition, is an increasingly investigated form of contextual information (Jiang & Wagner, 2004; Olson & Chun, 2002). Similar to the logic that rapid scene recognition implies our ability to extract scene properties in the periphery, there is evidence that coarse information in the form of a scene's low spatial frequency information, sufficient to convey the approximate spatial layout of a scene, is used to categorize scenes during early-stage processing (Schyns & Oliva, 1994). Assessing the extent to which objects embedded in scenes are able to be identified at eccentric retinal locations may shed light on the plausibility that scene-inconsistent objects could be targeted for processing by the visual system.

The Importance of Assessing Cue Interactions and Definitions

Two final points of consideration again concern the wide variability of contextual information sources thus far considered in the literature and their imprecise definitions. A complete picture of the temporal and peripheral components of contextual information must assess multiple cue types. For example, Pereira and Castelhano (2014) found that background information and object information extracted from the periphery interact, such that background information guides eye movements to regions likely to contain a target, whereas object information provides more localized information about where to search. Additional studies have also highlighted an interaction between background and object information (Davenport & Potter, 2004; Joubert et al., 2007; Võ & Schneider, 2010). Focusing on one cue type at a time in disparate studies makes it difficult to detect such interactions and compare effects of different cues and rule out the contribution of confounding cues. Additionally, this effort would be amiss without manipulating cues whose definitions are based on more than the researchers' intuition (Greene, 2013, 2016). As scene databases expand, and the accurate labeling of those scenes becomes more feasible using micro-task work forces such as Amazon Mechanical Turk, it is entirely possible and crucial to document and assess natural scene statistics and scene manipulations. One effort along these lines has been to quantify object-scene relations in two separate scene image databases (Greene, 2013), where it was also demonstrated that humans are prone to over-estimate the frequency of a particular object being present in those scene images (Greene, 2016). The latter result could reflect a divide between real-world and image-based object-scene relations, and over estimations of image statistics could arise from accurate estimations of real-world statistics. Here, the important aspect of contextual information we are concerned with is its informativeness during a target search task. Therefore, to experimentally assess our spatial cue manipulations, we verified the extent to which our cues are considered to be spatially informative of a target object location by an independent group of observers for each image used in the study.

III: Cue definitions and their spatial informativeness

Definitions

A centerpiece to this work is the differentiation of multiple contextual cues in scenes and the assessment of their individual influences to eye movements and decisions. Here, we define the three proposed types of context that provide spatial information concerning the location of a target.

Object co-occurrence. This is a form of contextual information that facilitates detection of a less conspicuous object by pairing it with a larger, more salient object that it often co-occurs with. Specifically, the two co-occurring objects are defined to be closely spatially coupled (physically near to one another in the visual scene). However, the particular configuration of the pair of objects is not a strict relationship. For example, a woman might frequently leave her car keys on the surface next to her purse. Her purse is likely to be large and easier to find than a small set of keys. Once she locates her purse, she can then narrow her search radius to locate her keys. The keys are not always to the left or right of the purse, but they are always near to one another. Removing the cued object will disrupt this form of contextual information, but moving the objects relative to one another should not alter the relationship as long as the distance separating them remains roughly uniform. That the two objects are semantically related is not necessary.

Multiple object configurations. Object configurations provide context similar to objects that co-occur, however obtaining contextual information from the configuration is dependent on a particular grouping of numerous items in an expected spatial arrangement. The objects could be spatially distant from one another, and the combination of all of them in a particular arrangement provides the contextual information. For example, bedrooms will almost always contain a bed with an adjacent nightstand and lamp, as well as a dresser and closet. Violating the spatial regularities

with which these configurations are observed by jumbling the objects will disrupt the context they provide.

Background Category. This cue consists of everything in the scene except for the foreground objects and can be indoor, natural outdoor, or urban outdoor. We believe the background information to be most closely associated to perceptual observervations usually attributed to scene gist. Thus, we believe that background information can be extracted before attentional selection occurs during scene viewing and therefore be the basis of our ability to quickly identify the general category of a scene (e.g., living room or campsite). Mismatching the background of a scene, but preserving the foreground objects in the scene should disrupt the context effect that background information provides in a scene.

Experiment to assess spatial informativeness of cues

With a basis of cue partitions set, we then manipulated real world scenes to selectively violate the chosen definitions. Our interest in manipulating these cues was to eventually assess their relative impacts on visual search guidance. Given our definition of scene context, and the specification that scene context is useful to the extent that it provides information to facilitate a certain task, we measured the extent to which images containing our cues increased the information about the target object locations. We also wanted to evaluate directly how spatially informative each individual cue was of the placement of observers' target location expectations.

Methods

Participants. 360 Amazon Mechanical Turk (AMT) workers who reported having normal or corrected-to-normal vision participated in the main verification task. An additional 81 undergraduate students from University of California, Santa Barbara who received course credit for

participation and were tested to have normal or corrected-to-normal vision participated in one of two follow-up target location expectation tasks. All participants provided informed consent.

Stimuli. The stimuli used in this experiment comprise images depicting natural indoor and outdoor scenes with manipulations of three types of contextual cues. A base set of 48 scenes was constructed in Unity 3D (Unity Technologies, Bellevue, WA, USA), a video game building and physics engine platform, each with a specified target object that would serve as the searched-for item in a subsequent visual search task. Each scene contained other objects that one might expect to find in a scene containing the target object and a background that was consistent with the target object. There were 16 unique target object categories (e.g., frying pan), each used three times, but corresponding to three different object exemplars (e.g., the viewing angle, design, color, or size was varied across the three exemplars of the category). Each base scene contains all three experimentally defined contextual cues which were manipulated to form versions of the scene missing certain cues. For the purpose of verifying the contextual cue manipulations, participants in this task saw versions of the scenes with completely isolated contextual cues, different from later chapters (see the relevant stimuli sections). Each base scene was constructed such that the normal, intact version contained a target, with a frequently co-occurring object placed near to it (constituting the object co-occurrence cue), a number of other objects that would typically also be present in the scene arranged in a typical way (the multiple object configuration cue), within a background that exemplified the scene category and was consistent with the target and other objects (the background category cue). Other versions of the scenes were created to isolate the various contextual cues or create target absent stimuli. For the main verification experiment, all participants viewed target absent versions of the scenes. Participants who verified the object co-occurrence manipulation (the “O” condition) viewed a version of the scene with all objects except the co-occurring object jumbled on a grey background. Participants who verified the multiple object

configuration manipulation (the “M” condition) viewed a version of the scene with and without the co-occurring object on a grey background. Finally, participants who verified the background category manipulation (the “B” condition) viewed versions of the scenes with all objects removed, i.e., just the backgrounds. Example stimuli are shown in Figure 1. There were two AMT quality assurance images included as well, described in the procedure.

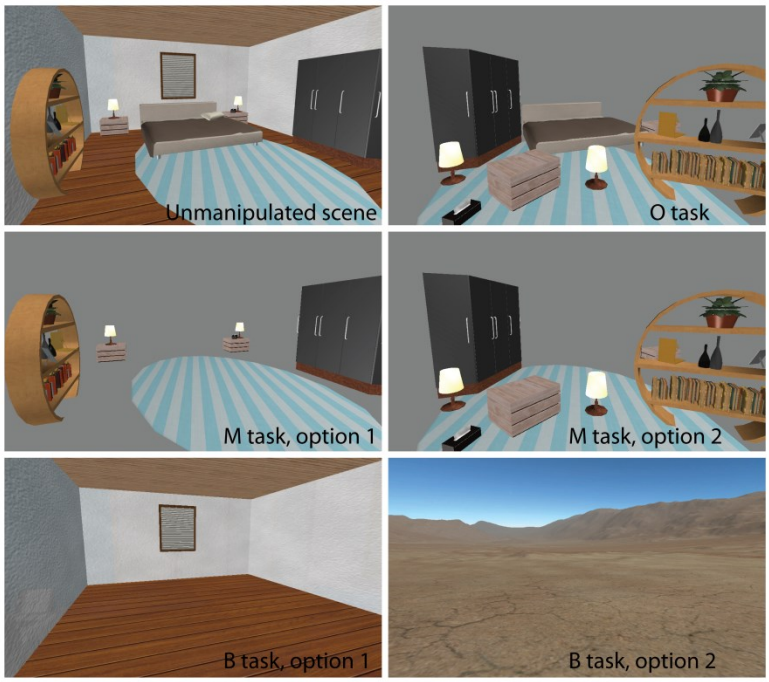

Figure 1: Example of stimuli presented to participants in the cue manipulation verification task for a sample scene. In this scene, the target was PILLOW. All stimuli in this task were target absent images. As labeled, observers in the O task viewed an image with all objects jumbled except the co-occurring object on a grey background, M observers viewed images without the co-occurring object present with all objects ordered in a typical way or jumbled, B observers viewed images with a matched or mismatched to the target background category. Observers’ tasks were to select the object (O condition) or image (M and B conditions) that would provide the most information about there the target object would be located.

Design. Separate groups of 40 observers each viewed a group of 16 images such that no observer saw the same target twice. Each observer was assigned to the O, M, or B condition, therefore a total of 120 observers viewed 48 images for each condition.

Procedure. After consenting to participate in a psychological study and indicating that they had normal or corrected-to-normal vision, participants were given a brief tutorial about how to use the experiment interface. Each participant performed two tasks. The first task varied by condition.

Observers in the O condition were required to click on the object that was most informative of the target object's location. Condition M and B observers were required to select the image that they thought was most informative of the target object's location between a jumbled and non-jumbled version of the objects (without the co-occurring object on a grey background) or between an indoor or outdoor background (with no objects), respectively. The second task required participants to click within the image (whichever they had previously selected in Task 1 for M and B participants) where they would expect the target object to be located. The task instructions were selected to be uniform across conditions and to explicitly assess the informativeness of the cue manipulations, however it is important to note in the first task that observers in the O condition were choosing to select one of the many objects (typically >10) in the scene whereas observers in the M and B conditions were choosing between only two possible options, therefore selections made solely by chance would result in drastically different rates of selecting our experimental manipulation. Image order was randomly determined, and two quality assurance trials were randomly mixed into the experimental trials. The first quality assurance trial was to indirectly assess overall understanding of the task instructions and mastery of the English language by using a simplified version of the stimuli to which there was an obvious correct answer. The second trial was to ensure that click recording was calibrated correctly within the browser window and required participants to click at the center of a target. At the end of the experiment participants filled out a short questionnaire indicating how well they felt they understood each tasks' instructions, their age, and their gender.

Given the differences in task 1 between the O and M and B condition, we opted to perform a follow-up to condition O, task 1 with a group of 21 separate undergraduate observers that more directly probed the basis of our object co-occurrence manipulation, but would have violated the uniformity of instructions and stimuli in the main task. These observers were asked to select the

object that they would expect to be physically closest to the target object while viewing a scene with all contextual cues present.

Finally, a separate group of 60 observers viewed images taken from the base scene (with all cues present, but target absent) and clicked within the image where they would expect the target object to be located. These judgments were compared to the selections made in task 2 by participants in the main verification experiment and are also used in later chapters as an empirical basis for assessing eye movement guidance on target absent trials in the visual search tasks.

Results

Verification of experimental contextual cue informativeness. Participants who reported understanding the tasks with a rating that was two standard deviations below the mean rating across participants were discarded from analysis. The average reported level of understanding among remaining participants was 9.2 for both tasks 1 and 2 on a 10 point scale, with 10 being the highest level of understanding. After discarding an additional 4 observers who failed the MTurk quality assurance task criteria, we analyzed the data of 110 participants for the O condition (image group 1, n = 36; image group 2, n = 36; image group 3, n = 38), 111 participants for the M condition (image group 1, n = 36; image group 2, n = 37; image group 3, n = 38), and 107 participants for the B condition (image group 1, n = 35; image group 2, n = 35; image group 3, n = 37). Shown in Figure 2 (left) is the proportion of participants who verified our manipulation of a particular cue for each image. A verification for the O task was taken to be an instance where the participant selected the experimentally defined co-occurring object as the most informative of the target object's location. We considered the manipulation of the M task to be verified when a participant selected the experimentally defined non-jumbled version of the multiple objects. Finally we deemed that the manipulation of the B condition to be verified if the participant chose the experimental background as the most informative of the target object's location. On average, our background category cue

was verified 96% of the time, multiple object configuration cue 84% of the time, and object co-occurrence cue 64% of the time. The right side of Figure 2 shows a histogram of the proportion of agreement for each contextual cue.

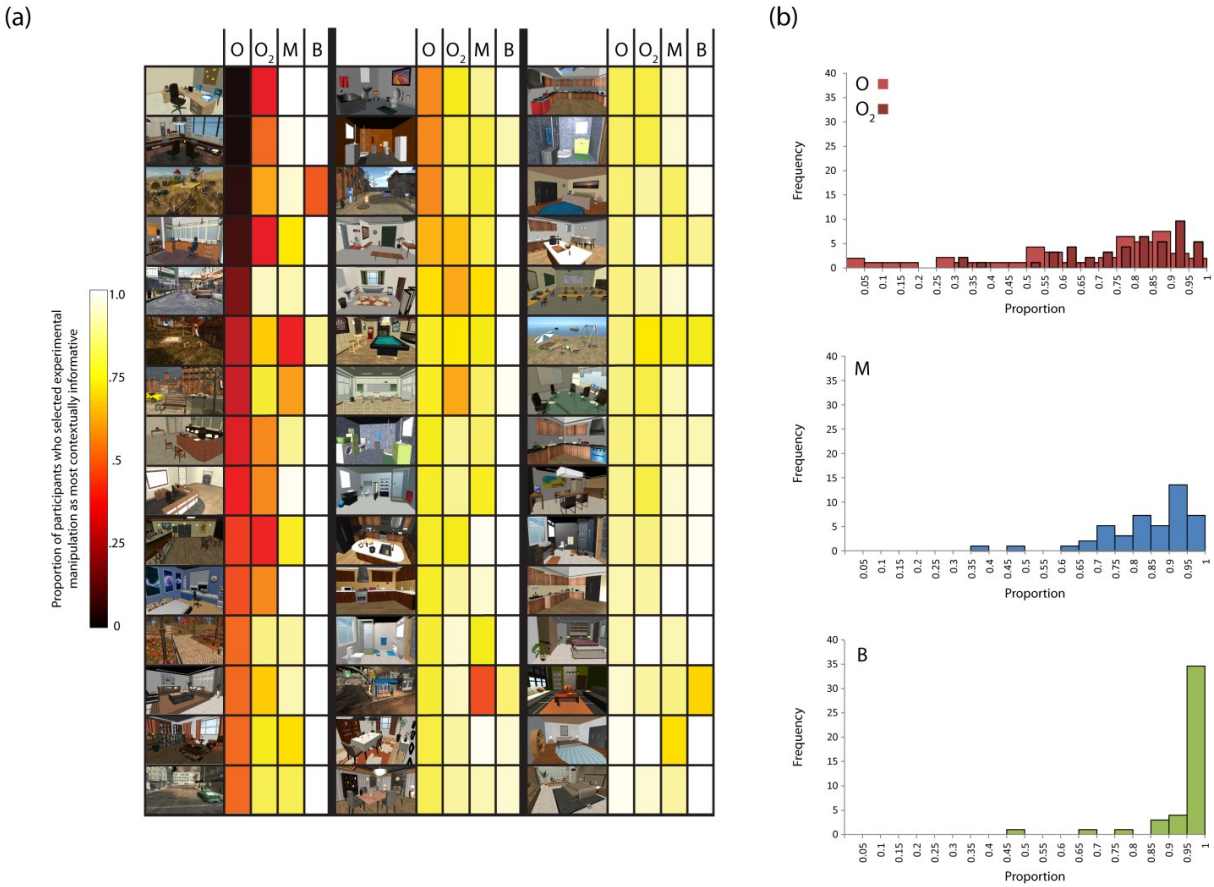

Figure 2: Part (a) of this figure depicts the proportion of observers that selected our chosen manipulation of a cue to be the most informative version of that cue for a target detection task for each of the 45 scenes. The O column corresponds to the object co-occurrence condition, the M to multiple object configuration, and B to background category information. The O2 column depicts the results from a follow-up task where we instead asked participants to click on the object that they would expect to be physically closest to the target object, the color of the cell representing the proportion of times the participants selected the co-occurring object in that task. Part (b) shows the histogram distribution of the proportions depicted in part (a).

There were many more instances of poor verification of the object co-occurrence (O) manipulation in the first task, likely because there were so many possible objects to choose from, justifying further exploration with our follow-up task. Figure 2, column O2, shows the proportion of times observers' selected our experimentally defined co-occurring object when instead asked to select the object they would expect to be closest to the target object (therefore most informative of

the target object's location) in a fully cued scene. The distribution of those proportions is more similar to that of the M and B verifications, with an average of 77% of observers selecting the experimentally defined co-occurring object. Three images with a television remote target were discarded completely from these analyses because their chosen co-occurring object (television) was never selected as the spatially closest object and instead an alternate object in the scene was selected more than 80% of the time.

Quantification of the relative contribution of each cue to spatial informativeness. Whereas the previous analysis answers the binary question of whether the cues are spatially informative (yes or no), we are also interested in how spatially informative each cue is of the target locations. Therefore we assessed the extent to which the selections made by participants in these tasks were related to and could predict the selections made by 60 participants who reported the expected target location in images containing all types of contextual cues. First, we evaluated the average distance of the expected target location judgments made by observers viewing images with a single type of contextual cue to the mode of the expected target location made by observers viewing images containing all types of contextual cues. Our expectation is that the more informative a cue is about the target location, the closer the expected target location selections for that cue will be to the mode of the selections made in images with all types of cues. For each image and cue type, we calculated the average distance of each expected target location for that cue type from the mode of the expected location in an image containing all cues. The baseline was calculated in the same way using the same number of randomly chosen selections from randomly selected images of the same cue type. We calculated the difference between the distance of actual selections and baseline selections for each image, and averaged that result, shown in Figure 3.

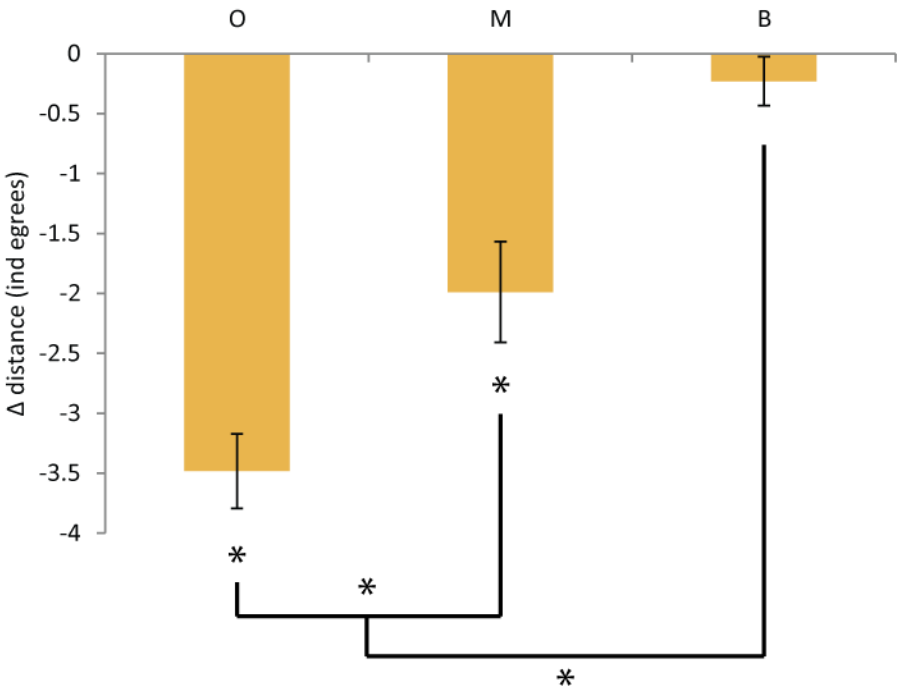

Figure 3: The spatial informativeness of a cue, calculated as the change in distance relative to a baseline calculation of expected target location selections in scenes containing one cue from the mode of expected target location selections in scenes containing all three cues.

A greater negative value in Figure 3 indicates a greater reduction from baseline of the distance of individually cued selections from the mode of fully cued selections, thus indicating a greater degree of spatial informativeness. There was a significant effect of cue type on overall increase in spatial informativeness, $F(2, 141) = 25.13$, $p<0.001$. The distances of expected locations resulting from object-based cues (O and M) were significantly closer to the mode of the fully cued expected target locations than the baseline control selections (O: -3.5°, $t = 11.16$, $p < 0.001$; M: -2.0°, $t = 4.72$, $p < 0.001$). This was not the case for the background-based cue (B: -0.23°, $t = 1.12$, $p = 0.13$). Furthermore, the increase in spatial informativeness of object co-occurrence information was greater than that for multiple object configuration information (post-hoc comparison difference = 1.49°, $p = 0.004$). Finally, we performed a contrast between the object- and background-based information, indicating that object-information was significantly more spatially informative than background information (difference = 2.37°, $p < 0.001$).

Second, we wanted to assess how well the mode of the expected target location selections in single cue images could predict the mode of expected target location s in images containing all cues. For each image, we obtained the x- and y-coordinates of the mode of all participants' expected target locations for the images with only individual contextual cues (O, M, and B tasks in the MTurk experiment) and the fully cued images (OMB). The O, M, and B coordinates were used as predictor variables of the OMB coordinates of the judgments in a linear regression model, summarized in Table 3. The x-coordinate of the individually cued click locations accounted for 79% of the fully cued x-coordinate location selections, $F(3,41) = 51.90$, $p < 0.001$, $R^2 = 0.79$. The y-coordinate selections for the individual cue images accounted for 77% of the fully cued y-coordinate click variance, $F(3,41) = 46.42$, $p < 0.001$, $R^2 = 0.77$. Importantly, the only individual cue that contributed significantly to predicting the fully cued expected target location judgments was the object co-occurrence cue (O), for both the x- and y-coordinates. This serves as another useful verification of our manipulation. Because we selected the co-occurring object to be spatially close to the target object, to the extent that observers are utilizing this information and selecting target locations that are proximal to the co-occurring object when present, these two measures will be highly correlated (see zero-order r between O and OMB for both the x- and y-coordinates in Table 3) and predictive of one another (see the partial correlations and coefficients for O in Table 3). The other manipulations were not as tightly spatially coupled with the target object, so we would not expect observers' judgments of the target location in those tasks to necessarily be as predictive as object co-occurrence of observers' judgments in the fully cued task.

	Zero-order *r*						β		pr		*b*	
	M		B		OMB							
	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
O	0.39**	0.50***	0.06	0.48***	0.88***	0.88***	0.84	0.84	0.86	0.82	1.07***	1.12***
M			0.14	0.52***	0.44**	0.49***	0.10	0.07	0.20	0.11	0.14	0.09

B					0.12	0.45**	0.06	0.01	0.13	0.02	0.09	0.01
									Intercept:		32.10	28.91

Table 3: Summary of results using the expected target locations collected from observers who viewed scenes with individual cues to predict the expected target location judgments of observers who viewed fully cued scenes. **: $p < 0.01$, †: $p < 0.001$

Discussion

We assessed observers' ratings of whether our contextual cue manipulations affected how informative the scene was for locating a target object estimated the amount of spatial informativeness provided by each cue. Images with the background category and multiple object configuration cues disrupted were selected to be less informative of a target object's location than images with those cues intact, thus supporting the conclusion that our cues were informative to the target search task. There was some disagreement between independent observers' ratings of the informativeness of the co-occurring object. One group of observers, when asked to select an object in the scene that provides the most information about the location of a target object, selected different objects than a group of observers who were asked to select the object in the scene that they would expect to be closest to a target object. Our assumption was that the former task instructions would indirectly assess participants' spatial expectations about object location relations, but that was not the case.

Furthermore, we were able to compare the expected target location selections from a task where observers viewed single cue images to the target location expectations of observers viewing a images containing all cues. Object information was more spatially informative of the expected target location than background information overall. Individually cued selections were also predictive of fully cued selections. Target location expectations resulting from the object co-occurrence cue alone accounted for the greatest proportion of variance in the fully cued target locations expectations, as would be expected given its tight spatial coupling with the experimentally selected target location. If

observers in the verification task were not using the co-occurring object as a basis for selecting the target location, it would not be as explanatory of the expected target location in the fully cued scenes.

IV: Manipulating the presence of individual spatial cues

With the informativeness of each contextual cue well established, we are now interested in manipulating the presence of the cues and scenes during observers' search for target objects within the scenes. We generated images that contained no cues, one cue, two cues, or all three cues in order to assess the relative influences and interactions of each contextual cue on visual search behavior. We evaluated the influence of the various contextual cues on target detectability, changes in decision criterion, and eye movement guidance. We also conducted a separate assessment of the inherent contextual information provided by a cue by utilizing observers' explicit judgments about expectations of target locations. We show that there is a relation between search eye movement behavior for each image and the scene's inherent information about expected target locations. In this chapter we also describe in greater detail the controls in place to ensure that the scene manipulations affected only the relative cue information and no other confounding factors (target/background contrast, physical plausibility of target location, target eccentricity, and crowding) that might influence visual search behavior.

Methods

Participants

Data for the main experimental object search task (n = 160) were collected from undergraduate students at the University of California, Santa Barbara who participated in exchange for course credit. All participants were verified to have or indicated they had normal or corrected-to-normal vision and provided informed written consent. Data for the explicit judgments about

expected target locations in full context (n = 60) the single context cue explicit judgments task (n = 360) were collected as described in Chapter III.

Design

The independent variable in this experiment was the type of contextual information contained in the image. There were three contextual cues: object co-occurrence (shortened to O in the statistical results and figures), multiple object configuration (M), and background category (B). Each scene either contained all cues (OMB), a combination of two cues (OM, OB, MB), just one cue (O, M, B), or no cues (None), resulting in eight levels of contextual information. The dependent variables analyzed were average sensitivity (d') and bias for the target search task , the average distance of each observers' closest fixation to the target or expected target location, and the average time it took for observers to fixate within two degrees of the target location (target present trials only). Images were Latin-square counter-balanced across conditions, resulting in eight possible display configurations. Twenty participants were run for each configuration, within which trial order and target presence on a given trial was randomized.

Stimuli

Images for the experiment depicted indoor and outdoor real scenes: living rooms, bedrooms, kitchens, beaches, city streets, etc. The images were created by taking screen captures from the virtual camera view of 3D scene models created using Unity 3D (Unity Technologies, Bellevue, WA, USA). Each image subtended 25 x 15.4 degrees of visual angle and was displayed in the center of a computer monitor on a grey background. Object cues were displayed as black text while the observers maintained their gaze on a fixation point preceding image appearance.

Each scene was associated with a specific target object that participants were instructed to search for during the target detection task. There were a total of 16 unique target objects, each used three times for a total of 48 unique scenes. Eight versions of each scene were generated

corresponding to each level of contextual information in the experimental design (subsequently labeled as None, O, M, B, OM, OB, MB, and OMB where O indicates object co-occurrence information was present, M indicates multiple object configuration information, and B indicates background category). The object co-occurrence cue was removed by deleting the co-occurring object. The multiple object configuration cue was removed by jumbling all objects (except the target and co-occurring object). The background category cue was removed by swapping the background with that of a scene from a different category. Consistent with previous scene perception research (Henderson & Hollingworth, 1999), any element of the scene that that could plausibly be moved or re-configured (e.g., table, bed, chair, cabinet) was treated as an object, whereas non-manipulable or structural elements (e.g., wall, floor, sky, mountain) were treated as part of the background. Each scene version also had a target present and absent screen capture, resulting in a total of 768 images. Figure 4 shows an example of the eight target present versions of one of the 48 scenes used in the study. Each participant saw each scene only once, and was therefore shown only 48 of the 768 images depending on their condition (which determined which contextual information condition they would see for each scene) and the random determination of target presence on each trial.

When creating the scenes across contextual conditions, we were careful to control for four possible low-level confounding factors that could influence behavior alongside contextual information:

Target/Background contrast. For scene changes eliminating scene gist that required a modification to the background immediately behind the target, the contrast between the target and background was held constant. We verified this by comparing the average saliency (Itti & Koch, 2000; Walther & Koch, 2006) of the target in the images with the original background (M = 0.074, SD = 0.22) to the conditions where the background was swapped (M = 0.076, SD = 0.24) and confirming

there was no significant difference in target saliency between them ($t(94) = -0.04$, $p > 0.25$, CI = [-0.096, 0.092]).

Physical plausibility of target location. Some target objects had co-occurring objects that provided them physical support (e.g., a frying pan target with a co-occurring stove). For these scenes eliminating object co-occurrence introduced violations of physical laws (i.e., floating target objects). We chose target and co-occurring objects such that this occurred in exactly 50% of the scenes. In such cases, the jumbled multiple object configuration conditions also contained floating objects so that the property "floating" was not uniquely associated with the presence of the target.

Target eccentricity. The target location never changed across contextual information conditions to ensure that the retinal eccentricity of the target across conditions was held constant. The initial fixation was below the image and was also held constant across conditions.

Crowding. We controlled for crowding around the target by jumbling multiple object configuration information instead of removing the objects completely and by swapping background categories rather than replacing the background with a uniform color for the target search task.

Figure 4: Example of the eight versions of one scene. The target is the cork (outlined in the top-left image) and the co-occurring object is the wine bottle. Each version of the scene contains different combinations of the contextual cues. O = object co-occurrence, M = multiple object configuration, B = background category.

Apparatus

Stimuli were displayed on a 1280 x 1024 pixel resolution Barco MDRC-1119 monitor where each pixel subtended 0.022 degrees of visual angle. Eye tracking data were recorded on an Eyelink

1000 (SR Research Ltd., Mississauga, Ontario, Canada) monitoring gaze position at 250 Hz. Each participant's gaze recording was calibrated and validated using a nine-point grid system. A velocity greater than 22°/s and acceleration greater than 4000°/s^2 classified an event as a saccade.

Procedure

Main experimental search task. Participants (n = 160) were instructed that they would be searching a series of images for a specific object while their eyes were tracked. A trial timeline is depicted in Figure 5. Every trial began with a fixation cross displayed in the horizontal center of the monitor, 0.8 degrees below the bottom edge of where the image would eventually appear. Participants initiated a trial with a key press once they had fixated the cross and were required to maintain fixation on that location for a randomly selected interval ranging from 500 – 1500 ms. During this time, the fixation cross was replaced with the name of the object (e.g., CORK) that participants were to search for. After successfully maintaining fixation, the image would appear for 1.5 seconds and participants could move their eyes freely around the image to search for the object. Once the image disappeared, participants judged whether or not the target object was present using a 10 point confidence rating scale collapsed into binary yes/no decisions for analysis. Each participant completed a total of 48 trials in randomized order with contextual information for each trial determined according to the counter-balancing described in the experimental design.

Figure 5: Timeline of the main experimental task. Participants aligned their fixation with a pre-cue (cross) and initiated the trial to see the name of the object for which they would be searching. They had 1500 ms to perform the visual search task, and then they indicated whether the target was present and confident they were with their judgment. Confidences were collapsed to binary yes/no decisions.

Statistical analysis.

Index of detectability and criterion. Some observers in some contextual cue conditions had perfect hit rates or false alarm rates, preventing computation of an index of detectability for those observers to be used to calculate the standard error. We thus utilized bootstrap resampling methods (Efron, 1979) from all conditions and observers to estimate the error in sensitivities and criteria for experimental conditions. For each condition, we obtained 10,000 bootstrap samples and calculated average hit rates and false alarm rates. The values that separated the center 68.29% (equivalent to a standard error for normal distributions) of the distribution of resampled sensitivities and criteria were used to estimate the confidence interval of the means. The statistical comparison

between the OMB and None conditions was calculated by taking the differences between each of the 10,000 samples for each condition and calculating the proportion of those differences that were greater than zero.

Eye movement guidance. To assess the guidance of eye movements to the target location, we calculated the average distance of the closest fixation on a given trial to the target location (target present trials). For target absent images we calculated the distance of the closest fixation to selections of expected target location for scenes with the joint presence of all cues (OMB; see above for description of procedure). We ran a repeated-measure ANOVA to assess the differences between contextual information conditions for this measure. The effect sizes for the planned comparisons were estimated using Cohen's d. We also performed three contrast analyses to assess the overall effect of each individual contextual cue. For example, to assess the O-effect, we conducted a contrast between the None, M, B, and MB conditions to the O, OM, OB, and OMB conditions. The effect size for these analyses was estimated with partial-eta squared. Identical analyses were performed to understand the time-course of the guidance of fixations to the target location. To estimate the time-course, we calculated the time until a fixation landed within two degrees of the target location on target present trials across images and observers.

Results

Contributions of Scene Context Cues to Target Detectability

We first looked at the effect of contextual condition on observers' sensitivity (d') for target detection. The index of detectability was estimated using the hit rates and false alarm rates across observers for a given contextual information condition. Figure 6a shows the average sensitivity (d') across participants for detecting the target object and the average observer criterion for each contextual information condition. Observers' accuracy at detecting the target was 0.90 d' units (50%) higher when searching for targets in images containing all contextual cues than those

containing none (None vs. OMB; $p < 0.001$).When evaluating the statistical effects of individual types of context, we calculated the difference between all four conditions with a particular type of contextual cue present and all four with that cue absent (analogous to an ANOVA contrast). For example, to evaluate the contribution of object co-occurrence we evaluated the differences between O, OM, OB, OMB and None, M, B, MB respectively. Adding object co-occurrence information significantly increased observers' sensitivities by 0.27 d' units on average and close to significance (15% increase, $p = 0.051$). Adding multiple object configuration information significantly increased observers' sensitivity by 0.41 d' units on average (23% increase, $p = 0.011$). Finally, adding background category information did not significantly affect observers' sensitivity by 0.15 d' units on average (9% increase, $p = 0.207$).

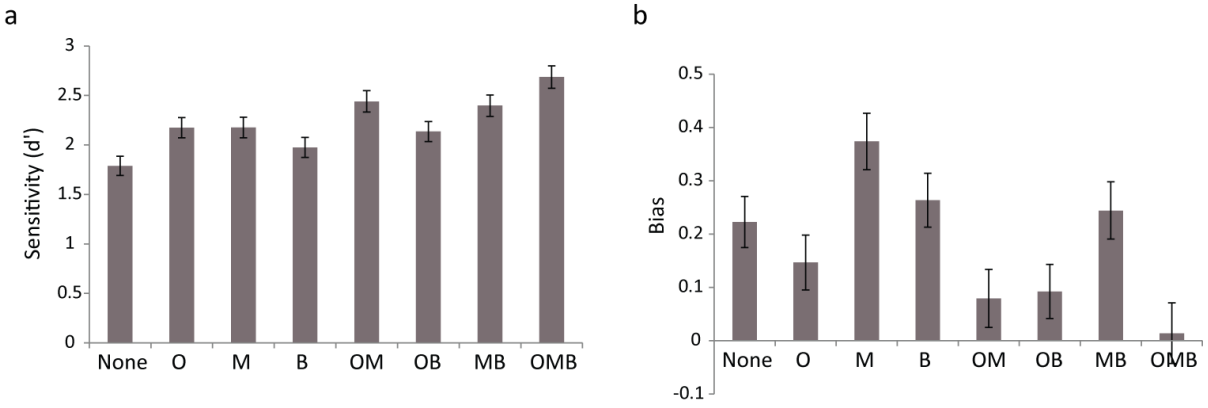

Figure 6: Observers' average sensitivity (d') and bias for detecting target objects in the scenes across each contextual information condition. Error bars represent the center of 68.29% of the distribution of bootstrap resampled measures as an approximation to the standard error of the mean. O = Object co-occurrence, M = Multiple object configuration, and B = Background category.

Contributions of Scene Context Cues to Decision Bias

Because the index of detectability varies across conditions a change in decision criterion is expected for an observer that tries to maximize proportion correct by placing the criterion optimally at d'/2 (for 50 % target prevalence). We thus estimated the bias relative to an optimal decision criterion for 50 % target prevalence. A bias of zero corresponds to an optimal placement at d'/2. In

general Figure 6b shows that across all conditions observers were biased to more frequently say "target absent" placing the criterion above d'/2.

The joint presence of all the contextual cues significantly decreased the bias by 0.21 units ($p = 0.003$) and led to a criterion placement close to the optimal. We also evaluated the independent contributions of adding each individual contextual cue to the decision bias. We contrasted the bias for all conditions with an additional contextual cue to the conditions without that cue (e.g., OM vs. M; OB vs. B; etc). Adding object co-occurrence decreased observer bias by 0.08 units ($p < 0.001$) making observers' criterions closer to optimal. Multiple object configuration and background cues had smaller effects on bias which did not reach statistical significance ($p = 0.47$ and $p = 0.08$, respectively). These results provide evidence for an interaction between the individual contextual cues given that the effect on bias was much stronger when multiple object configuration and background category information were added along with object co-occurrence information (a .21 unit shift) than when object co-occurrence information was added on its own (a 0.08 unit shift).

Eye Movement Guidance

In addition to exploring observers' target detection accuracy based on their perceptual decisions, we also investigated the effect of manipulating different types of contextual information on eye movement guidance. We first evaluated whether contextual cues increased the guidance of eye movements towards the target. For each image, for each observer, we calculated the distance of the closest fixation to the target location (for target present trials) or expected target location (for target absent trials). For each observer, we averaged this value across the 8 trials for each contextual information condition. Figure 7 shows the average distance of the closest fixation on a given trial to the target location for each of the contextual cue conditions. The joint presence of all contextual cues (OMB) reduced the distance of the closest fixation to the target relative to the no

contextual cue (None) condition (mean difference = 0.703°, CI = [0.313 1.094], $p < 0.001$, $d = 0.714$). We evaluated the independent contribution of adding each individual contextual cue by using a contrast analysis across conditions (O vs. None; OB vs. B; OM vs. M, etc.) similar to the analysis utilized for perceptual performance (see Methods for details). Adding object co-occurrence ($F(1,139) = 21.682$, $p < 0.001$, $\eta^2_{partial} = 0.135$) and multiple object configuration ($F(1,139) = 24.269$, $p < 0.001$, $\eta^2_{partial} = 0.149$) strongly aided guidance toward the target location, whereas adding background category information more weakly contributed to the guidance of eye movements ($F(1,139) = 4.785$, $p = 0.030$, $\eta^2_{partial} = 0.033$). Figure 8 shows a representative example of observers' fixations (all fixations included) for each condition for a sample image.

We also investigated the influence of the contextual cues on eye movements in the target absent images which provide a measure of guidance toward expected target locations in the absence of any guidance provided by the physical presence of the target. We obtained explicit expectations of the target location from 60 independent observers (not participating in the main search experiment) which viewed the scenes with all contextual cues and were asked to select the location most likely to contain the target. The joint presence of all contextual cues (OMB) significantly reduced the average closest fixation distance to the expected target location (mean difference = 1.821°, CI = [1.396 2.247], $p < 0.001$, $d = 1.60$)

We evaluated the individual effects of each type of contextual cue on eye movement guidance. Adding object configuration information significantly improved the guidance of observers' eye movements toward the expected target location ($F(1,131) = 287.57$, $p < 0.001$, $\eta^2_{partial} = 0.687$), as did adding multiple object configuration information ($F(1,131) = 71.602$, $p < 0.001$, $\eta^2_{partial} = 0.353$). Adding background category information did not significantly affect the guidance of observers' eye movements ($F(1,131) = 1.053$, $p > 0.250$). Together the eye movement analysis is in agreement with the index of detectability results showing that adding the background did not

contribute to benefits in detectability and eye movement guidance as much as object co-occurrence and multiple object configuration.

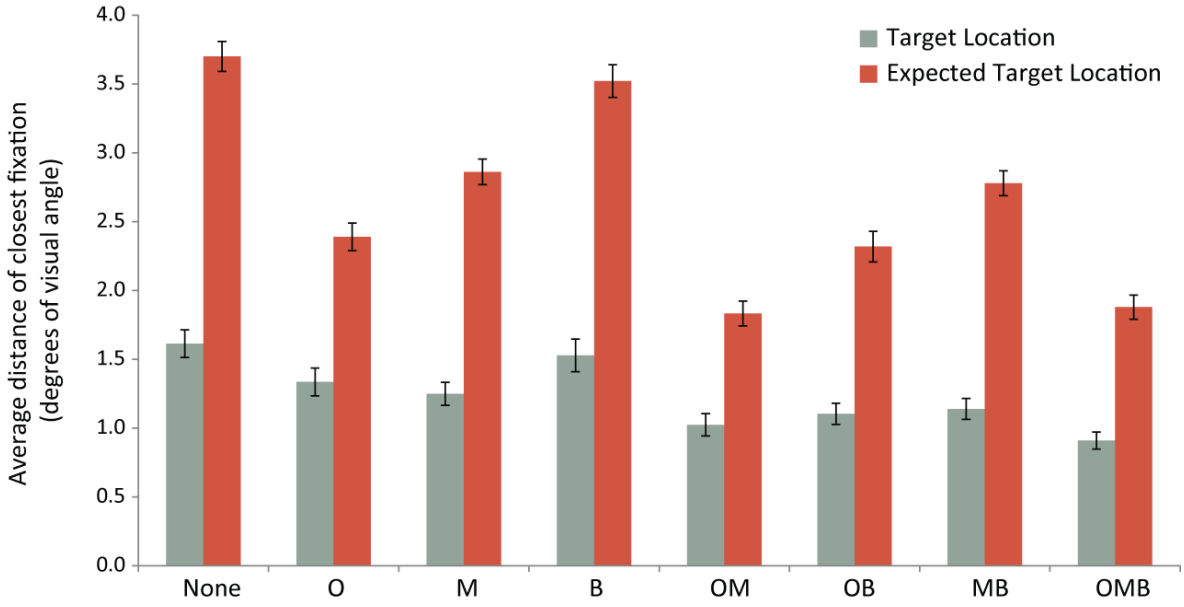

Figure 7: Average distance of the closest fixation on a given trial to the target (target present trials) or expected target (target absent trials) location. Error bars represent the standard error of the mean.

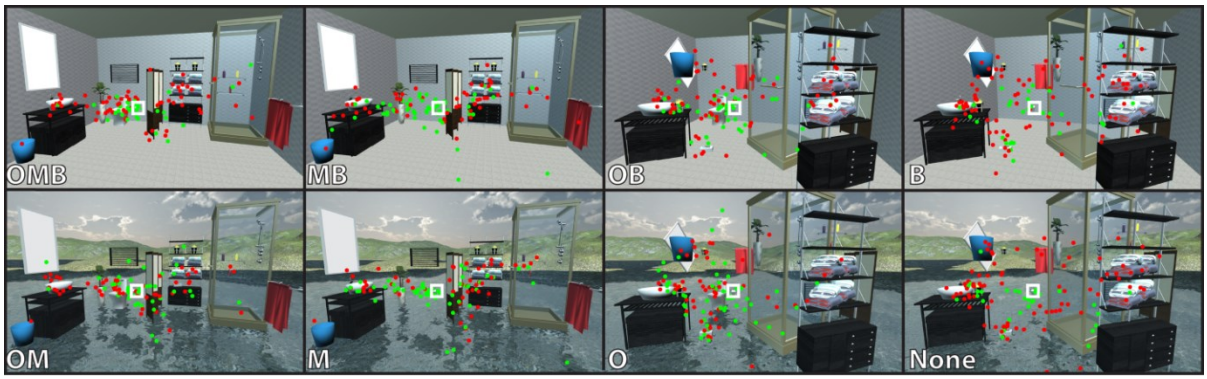

Figure 8: Every observer's fixations for each of the 8 conditions for a sample scene. Green fixations were for target present trials, red fixations were for target absent images, and the white square indicates the target (toilet paper) location.

Time to foveate the target.

In addition to exploring the guidance of eye movements in terms of proximity to the target location, we also investigated whether the addition of each contextual cue prompted eye

movements toward the target location more quickly. We assessed this by calculating the average time it took for participants to fixate within two degrees of the target location. Trials in which the observer did not fixate within two degrees of the target were removed from analysis. We assessed the differences between contextual conditions using a repeated-measure ANOVA with the same planned comparison and contrasts as the closest fixation distance analysis. Target absent trials are excluded in this analysis because only 47% of trials contained fixations within two degrees of the expected target location, resulting in only 8 comparison rows after listwise deletion in the repeated-measures ANOVA. The joint presence of all context cues significantly reduced the time to foveate the target (OMB vs. None, mean difference = 116 ms, CI = [42.39 189.853], $p < 0.001$, $d = 0.60$). Adding object co-occurrence information and multiple object configuration information significantly decreased the time until a fixation landed within two degrees of the target location ($F(1,120) = 29.366$, $p < 0.001$, $\eta^2_{partial} = 0.197$; $F(1,120) = 9.114$, $p = 0.003$, $\eta^2_{partial} = 0.071$, respectively), whereas adding background category information failed to illicit a significant effect ($F(1,120) = 0.997$, $p = 0.320$). See the supplementary information for a figure depicting the average time for participants to fixate the target across all eight conditions and refer to figure 9(e) for a depiction of the average contrast effect for each contextual cue.

Summary of single cue effects

For each metric we have considered thus far, Figure 9 shows the effect of each cue on that metric in isolation. For comparison, figure 8a depicts the increase in spatial informativeness provided by each cue, as discussed in Chapter III. Figure 9b-e display the benefits of each cue by averaging the difference between all contextual information conditions with that cue versus without that cue (e.g., the average of None – O, M – OM, B – OB, and MB – OMB for the "Add O" effect) for sensitivity (d'; 9b), the closest fixation distance to the target (9c) and expected target (9d) location,

and the time to fixate within the target region (9e). A significant result from our previous contrast analyses would be reflected in this figure as a value significantly different from zero.

A clear trend can be seen such that adding object co-occurrence and multiple object configuration information has a stronger effect compared to that of adding background category information (statistically analyzed via a contrast where O and M were each weighted by 0.5 and B by 1). Although background category information in isolation sometimes provides information or significantly increases observer sensitivity at target detection ("see Add B" in Figure 9a and b), object co-occurrence and multiple object configuration do so with much greater strength. We tested this by performing a contrast between the average effect of the object-based cues (O and M) and the non-object cue (B). The object-based cue effects were significantly larger than the non-object-based cue effects in all cases (O/M vs B: 9a, $p < 0.001$; 9b, $p = 0.018$, estimated from bootstrap re-sampling; 9c, $p = 0.033$; 9d, $p<0.001$; 9e, $p = 0.026$).

We also compared the two object-based cues to the background category cue (i.e., O vs B and M vs B) for the five metrics shown in Figure 9 a-e. Of the ten possible comparisons, four of the pairwise comparisons (controlling for false discovery rate) did not reach significance (Figure 9b, d': O vs B; Figure 9c, closest fixation distance to target location: O vs B and M vs B; Figure 9e: time to fixate within the target region: M vs B). A final important note is that the more spatially informative a cue is of the target location (9a), the more contextual guidance it provides during target search (9d).

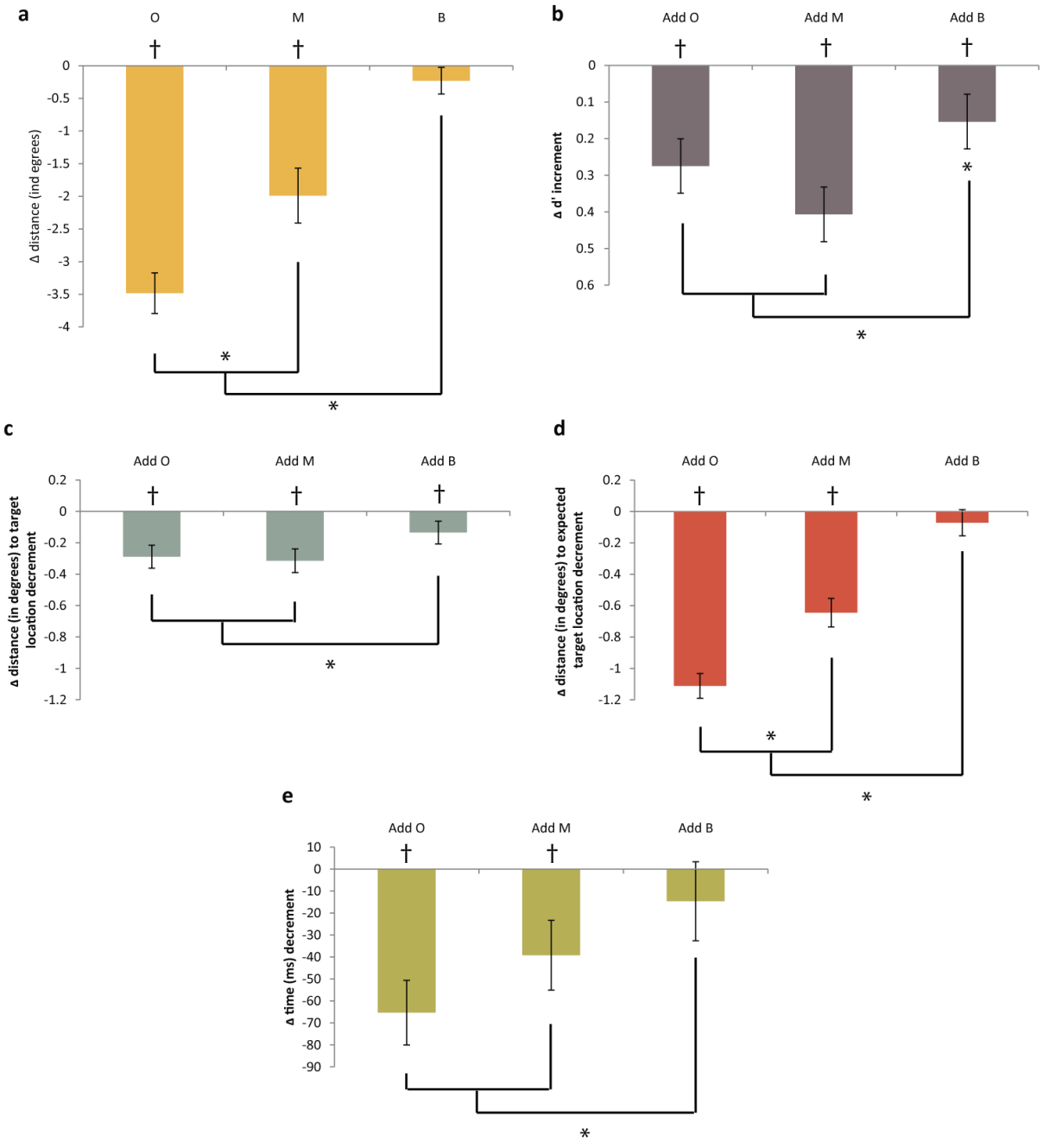

Figure 9: A comparison of the average effect on each metric of adding a single type of contextual cue relative to a baseline measure without that cue. Panel A displays the increase in spatial informativeness relative to baseline (random selections) of explicit observer judgments of expected target location for each individual cue. Error bars represent the standard error of the mean difference calculated across images. Panel B shows the average increase in d' for each cue relative to the complementary condition without that cue (e.g., the average of O-None, OM-M, OB-B, and OMB-MB). Error bars represent the inner 68.29% of complement averages calculated for each of the 10,000 bootstrap re-samples. Similarly, panels C, D, E show the same average decrease in distance of the closest fixation to target location (c), expected target location (d), and time to fixate the target region (e) for each cue relative to the complementary condition without that cue. Error bars represent the standard error of the mean of the four complement means for each observer. All cases marked (†, $p < 0.05$)

above the individual cue bar are significantly different from zero. Comparisons marked (*, $p < 0.05$) between individual conditions are significantly different from each other or the contrast between O and M (each weighted by 0.5) and B (weighted by 1) is significant.

Relating scene-specific eye movement guidance with the scene's expectations of target locations

Finally, we investigated whether we could find a relationship betweeen the explicit judgments about expected target locations and search guidance of eye movements for each scene. We hypothesized that scenes for which a contextual cue provided much information about the target location would also have an associated higher degree of eye movement guidance toward the target. To quantify the information about target location inherent in a contextual cue we calculated the average distance of observer selections of likely target locations for target absent scenes with the individual contextual cues (O, M, B) to that of the target location. We then correlated the distance of observers' explicit expectations with the average distance (across observers) of the closest fixation to the target location in both target present and absent images. Figure 10 shows a scatter plot of these two measures for each scene with an individual contextual cue. All three correlations were significant (O: $r(43) = 0.53$, $p < 0.001$; M: $r(43) = 0.41$, $p = 0.005$; B: $r(43) = 0.57$, $p < 0.001$), indicating that the amount of information provided by a single type of context as to the expected target location is a predictor of the extent to which observers' eye movements will be guided to the target location. Three outlier images were removed from analysis because their mean distances were over 15° (more than two standard deviations) away from the target location, suggesting that the target was placed at an unexpected location for these scenes. Including such outliers diminished the strength of the correlation, but all three correlations remained significant.

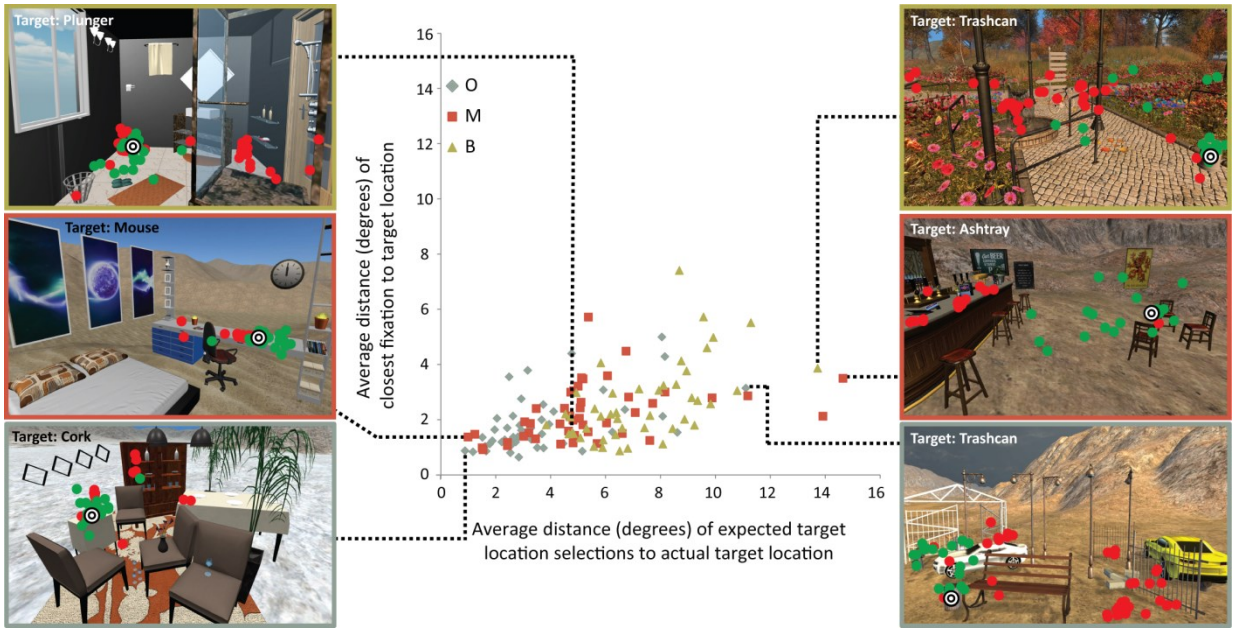

Figure 10: Scatterplot showing the relationship for each image between the average distance of a group of observers' expectations of target location from the actual target location and the average distance of a separate group of observers' fixations to the actual target location across all trials. Representative sample images for each contextual cue are shown for various points on the scatterplot to visualize the expectations of target locations (bright red points within sample images) and the closest fixations to the target location (bright green points within sample images).

Contextual cue combination: A classic test when many cues are available to an observer evaluates whether the combination of multiple cues is consistent with an optimal combination (Trommershauser, Kording, & Landy, 2011). It is common to first assume that the cues are statistically independent and assess how an observer benefits from multiple cues compared to a theoretical prediction based on their single cue performance benefits. We first calculated the isolated effect (d'_{cue}) of each contextual cue relative to the condition where no cues were present according to Equation 1 (see the appendix for the derivation of this equation). We then used Equation 2 (also derived in the appendix) to calculate the predicted d' from the joint presence of the contextual cues assuming that the cues are statistically independent and combined optimally (i.e., linearly combined with weights set optimally). We compared this value to the observed effect on d' with the conditions where two or more contextual cues were present. Note that the observed d'

values for a given cue also contain target feature guidance and possible other sources of guidance that contribute to d' when no cues were present (the "None" condition). Therefore, the observed value is thought to contain a d'_{None} effect, treated in the derivations as an independent cue, that is also added onto the predicted value. Equation 3 therefore shows an example calculation of the predicted d'_{OMB} effect.

$$d'_{cue_i} = \sqrt{d'^2_{cue_i,none} - d'^2_{none}} \quad \text{[Equation 1]}$$

$$d'_{predicted} = \sqrt{d'^2_{cue_1} + d'^2_{cue_2} + d'^2_{cue_3}} \quad \text{[Equation 2]}$$

$$d'_{OMB,predicted} = \sqrt{d'^2_{O,none} + d'^2_{M,none} + d'^2_{B,none} - 2d'^2_{none}} \quad \text{[Equation 3]}$$

Figure 11 displays the average predicted summation of the individual cues (from Equation 2) in comparison to the observed experimental result using the average d' for each combined cue condition. The points lie generally along the identity line, suggesting that observer benefits with multiple contextual cues are consistent with the benefits expected from optimal integration of independent cues. We calculated individual slopes for the 10,000 bootstrap sample point sets while forcing the intercept to be zero. The average slope was 1.01 with 56.77% of the slopes greater than 1, therefore we fail to reject the hypothesis that the cues are being combined linearly optimally.

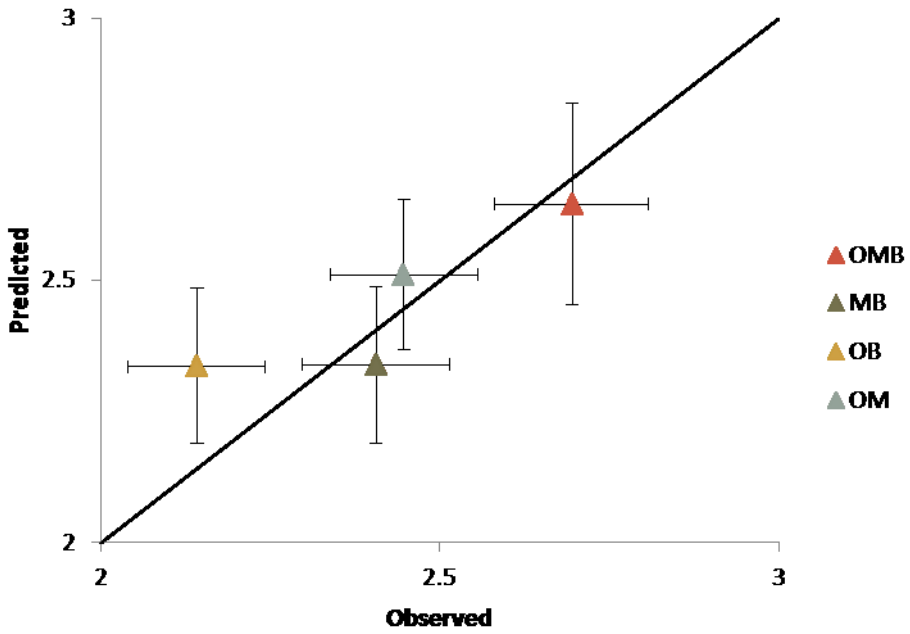

Figure 11: The derived summation of individual cue effects compared to the experimentally observed combined cued effects on d'. These calculations were made using the average d' for each condition with more than one type of contextual cue. The error bars represent the inner 68.29% of the distribution of 10,000 bootstrap re-sampled average derived and observed d' values.

Discussion

The goal of this chapter was to assess the independent contributions of three contextual cues on visual search task decisions and eye movement guidance. We experimentally manipulated these distinct types of contextual information while controlling for many low-level visual properties and other scene properties known to influence visual search (retinal eccentricity, physical plausibility, crowding of the target and saliency of the target against its local background) to demonstrate that object-based sources of context (object co-occurrence and multiple object configuration) illicited stronger benefits to the accuracy of perceptual judgments and the guidance of eye movements than did background information. When provided with object-based contextual information, observers were more sensitive to detecting the target, fixated nearer to the target location (or expected target location), and fixated earlier within the target region.

The contributions of contextual information to eye movement guidance was even larger during target absent trials where the physical presence of the target cannot improve guidance (Malcolm & Henderson, 2009), underscoring its importance in attentional guidance. Furthermore, object co-occurrence was the only source of contextual information that in isolation significantly reduced observers' decision bias, adding to the evidence that related object information is more tightly coupled with target expectations than background information.

Our findings are interesting to consider in concert with other work that has directly explored the influence of semantic information about a scene on eye movement behavior. One study showed that spatial expectations derived from object and surface placements within scenes can guide target search even when the searched for object does not belong in the scene. Information as to whether an object belonged in a scene was taken to be scene gist, therefore if an object did not belong in a scene, it was interpreted as there being an absence of relevant semantic scene gist information for that object (Castelhano & Heaven, 2011). We have not determined which of our manipulations affect observers' perceptions of scene gist as defined in the Castelhano and Heaven study, but our work clarifies what information within a scene is helpful for eye movement guidance. The Castelhano and Heaven study placed unexpected objects within fully in tact scenes, i.e., the backgrounds and objects in that scene were normal, but a random other object was inserted somewhere and searched for. Our work elucidates that the useful scene property that guides eye movement behavior is the configuration of objects in the scene as opposed to the background or scene environment alone. Our findings are also consistent with evidence suggesting little contribution of background information (defined to be the content of the image that portrays scene gist) to the guidance of attention between semantically related objects on an image memory task (Wu, Wang, et al., 2014). Together our results suggest that, although scene gist information as it is defined in Wu et al. (as the background of a scene) provides useful scene category information to an

observer that can be used in the absence of other sources of information, under natural search conditions, object information drives an observers' perceptual decsisions and eye movement behavior more than background information. Demonstrations of the relatively minmized role of background information relative to object information may seem in contrast with a study that demonstrated fewer required fixations and shorter search times overall to localize a target when observers had access to background information (referred to in the study as "scene context") than when they had access to object information (referred to as "object content"; Pereira & Castelhano, 2014). Critically, Pereira and Castelhano manipulated the presence of object information, but not the relative configuration of objects, nor the inclusion of objects that frequently co-occur with search targets in real scenes. In our scenes, as in natural scenes (Greene, 2013), objects provided much statistical information about the location of other objects in the scenes.

Additionally, our results also show a tight relationship between search behavior and observer judgments about expected target locations. Previous work has shown that image-to-image variations in eye movement guidance could be predicted from explicit target location judgments (Droll & Eckstein, 2010; Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009). Here we show that there is also a relationship between the spatial informativeness (closeness to expected target locations in fully contextually cued images) of the explicit judgments provided by scenes with the individual contextual cues and their benefits to search accuracy and eye movement guidance.

Finally, we have also demonstrated that combining cues is not simply supplying redundant information to observers and that when all are present simultaneously observers are able to perform at a level that is consistent with the possibility that they are able to still independently process each cue. This could suggest a modularity of organization in the brain for dealing with various scene and object-based contextual cues. Especially interesting is that two of the three studied cues were object-based, suggesting either that we are able to sub-divide object information

into separately useful cues, multiple units in the brain are responsible for processing different types of object information, or that the multiple object configuration cue is not extracted by object processing units.

V: Temporal Effects of Contextual Cues

Given our understanding of how observers use and combine the presented contextual cues during a visual search task, we are now interested in better understanding when each cue begins to exert its influence. In this chapter we will, (1) assess the temporal course of the cues' influence on a variety of perceptual judgments and eye movement behaviors and (2) again investigate how the cues are combined to facilitate search performance and how the measured accuracy benefits compare to that predicted from a theoretical prediction of statistically independent cues utilized extensively in the fields of cue combination (Ernst, 2006; Steven S. Shimozaki, Eckstein, & Abbey, 2003; Trommershauser et al., 2011).

We used the images from Chapter IV in which no contextual cues, one type of cue, or all cues were present in combination with a gaze contingent viewing paradigm so that observers could view each image for only one, two, or three fixations. We assessed the index of detectability, bias, and proximity of fixations to the target location as a function of viewing time.

Methods

Participants. A total of 300 undergraduate students at the University of California, Santa Barbara participated in the experiment in exchange for course credit. All participants provided informed written consent and were verified to have normal or corrected-to-normal vision.

Stimuli. The scenes as described in previous chapters were used for this experiment with a few differences. In order to preserve the overall difficulty of the search task between conditions,

instead of using a grey background in trials where the B cue was absent, we mismatched the background on such trials. Ten versions of each scene were created corresponding to the contextual information levels described in the experimental design. An example of each scene is shown in Figure 12.

Figure 12: Sample scene images for a trial in which the participant searched for CORK. The top image shows the scene with all three cues, the middle left shows the scene with only the object co-occurrence cue (O), middle-right with only the multiple object configuration cue (M), bottom-left with only the background category cue (B), and the bottom right with no cues. The sample scenes contain the target. There were five additional complementary scenes with the target object removed. Participants saw one of the ten scenes and their task was to determine if the target object was present, with a known 50% likelihood of target object presence.

Design. We manipulated the type of contextual information present in the stimulus (five levels: None, O, M, B, and OMB) and the number of allowed saccades (three levels: one, two, or three) while completing the task. Each participant served in all of the conditions, resulting in a two-

way (3x5) repeated-measures design. In order to determine which set of images a particular observer would see, we Latin-square counterbalanced the 45 images into groups of 3 across the 15 possible condition combinations. Observers were randomly assigned to one of the 15 image assignment groups and image presentation order was randomized.

Apparatus. Stimuli were displayed on a Barco MDRC-1119 monitor with 1280 x 1024 pixel resolution. Participants positioned themselves on a chin and forehead rest 76cm aware from the monitor so that a single pixel subtended 0.022° of visual angle. Eye tracking data were recorded on an Eyelink 1000 (SR Research Ltd., Mississauga, Ontario, Canada) monitoring gaze position at 250 Hz using a nine-point grid calibration procedure. A velocity greater than 22°/s and acceleration greater than 4000°/s^2 classified an event as a saccade.

Procedure. Participants were instructed that they would be viewing a series of images on the computer monitor and determining whether or not various objects were present in those images. They were told that were was a 50% likelihood that the target would be present in the images. The time course of a single trial is shown in Figure 13. At the beginning of each trial, participants were required to fixate a cross at the bottom-center portion of the display monitor. They initiated a trial by pressing the space bar, at which point the name of the object (e.g. TRASHCAN) they were to search for appeared. They were required to read the object name without moving their eyes. After 500-1500 ms, the image appeared. After the requisite number of saccades (one, two, or three) had been made within the image it was removed and a response screen appeared where participants could indicate how confident they were that the target object was present. Responses of 1-5 indicated the object was absent, 1 being highest confidence, whereas a response of 6-10 indicated the object was present, 10 being highest confidence. Participants' first saccade from the initial fixation location into the image was not counted as part of their allowance and they were not explicitly told that the image display time was dependent on their eye movement

behavior. Instead, they were instructed that the image would appear for a variable amount of time on each trial. No participant reported knowledge or discovery of the display timing criterion.

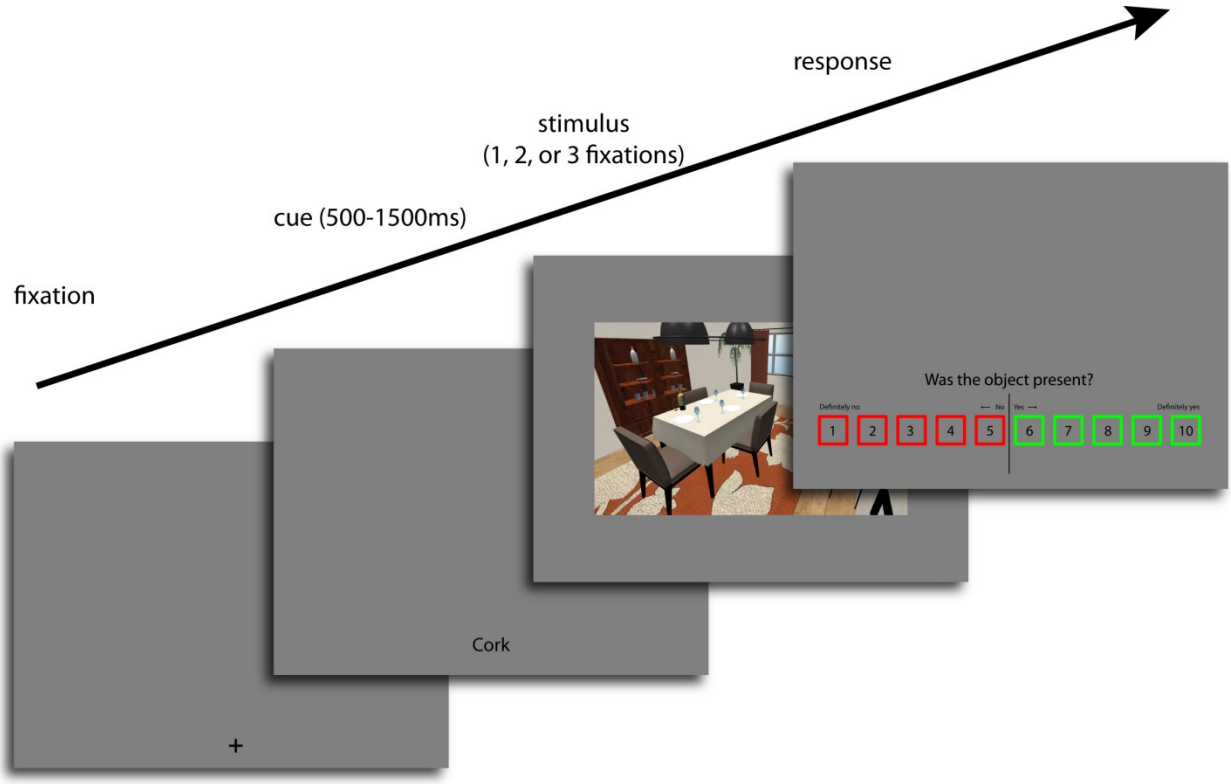

Figure 13: Sample timeline of a single trial during the experiment. The trial initiated once the participant fixated a crosshair and pressed a button, after which they were cued with the target they were to search for. In the experiment, after participants made their first fixation within the image, they were then given either one, two, or three additional fixations to explore the scene. Once they exhausted their allowance, a response screen appeared where the participant indicated whether the target was present and how confident they were in their decision.

Statistical Analyses. In order to quantify observers' performance on the visual search task, we estimated their index of detectability (d') from each recorded hit rate and false alarm rate after collapsing their confidence rankings into binary yes/no decisions about target presence. Because some observers had perfect false alarm or hit rates, we utilized bootstrap re-sampling methods to estimate the variability of d' across observers and perform statistical analyses of differences between the experimental conditions. Specifically, we assessed the distributions of sample mean differences between each condition, and generated p-values from the proportion of differences in the tail above or below zero.

We also analyzed the guidance of observers' eye movements toward the target location using the recorded eye-tracking data. We assessed the distance of the observers' closest fixation to the target location on each trial for target present trials to the expected target location on target absent trials. The expected target location was the mode of selections made by 60 separate observers who freely viewed the full-cued target absent images and clicked where they believed the target would be located from Chapter III. We analyzed the target present and absent data separately, using a two-way repeated measures ANOVA with post-hoc comparisons while controlling for the false discovery rate.

Results

Before considering the effect of various contextual cues on behavior across fixation allowances, we first assessed whether there appeared to be significant differences between fixation latencies in various conditions. Given that the overall viewing time of the scene was dependent on observers' fixations, if they exhibited shorter fixation latencies in a particular contextual condition, a drop in visual search performance may be attributable to having less time to search the image overall relative to other contextual information conditions of equal fixation allowance. We compared the average fixation latencies in each condition using a repeated measure ANOVA and found that there was no significant effect of contextual condition on fixation latency, $F(4,2392) = 2.12$, $p = 0.08$.

Target Detectability. We first assessed observers' ability to detect the target object as a function of the number of fixations manipulated using the saccade-contingent display termination. Figure 14 depicts the sensitivity (index of detectability, d') across each fixation allowance condition for each of the single and multiple contextual cue conditions. First, assessing the main effect of contextual information, we found a significant increase in observers' sensitivity when the multiple object configuration cue was present, or when all cues were present compared to when no cues

were present across all fixation allowance conditions (None v. M, M = 0.27, $p < 0.001$; None v OMB, M = 0.46, $p < 0.001$). After controlling the false discovery rate (Benjamini & Hochberg, 1995) to correct for multiple comparisons, there was no significant difference in the index of detectability for observers searching with the object co-occurrence cue, the background category cue, or no cue at all (None v. O, Mean = 0.11, $p = 0.06$; None v. B, Mean = 0.07, $p = 0.18$). Although this result may suggest that the object co-occurrence and background category cues do not influence target detection task performance within the first three fixations overall, it is important to note that there is a significant increase in observers' index of detectability when the object co-occurrence and background category cues are added to multiple object configuration (M v. OMB, M = 0.19, $p = 0.007$). This demonstrates that, whereas in isolation, neither cue's effect on sensitivity reached statistical significance, when combined they significantly increased task performance.

The more interesting analysis probes the interaction between contextual information and fixation allowance to assess whether a particular cue type is utilized to varying degrees on different fixations. Given that we cannot directly assess the interaction in an ANOVA, first we assessed the increase in the index of detectability across fixation allowance conditions for each type of contextual cue. The increase in d' between the n^{th} — $(n-1)^{th}$ fixation allowance condition was significant in all cases except between the 1^{st} and 2^{nd} M and 2^{nd} and 3^{rd} O fixation allowances. Overall, participants' index of detectability increases as they are given more time to explore the image. In light of these two insignificant increases in performance across fixation allowance, and from visual inspection of the results, we chose to assess the crossed behavior of O and B between the second and third fixation allowances. We calculated the difference between O and B in the third fixation (OB_3), the difference between O and B in the second fixation (OB_2), and then assessed the distribution of $OB_3 - OB_2$ across all 10,000 bootstrap re-sampled indexes of detectability. If there was evidence of a significant interaction, we would expect fewer than 5% of the differences to be greater than zero.

Furthermore, if this distribution fails to show significant evidence of an interaction effect, it is impossible that any other distributions would. In total, 8.4% of the differences were greater than zero, therefore we conclude that there is no interaction between contextual information and fixation allowance. This conclusion is supported by running a two-way repeated measures ANOVA on the PC (proportion of trials correctly classified as target present/absent) data, $F(8,2392) = 0.744$, $p = 0.65$. Therefore, while there are clear differences in utilization of contextual information across all fixations, their change in facilitation of sensitivity at detecting the target is similar as scene exploration unfolds.

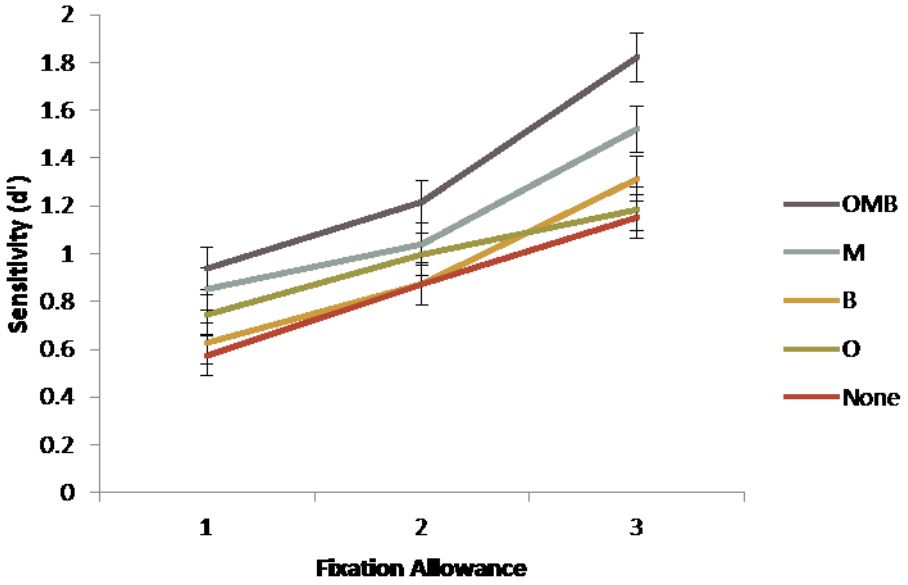

Figure 14: The average sensitivity index of detection as a function of fixation allowance for each contextual cue condition. Error bars represent an estimate of the standard error of the mean, as calculated from the sensitivity indexes delineating the inner 68.29% of the distribution of sensitivity indexes from 10,000 bootstrap re-sampled samples.

Bias. We also explored the change in a participant's bias to make a target absent judgment given that the index of detectability varied across conditions. Figure 15 portrays our measurement of bias, which indicates how far from optimal (d'/2) the average observer criterion was for making target present and absent judgments. There is a clear effect, such that the presence of the co-occurring object was the important cue for reducing bias toward optimality (average overall bias

reduction: 0.23, None v O; 0.25, M v O; 0.24, B v O; fewer than 0.1% of the bootstrapped bias differences were less than 0, i.e., $p < 0.001$). Of note is the increase in bias, corresponding from a slight tendency to over-detect the target to a slight tendency to under-detect the target, in the fully cued condition from the 1st and 2nd to 3rd fixation ($p = 0.007$ and 0.008, respectively; not significant after FDR correction). This could be the result of participants initially perceiving contextually in-tact scenes, consistent with the target object, and thus being likely to assume the target was present when having very few exploratory fixations, but then becoming more confident in rejecting target presence upon further exploration of the scene.

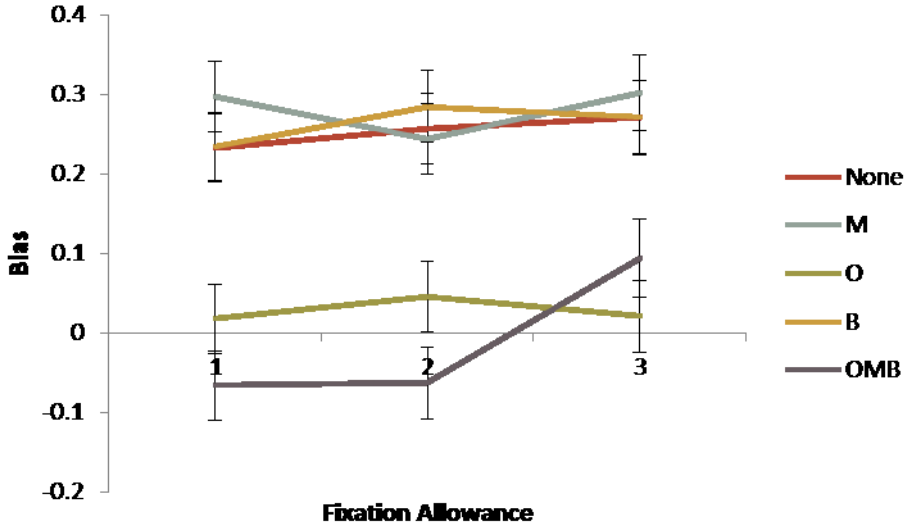

Figure 15: Average bias, where zero corresponds to optimal behavior and a positive score indicates a greater tendency to make a target absent judgment as a function of fixation allowance for each contextual information condition. Error bars represent an estimate of the standard error of the mean, as calculated from the biases delineating the inner 68.29% of the distribution of sensitivity indexes from 10,000 bootstrap re-sampled samples.

Contextual cue combination: These data present another opportunity with which we can compare the actual combination of cue information on detectability index with the optimal combination of cue information. Figure 16 displays the average predicted summation of the individual cues (from Equation 2 in Chapter IV) in comparison to the observed experimental result using the average d' for each fixation allowance condition and also the average d' across all fixation

allowances (from Chapter IV). Again, we see agreement of the points with the identity line, suggesting that benefits to observers' sensitivity at target detection are optimally linearly combined. We again calculated individual slopes for the 10,000 bootstrap sample point sets while forcing the intercept to be zero. The average slope was 0.994 with 45.67% of the slopes greater than one, therefore we again fail to reject the hypothesis that the cues are being combined linearly optimally.

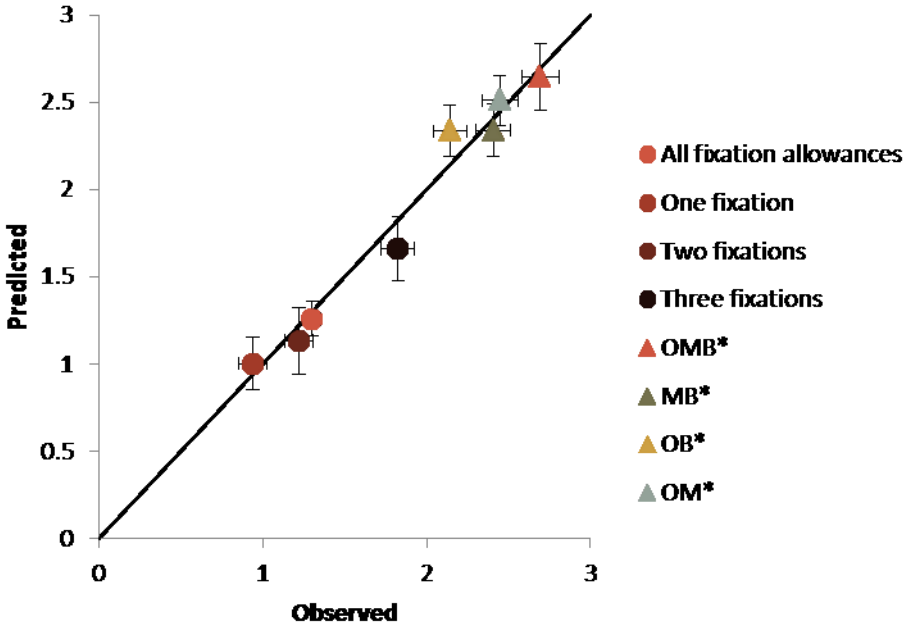

Figure 16: The derived summation of individual cue effects compared to the experimentally observed fully cued effects on d'. These calculations were made using the average d' for each fixation allowance condition (labeled as one, two, and three fixations in the legend) and by averaging across the fixation allowance conditions (labeled as 'all fixation allowances'). The error bars represent the inner 68.29% of the distribution of 10,000 bootstrap re-sampled average derived and observed d' values. The large negative error bar on the two fixation derived d' is a result of instances where performance was better on the no cue condition than on a cued condition for a proportion of samples (see appendix for more information). *These points were calculated identically, but taken from Chapter IV.

Eye Movement Guidance. In order to assess the extent of eye movement guidance offered by each cue on the visual search task, we computed the average distance of the closest fixation to the target (on target present trials) or expected target (on target absent trials) location for each fixation allowance and contextual cue condition. First, we will consider the results for target present

trials, shown in Figure 17. The minimum distance of the closest fixation to the target location was analyzed using a 2-way ANOVA[1].

The effect of contextual condition was significant, $F(4,3931) = 27.97$, $p < 0.001$, as was the effect of fixation allowance, $F(2,3931) = 345.05$, $p < 0.001$. The interaction between fixation allowance and contextual condition was not significant, $F(8,3931) = 1.003$, $p = 0.43$. In order to understand the overall benefits of the various contextual cues to eye movement guidance, irrespective of fixation allowance, we performed post-hoc comparisons controlling for the false discovery rate between the fully cued and no-context conditions as well as between each of the singly-cued conditions and the no-context condition. Eye movements were significantly closer to the target location when all contextual cues were present than when none were present (mean difference = 1.01°, $p < 0.001$). Compared to when no cues were available, eye movements were also significantly closer overall when multiple object configuration (mean difference = 0.41°, $p < 0.001$) and object co-occurrence (mean difference = 0.58°, $p < 0.001$) information was present, but not when background category information was present (mean difference = 0.08°, $p > 0.25$).

We were also interested in assessing the time course of contextual guidance of each contextual cue. For all fixation allowance conditions, background category fails to have a significant effect on eye movement guidance ($p > 0.25$ in all cases). Surprisingly, the facilitative effect of the object co-occurrence and multiple object configuration cues is present upon the first fixation within the image (First fixation: None vs O, mean difference = 0.93°, $p < 0.001$; None vs M, mean difference: 0.62°, $p = 0.002$).

After the first fixation, multiple object configuration information fails to illicit a significant effect on eye movement guidance (None vs M: Second fixation, mean difference = 0.31°, $p = 0.112$;

[1] Although the experimental design was repeated-measures, we have elected to analyze the design as if it were between subjects, sacrificing some experimental power, because there were many instances where for a given context type and fixation allowance there were either no target present or target absent trials (resulting in many empty cells in the repeated measures design).

Third fixation, mean difference = 0.33°, p = 0.09). Object co-occurrence information has a diminished effect on subsequent eye movements, trending toward significance on the second fixation when controlling for the false discovery rate, and significantly greater than None on the third fixation (None vs. O: Second fixation, mean difference = 0.41°, p = 0.03; Third fixation, mean difference = 0.49°, p = 0.01). Given that these are target present trials, these findings are consistent with the possibility that contextual information guides initial eye movements, after which target-feature guidance reduces the contextual guidance. Alternatively, in cases where observers have already located the target and thus completed the task by later fixations, eye movement behavior may no longer be target or context oriented.

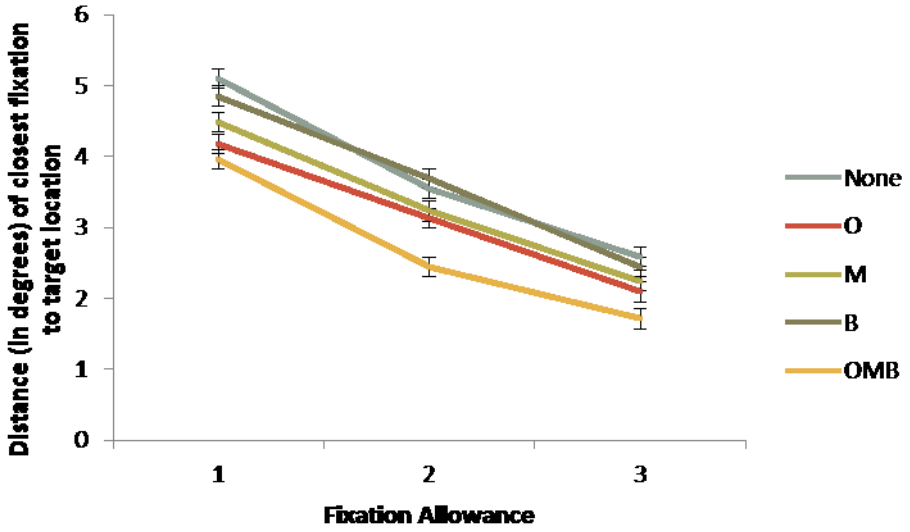

Figure 17: Average distance of an observers' closest fixation to the target location as a function of fixation allowance for each contextual cue condition. Target present trials only are included in this analysis. Error bars represent standard of the mean.

Next, we turn to the results of target absent trials (Figure 18), where target feature information is removed, isolating the contribution of contextual information to eye movements guidance. For these trials, we analyzed the distance of fixations from the mode of the expected target location, as reported by the observers from Chapter III. The minimum distance of the closest

fixation to the expected target location was analyzed using a 2-way repeated measures ANOVA. The effect of contextual condition was significant, $F(4,1188) = 60.93$, $p < 0.001$, was the effect of fixation allowance, $F(2,594) = 194.25$, $p < 0.001$. The interaction between fixation allowance and contextual condition was also significant, $F(8,2376) = 2.033$, $p = 0.039$.

We again performed post-hoc comparisons controlling for the false discovery rate between the fully cued and no-context conditions as well as between each of the singly-cued conditions and the no-context condition. Eye movements were significantly closer to the target location when all contextual cues were present than when none were present (mean difference = 1.26°, $p < 0.001$). Compared to when no cues were available, eye movements were also significantly closer overall when multiple object configuration (mean difference = 0.44°, $p < 0.001$) and object co-occurrence (mean difference = 1.01°, $p < 0.001$) information was present, but not when background category information was present (mean difference = -0.06°, $p > 0.25$).

We were again interested in assessing the time course of contextual guidance of each contextual cue, but also in interpreting the significant interaction between contextual cue and fixation allowance. Again, for all fixation allowance conditions, background category fails to have a significant effect on eye movement guidance ($p > 0.15$ in all cases). In contrast to the results for target present trials, the facilitative effect of the object co-occurrence is present throughout all fixation allowances, (None vs O: First fixation, mean difference = 0.60°, $p = 0.001$; second fixation, mean difference = 1.34, $p < 0.001$; third fixation, mean difference = 1.07, $p < 0.001$), whereas the multiple object configuration cue does not have a significant influence on eye movement guidance until the second fixation within the image (None vs M: first fixation, mean difference: 0.24°, $p > 0.20$; second fixation, mean difference = 0.62, $p = 0.001$; third fixation, mean difference = 0.60, $p = 0.001$).

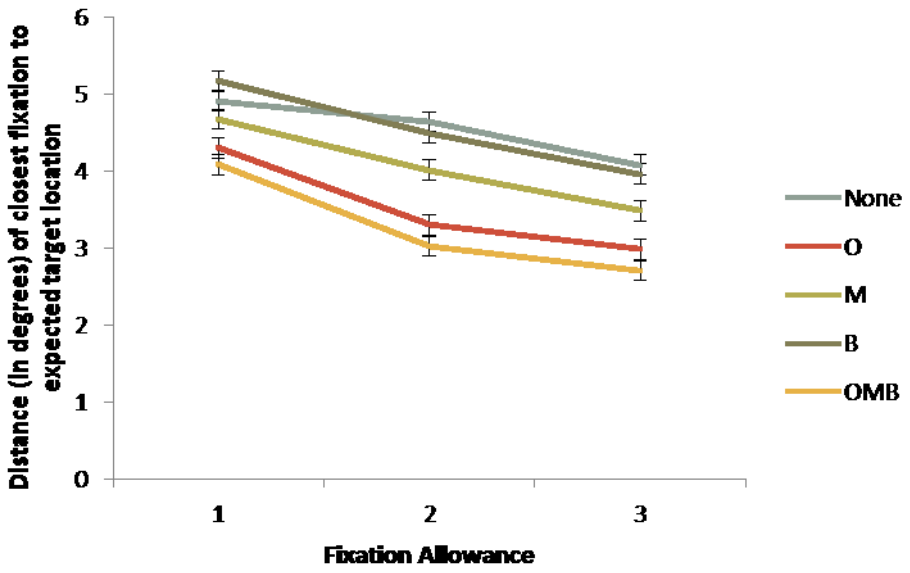

Figure 18: Average distance of an observers' closest fixation to the expected target location as a function of fixation allowance for each contextual cue condition. Target absent trials only are included in this analysis, therefore these data illustrates participants' behavior in the absence of target feature information guidance. Expected target location was calculated as the mode of the location where a separate group of observers expected the target to be located for a given scene. Error bars represent standard of the mean.

Contributions of individual contextual cues to eye movement behavior

In order to further quantify the relative degree to which each contextual cue was being utilized for fixation guidance across fixations, we explored how well the x- and y-coordinates of fixations to scenes with single cues and observer explicit judgements about target location could predict the x- and y- coordinates of the fixations in scenes with all three cues. This will help us understand the individual contributions of each cue to eye movement guidance (relative to the guidance demonstrated when all cues were present) fixation-by-fixation.

For each fixation allowance condition, and contextual cue condition, we calculated the mode of the x-coordinate of all observers' closest fixations to the target or expected target location (for target present and absent trials, respectively) for each image. We did the same to obtain a y-coordinate mode for each fixation allowance and contextual cue condition across images. In this way, we used the x-coordinate fixation modes for the O, M, and B conditions for each image to

predict the x-coordinate fixation modes for each image in the OMB condition in a general linear regression model (and similarly for the y-coordinate) for each fixation allowance. Our expectation was that the amount of information contributed to eye movement guidance during a particular fixation allowance for a given cue type will be captured by its squared partial correlation, i.e., its proportional contribution to the variance in fully cued fixations with the other cue contributions removed. Note that it is likely the case that eye movements between conditions may be collinear, so the model specified here may be under-powered, but this should not affect our interpretations of the partial correlations.

The appendix shows a full table of zero-order correlations, partial correlations, standardized, and unstandardized model coefficients for each x/y-coordinate and each fixation allowance condition. The proportion of variance accounted for in the fixations in scenes with all cues by the fixations in scenes with single cues (the coefficient of determination) was significant for both the x- and y-coordinates for all fixation allowances (one fixation, x: $F(3,41) = 12.60$, $R^2 = 0.48$, $p < 0.001$; one fixation, y: $F(3,41) = 16.22$, $R^2 = 0.54$, $p < 0.001$; two fixation, x: $F(3,41) = 44.80$, $R^2 = 0.77$, $p < 0.001$; two fixations, y: $F(3,41) = 11.62$, $R^2 = 0.46$, $p < 0.001$; three fixations, x: $F(3,41) = 58.03$, $R^2 = 0.71$, $p < 0.001$; three fixations, y: $F(3,41) = 28.27$, $R^2 = 0.64$, $p < 0.001$). Plotted in Figure 19 are the squared partial correlations of each individual cue with x- and y-coordinates of fixations in scenes containing all cues. Error bars represent the inner 68.29% of the distribution of squared partial correlations for each cue from 10,000 bootstrap re-sampled linear regression models. Of note, is the overall lack of explanatory power along the vertical dimension (y-coordinate) by the object co-occurrence cue during the first fixation and by the background category across all fixations.

We performed a contrast-like analysis using the bootstrapped squared partial correlation distributions to assess the differences between the correlations for each condition. We calculated the difference of the summed x and y cue correlations between cues for each fixation allowance

condition (or across fixation allowance conditions) and assessed the proportion of differences above or below zero (depending on the direction of the difference). The results demonstrate that the proportion of variance in eye movements within images containing all cues associated with the multiple object configuration cue is significantly greater than that associated with object co-occurrence and background category on the first fixation (M v B, $p < 0.001$; M v O, $p = 0.046$). Across all fixations, the multiple object configuration and object co-occurrence cues uniquely accounted for a greater proportion of the fully cued eye movement variability than the background category cue (M v B, $p = 0.009$; O v B = 0.001). Therefore, the multiple object configuration cue accounts for the most variance as compared to other cues in the fully cued condition during the first fixation overall, and is the only cue to show diminishing explanatory power overall across fixations. The other cues generally plateau or increase in explanatory power across fixations, suggesting a differential utilization of cue information as time progresses.

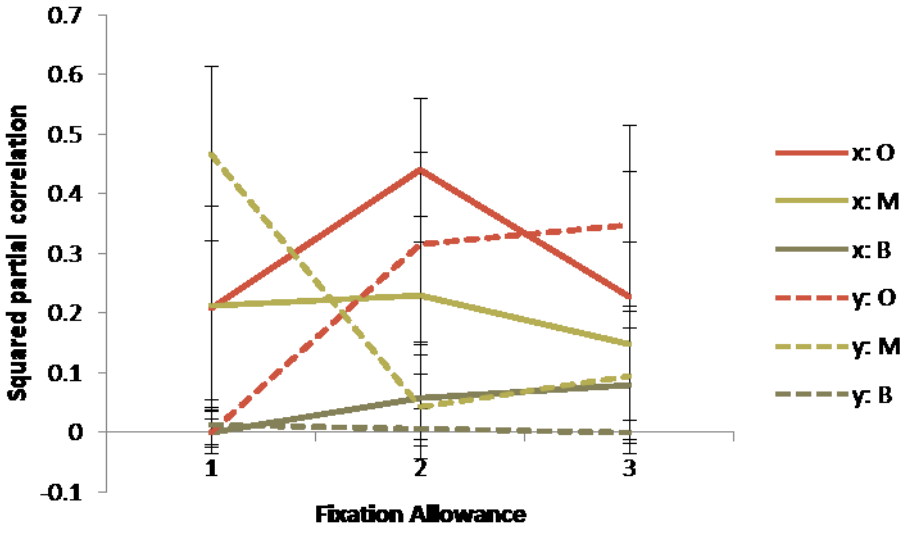

Figure 19: The squared partial correlations of each individual cue with the fully cued coordinates for the x- and y-coordinates. Error bars represent the inner 68.29% of the distribution of partial correlations for each cue from 10,000 bootstrap re-sampled linear regression models.

From Chapter III we know the mode of the expected target location of images containing a single cue, taken from observers who had unlimited time to study the image and make a selection. We can use this result as an upper bound of the information provided by each cue concerning the expected target location. Combined with the results from this experiment, we can thus assess the temporal dynamics of the acquisition of information from each cue relative to the upper bound of information available for each cue. We did this by calculating the correlation between the mode of observers' closest fixations to the target location for each fixation allowance with the mode of observers' expected target locations when viewing images containing a single cue (in both cases).

Figure 20 presents the squared correlations of each individual cue fixation mode to the individual cue expected target locations in the x- and y-coordinate space. Error bars represent the inner 68.29% of the distribution of bootstrapped squared correlations. The results indicate that object information is the only cue information to be increasingly extracted across fixations relative to the upper bound of information available (difference in r^2 between third and first fixations: x-coord = 0.48, $z = 3.17$, $p < 0.001$; y-coord = 0.36, $z = 2.21$, $p = 0.01$). All other cues were not significantly differentially utilized across fixations. This may suggest that, although multiple object configuration information shows earlier influences on eye movement guidance, the information it provides may not be fully utilized until later fixations.

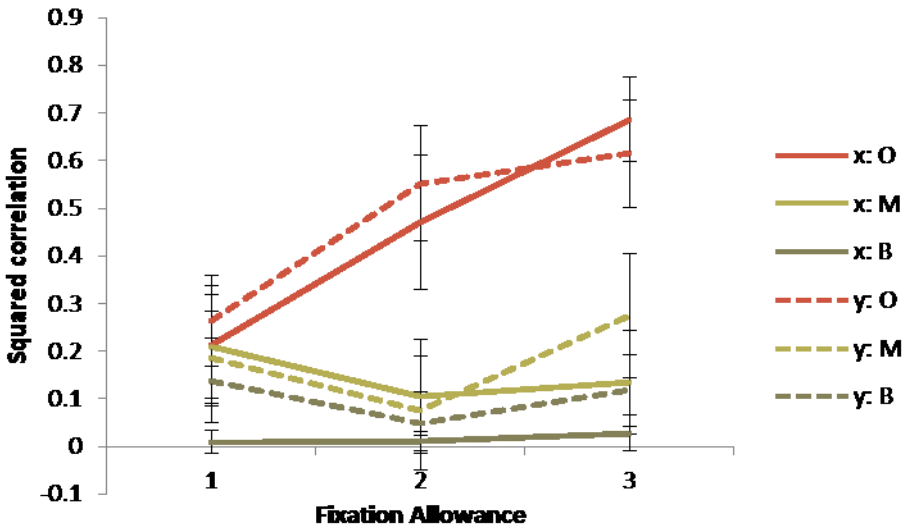

Figure 20: The squared correlations of observers' expected target locations when cued with one type of contextual information with the expected target locations of observers viewing images containing all contextual cues (x- and y-coordinates considered separately). Error bars represent the inner 68.29% of the distribution of squared correlations for each cue from 10,000 bootstrap re-sampled correlations.

Discussion

Temporal Dynamics of the Influences of Contextual Cues

Regarding the temporal influences of various cues, we found that the performance benefits from each cue across fixations increased similarly for each cue (i.e., there was no interaction), but that object-based cues provided greater facilitation of search and perceptual performance behavior overall, consistent with the previous chapter. Although we did not identify a significant interaction between the individual cues on our measure of object detection performance and eye movement guidance on target present trials (index of detectability and distance of closest fixation to target location, respectively), we did discover a significant interaction in target absent trials and significant differences between each cues' usefulness as a predictor of fully cued behavior. Further exploration of the interaction between cue type and fixation allowance on the distance of an observers closest fixation to the expected target location revealed that observers fixated significantly closer to the target region across all fixations when provided with the object co-occurrence cue than when

provided with no cue. Not until the second fixation did observers show similar benefits when provided with the multiple object configuration cue. However, we also observed that fixations within scenes containing only multiple object configuration information account for the most variance in x- and y-coordinate fully cued fixations than other cues alone during the first fixation within a scene. However, multiple object configuration cued fixations' explanatory power decreases as scene exploration unfolds, with object co-occurrence information providing the most explanatory power overall by the third fixation. We take this as evidence that the spatial configuration of objects is initially perceived and utilized by the visual system to guide eye movements to likely general target locations, at which point information about specific objects can be utilized to further localize the target, in agreement with past work (Pereira & Castelhano, 2014).

Contextual Cue Combination Consistent with Statistically Independent Cues

We assessed whether the performance (d') benefits measured in humans was consistent with an optimal linear combination of each cues' independent effect on behavior. We found consistent evidence that the observed performance of humans who viewed full cue images was equivalent to that which would be expected by an additive linear combination of single cues. This result is consistent with that from Chapter IV.

Contributions of Each Contextual Cue to Eye Movement Behavior, Sensitivity and Bias

Again, object information was shown to be more influential in all measures of task performance than background information. Specifically, the object co-occurrence cue accounted for the greatest proportion of variance in the full cue target location expectations, as would be expected given its tight spatial coupling with the experimentally selected target location. Our finding that the object co-occurrence cue was most effective in reducing observer decision bias to make a target absent judgment is also in agreement results from Chapter IV demonstrating the same effect during longer viewing times with a greater variety of contextual information comparisons.

Past work that used contextual information in conjunction with saliency information to predict human fixations in a series of images during a target search task demonstrated that scene context (as extracted from global image features) was influential in determining a region along the vertical dimension of an image (that spanned the entire width of the image) in which a target was likely to be located (Torralba et al., 2006). Critically, there were only three types of targets used in their experiment: people, mugs, and paintings. Our results demonstrate that background category information alone provides very little guidance in observers' fixations along the vertical dimension alone when a wider variety of targets are explored. This could suggest that multiple object configurations contribute to global scene statistics, and thus that object information is a part of what has historically been referred to as "scene gist" or "scene context". In combination, these two results emphasize that the influences of various types of context likely interact with the specific target being searched for. For example, you would expect background information to be much more useful in helping an observer localize an airplane, which will typically be found in easily identifiable sky regions, than a pencil, which will be easier to localize relative to other objects with which it frequently co-occurs.

VI: Extraction of Cues in the Periphery

Our final consideration of the three spatial contextual cues discussed thus far is how each cue can be extracted across the visual field. Given that the category of a scene has been shown to be extractable prior to the initiation of an eye movement (Larson et al., 2014), we expect that the background category cue will be most robust to extraction in the far visual periphery. The other cues will likely be more susceptible to degraded extractability in the periphery due to our diminishing resolution in retinal locations distant from the fovea.

Assessing the extraction of the contextual cues in the visual periphery is important to understand the constraints that a foveated visual system imposes on the use of the individual

contextual cues to guide eye movements. For example, there are two reasons why a cue may not be found to provide visual search guidance: (1) it simply does not provide information that is useful in facilitating visual search performance or (2) the cue could provide useful information, but is not easily extractable in the periphery, and is therefore never utilized as an information providing source.

We assessed each cue's extractability across the visual field by displaying images with or without each contextual cue present at various eccentricities from a forced fixation location. The observer's task was to indicate whether a particular cue was present in the image. To assess the possible interaction of multiple cues, we also manipulated whether observers detected the presence of the cue of interest within an image containing no other cues or all other cues. We measured observers' cue detection performance as a function of eccentricity and whether the image contained no or all other cues.

Methods

Participants. Undergraduate and graduate students (n = 360) at the University of California, Santa Barbara participated in the experiment in exchange for course credit or cash payment. All participants provided informed written consent and were verified to have normal or corrected-to-normal vision.

Design. We manipulated the type of contextual cue that participants were instructed to detect (three levels: O, M, or B, between-subjects), the eccentricity of the center of the image from the observers' fixation point (five levels: 0°, 4°, 8°, 12°, or 16°, within-subjects), and whether all or no other types of contextual cues were present in the stimuli (two levels, between-subjects), resulting in a three-way mixed design. Sixty participants each were randomly assigned to the six between-subjects experimental conditions. The eccentricity of the scene was again Latin-squares

counterbalanced across levels using groups of nine images. Image presentation order was randomized.

Stimuli. The same base-set of scenes used in past chapters were used to create the stimuli for this experiment. Different stimulus sets were used for each of the six (3 cue types x 2 other cue presence) between-subjects conditions. The most prominent stimulus differences arise depending on whether all or no other cues were present besides the cue of interest (i.e., the cue that the participant was to detect as part of their task). In the condition where no other cues were present besides the cue of interest, the cue-present version of the image contained the cue of interest alone. More precisely, O-present images were taken from scenes where the background was mismatched and all objects except the co-occurring object were jumbled, M-present images contained scenes with mismatched background and no co-occurring object, and B-present images were scenes with jumbled objects and no co-occurring object. The cue-absent images were taken from the scene with no contextual cues present. Alternatively, when all other cues were present, the O-present, M-present, and B-present images were identical, taken from scenes containing all three cues. For the O-absent image we deleted the co-occurring object, for the M-absent image, we jumbled the objects, and for the B-absent image, we modified the background of the image to correspond to the mismatched (either the indoor or outdoor) category. The images were circularly cropped to a 700 pixel (11.9°) diameter and the targets were never present in the images. Each participant viewed a set of 45 images total. An example of each image type is shown below in Figure 21.

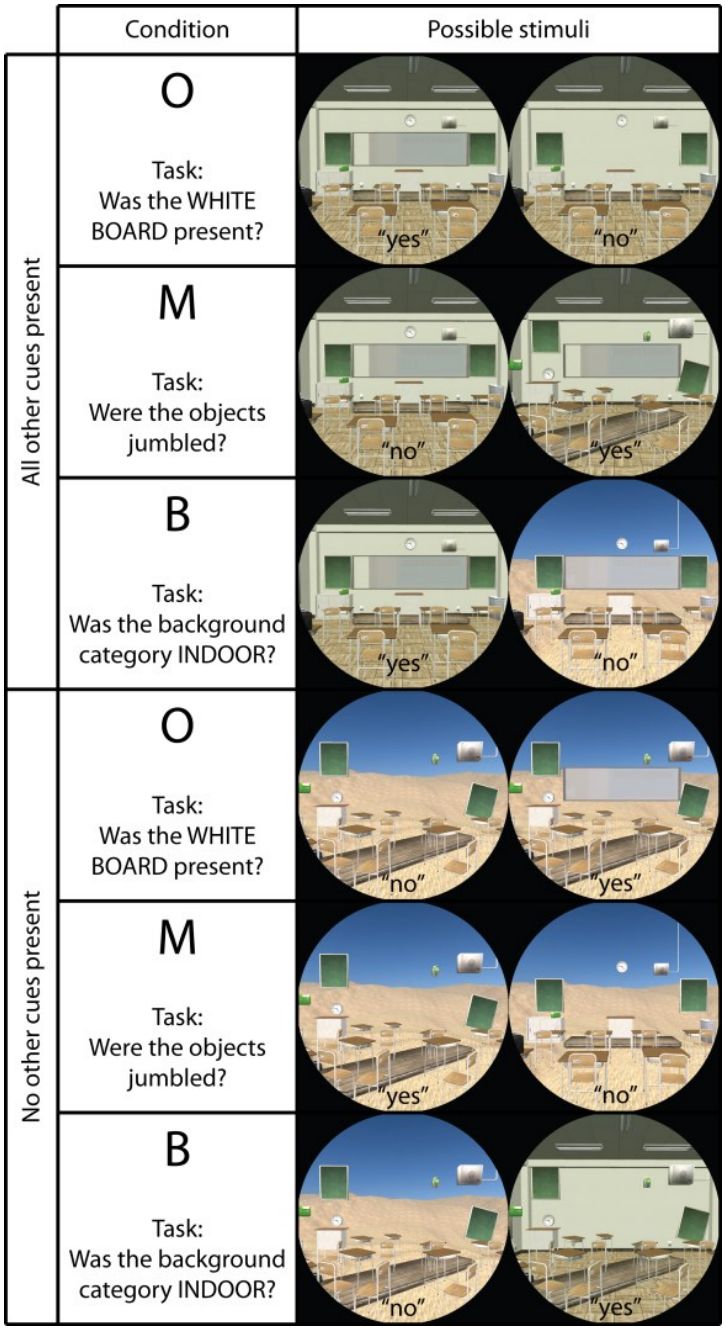

Figure 21: Task instructions and sample stimuli for Experiment III. The first two columns indicate the condition corresponding to the stimuli in the rightward columns and the specific task that participants performed in that condition. Overlaid on the possible stimuli are the correct responses to the task question. As indicated by the tasks, only one of the two images for each condition appeared on screen, chosen randomly with equal probability. Note the difference between stimuli for when all other cues are present alongside the cue that defines the observers' condition versus when no other cues are present alongside the cue relevant to the condition.

Apparatus. Stimuli were displayed on a 3440 x 1440 pixel resolution LG 34UM95 LED monitor. Participants used a chin and forehead rest to stabilize their heads 77 cm from the monitor,

resulting in a single pixel subtending 0.017° of visual angle. The same eye-tracking equipment, settings, and procedures as Experiment 1 were again used to ensure that participants did not initiate any eye movements during the experimental trials.

Procedure. Participants were instructed that they would be viewing a series of images and determining certain properties about those images without moving their eyes. Every task required observers to make a yes/no judgment about whether a particular cue was present in the images, and they were informed that there was a 50% likelihood that it would be present. Observers in the O condition had to determine with the co-occurring object was present in the image. Observers in the M condition determined whether the objects in the image were jumbled. Observers in the B condition determined whether the background of the images matched a category cue.

Participants' initiated a trial by fixating a cross and pressing the space bar. To manipulate the distance of the fixation from the image, the initial fixation cross appeared in one of five locations. If participants were assigned to the O or B condition, after pressing the space bar, the cross would be replaced with the name of the object (e.g. BENCH) they were to search for of the category cue (e.g. INDOOR) they were to determine if the image matched, respectively. Participants in the M condition simply maintained fixation on the cross. After 500-1500 ms, if the initial fixation location (and thus the cross indicator on the computer monitor) was not within the boundaries of where the subsequent image would appear, the cross would re-appear for 200 ms before the image appeared. If the fixation cross was located within the subsequent image boundaries, the fixation cross and cue would disappear 200 ms prior to the image appearance to eliminate masking effects. The image was displayed for 500 ms, after which a response screen appeared identical to that of the previous experiment where participants indicated their confidence as to the presence or absence of the contextual cue.

Statistical Analyses. We analyzed how performance at the yes/no task changed as a function of image eccentricity between contextual cue conditions by measuring the proportion of trials in which participants' correctly detected the presence of the contextual cue (PC). We conducted a three-way mixed ANOVA and used false discovery rate controlled post-hoc comparisons to assess pairwise differences and interpret significant interactions.

Results

Index of detectability as a function of image center eccentricity. To test for differences between levels of eccentricity, type of contextual information, and the manipulation of other cues being present alongside the cue of interest, we assessed the index of detectability and the proportion of trials in which the participant correctly detected the target. Figure 22 depicts the index of detectability. Similar to Chapter V, some observers had perfect hit or false alarm rates not allowing for the calculation of the individual index of detectability for an ANOVA. We chose not to statistically assess the data using bootstrapped distributions of global d' scores because we were specifically interested in evaluating the interaction between cue type and eccentricity. Thus we have focused our analyses on the PC results (Figure 23) where we could statistically assess the interaction. We performed a three-way mixed ANOVA with eccentricity as a within subjects factor and context type and other context presence as between subject factors. All three main effects were significant (eccentricity, $F(4,1416) = 49.47$, $p < 0.001$; context type, $F(2,354) = 200.63$, $p< 0.001$; other context presence, $F(1,354) = 6.22$, $p = 0.013$). There was also a significant interaction between context type and other context presence, $F(2,354) = 5.29$, $p = 0.005$, as well as between eccentricity and context type, $F(8,1416) = 2.76$, $p = 0.005$. All other interactions were not significant.

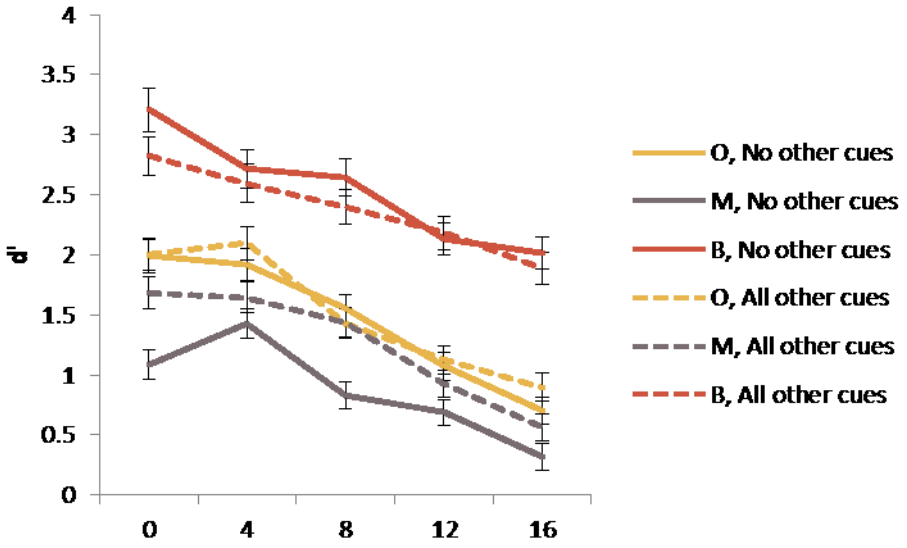

Figure 22: D**etectability** index as a function of image eccentricity from fixation for each of the contextual cue conditions. Error bars represent an estimate of the standard error of the mean, as calculated from the **detectability** indexes delineating the inner 68.29% of the distribution of detectability indexes from 10,000 bootstrap re-sampled samples.

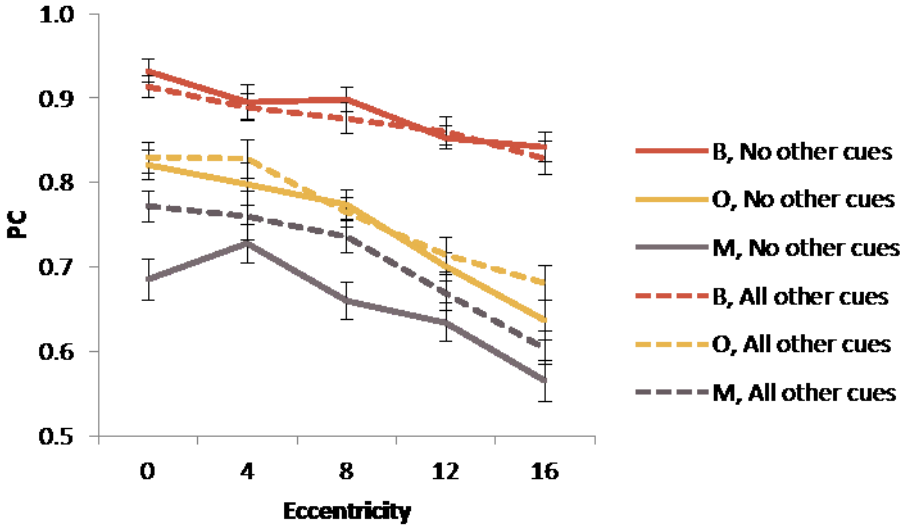

Figure 23: Average proportion of trials where the participants correctly determined the presence/absence of the target as a function of image eccentricity from fixation for each contextual cue condition. Error bars represent the standard error of the mean.

We performed post-hoc comparisons while controlling the false discovery rate to interpret the main effects in light of the significant interactions. First, we wanted to understand whether performance at detecting each contextual cue changed as a function of eccentricity for each type of contextual cue, shown in Figure 24. For all three cues, detection performance was significantly better for the nearest eccentricities than for the farthest (B, mean difference = 0.09, $p < 0.001$; M, mean difference = 0.14, $p < 0.001$; O, mean difference = 0.17, $p < 0.001$). The slope of the drop-off for background category was shallower than that of multiple object configuration and object co-occurrence information. Second, we wanted to better understand the effect of the presence of other contextual cues on observers' performance, shown in Figure 25. When no other contextual cues were present, participants performed best at determining the presence of the background category cue, second best at detecting the co-occurring object, and third best at detecting multiple object configuration information (B vs O, mean difference = 0.14, $p < 0.001$; B vs M, mean difference = 0.23, $p < 0.001$; O vs M, mean difference = 0.09, $p < 0.001$). The pattern of results is identical when

all other contextual cues were present as well (B vs O, mean difference = 0.11, $p < 0.001$; B vs M, mean difference = 0.17, $p < 0.001$; O vs M, mean difference = 0.06, $p < 0.001$). However, critically, the only contextual cue that was significantly affected by the presence or absence of other cues was multiple object configuration information (all vs no other cues present, mean difference = 0.06, $p < 0.001$).

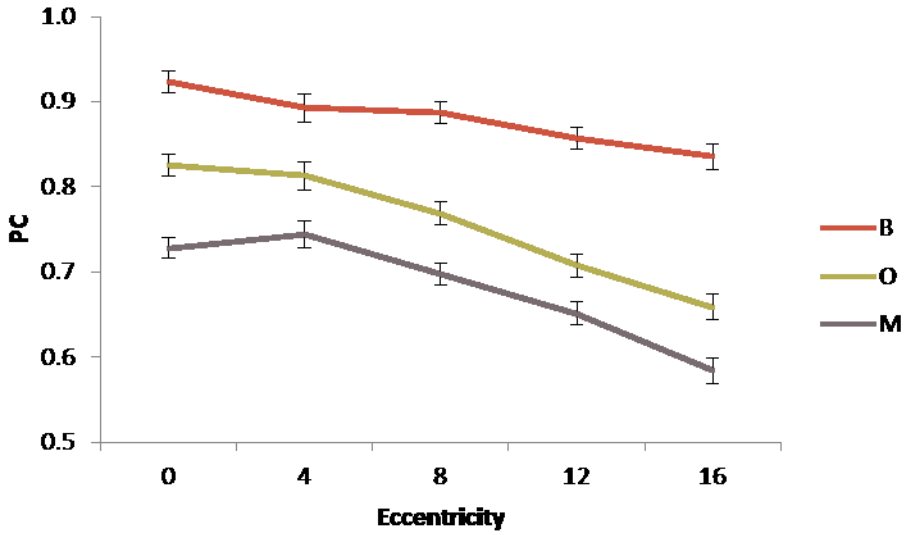

Figure 24: Average proportion of correct judgments about target presence as a function of image eccentricity from fixation for each contextual cue condition, irrespective of the presence of other cue information, i.e., an illustration of the interaction between eccentricity and contextual cue type.

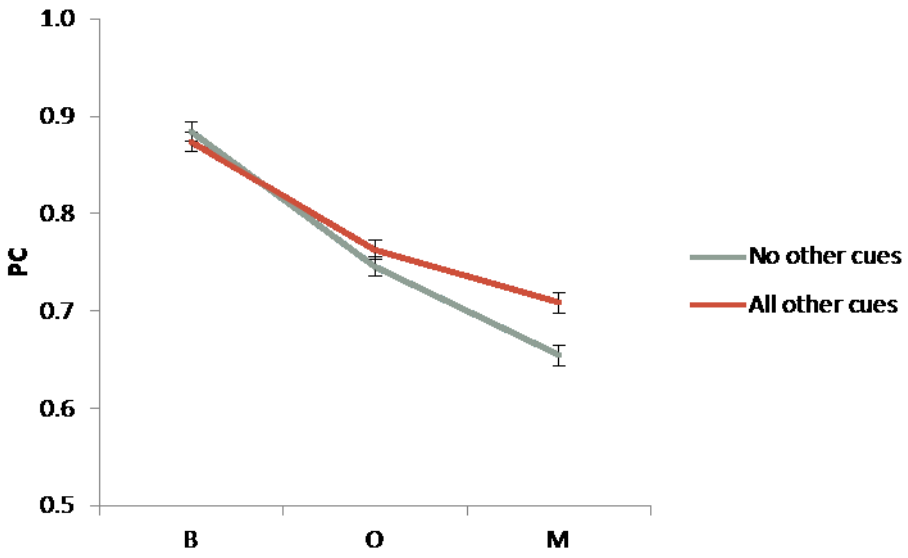

Figure 25: Average proportion of correct judgments about target presence as a function of contextual cue type depending on the presence of other cue information, irrespective of image eccentricity from fixation, i.e., an illustration of the interaction between the manipulated cue type and the presence of other information.

Target detectability as a function of co-occurring object eccentricity and area. For the object co-occurrence cue, recall that observers' task was to determine whether the co-occurring object was present in an image at various distances from fixation. We performed a more fine-grained analysis of performance at detecting the co-occurring object as a function of eccentricity and discriminability in the periphery given that the cue itself is localized to a single object of known size within the image. We calculated the distance of the center of the object (specifically, the center of a rectangular bounding box that contained the object and overlapped with its widest and tallest points) from the fixation location on a given trial. We also calculated the area (in square-degrees of visual angle) of the object by using LabelMe (Russell, Torralba, Murphy, & Freeman, 2008) to draw a polygon around the object and then calculating the area of that polygon given the vertices recorded by LabelMe. Figure 26 shows a scatterplot of the relationship between performance (PC) at detecting the co-occurring object cue and co-occurring object area. Performance and co-occurring object area were significantly positively correlated $r(43) = 0.40$, $p < 0.01$, suggesting as expected that

the co-occurring object is easier to detect when larger. Note that removing the two outlier cases does not affect the strength of the correlation, $r(41) = 0.43$, $p < 0.01$. Figure 27 shows a scatterplot of the relationship between performance (PC) at detecting the co-occurring object cue and co-occurring object eccentricity from fixation. As expected, performance at detecting the co-occurring object was significantly negatively correlated with the distance of the object from fixation, $r(223) = -0.319$, $p < 0.001$.

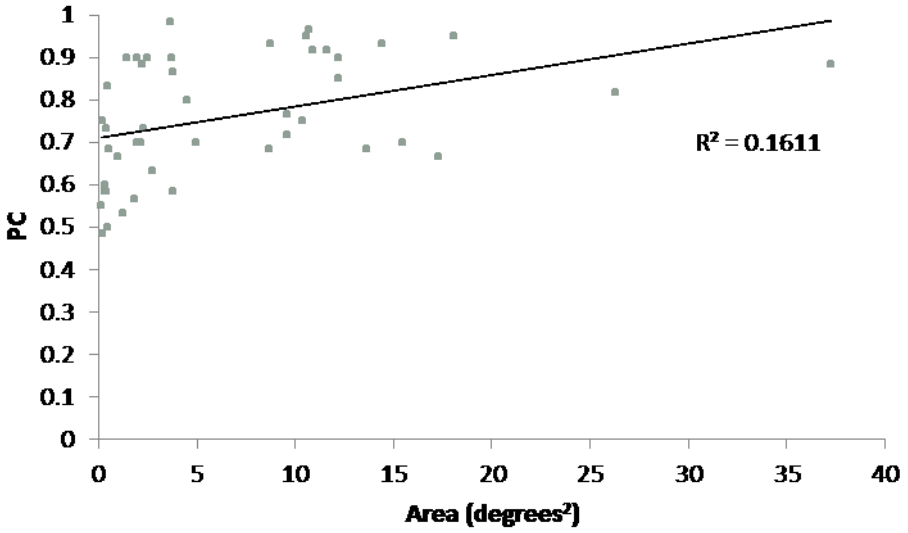

Figure 26: A scatterplot demonstrating the positive correlation between performance at detecting the co-occurring object and its size.

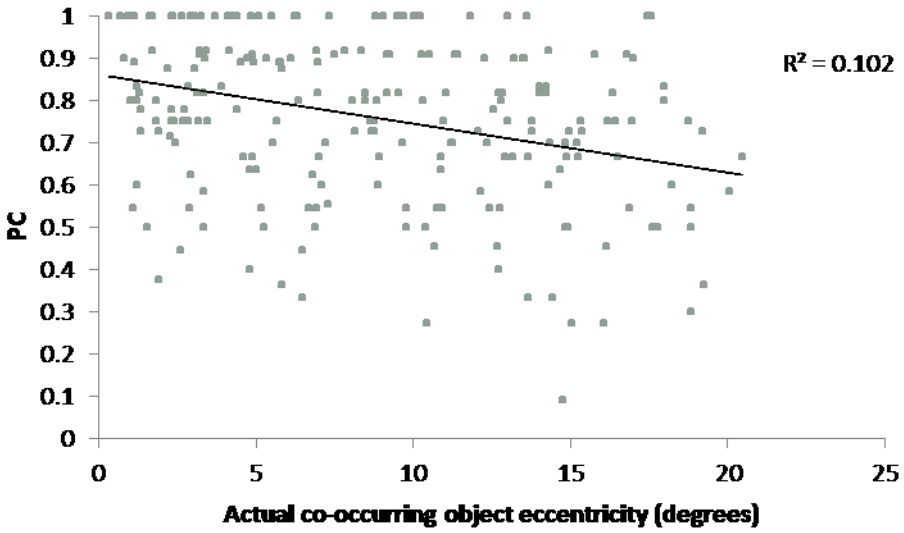

Figure 27: A scatterplot demonstrating the negative correlation between performance at detecting the co-occurring object and its distance from fixation. Note that because there were 45 images displayed at five different retinal eccentricities, there were 45x5=225 different co-occurring object retinal eccentricities.

Note that in the relation between PC and object area (Figure 26), each point in the scatterplot corresponds to PC averaged across all trials for a given co-occurring object (i.e., image), thus the averages are collapsed across all eccentricities. The relation between PC and object area may be better explained when controlling for eccentricity. Therefore, to better understand the interplay between retinal eccentricity of the cue and its size, we chose to perform a linear regression with co-occurring object retinal eccentricity and area as predictor variables for performance at detecting the co-occurring object. Retinal eccentricity and area were significant predictors of detection performance, $R^2 = 0.20$, $F(2,222) = 27.15$, $p < 0.001$. The zero-order correlations, partial correlations, and coefficients are summarized in Table 4 below. Note that that zero-order correlation between area and PC is different here than in Figure 26 because each area was used five times to be paired with retinal eccentricity given that the same object was shown at five different eccentricities (resulting in 45x5 = 225 PC values). The correlation between PC and area is not strengthened when controlling for eccentricity ($pr^2 = 0.10$), therefore another image-based factor

(such as saliency of the object) may be accounting for the additional variability in detection performance.

	Zero-order		β	pr	*b*
	Eccentricity	PC			
Area	0.06	0.29***	0.31	0.32	0.01***
Eccentricity		-0.32***	-0.34	-0.35	-0.01***
	R^2:	0.20***		Intercept:	0.81

Table 4: Results of linear regression analysis using co-occurring object retinal eccentricity and area as predictors of observers' performance at detecting the co-occurring object. *** indicates $p < 0.001$.

Discussion

Background-based information was much easier to extract in the periphery compared to object-based information. This is consistent with past research that has shown that background information in the form of scene gist can be extracted rapidly during scene viewing. However, this result is important because, contrary to the intuition that a cue which may be easily extractable in the periphery would be quicker and more impactful in its influence on behavior, a growing body of work is consistently demonstrating that object-based information that is relatively less extractable in the periphery provides much more information for visual search tasks. Recall that we specified two possibilities as to why a cue may not influence visual search: (1) because it simply does not provide information that is useful in facilitating visual search performance or (2) because the cue could provide useful information, but is not easily extractable in the periphery, and is therefore never utilized as an information providing source. Previous chapters have illustrated that background information facilitates visual search to a much lesser degree than object information. The results here indicate that background information fails to facilitate visual search because it is not informative of the target search task rather than because the cue is unable to be extracted in order to be utilized.

The results do indicate a possible useful role for background information. Multiple object configuration information was easier to detect when paired with background category and co-

occurring object information. Given that the presence of a single additional object (the co-occurring object) among other jumbled objects likely does very little in helping observers determine whether the remaining objects are jumbled, it is likely that background information is contributing more to improved detection of the multiple object configuration cue. This suggests that background information facilitates the extraction of other cues. Although background information in isolation does not provide as much localization or perceptual performance benefit as object-based information, it certainly facilitates observers' ability to interpret the spatial arrangement of objects and presumably then utilize object information for eye movement guidance.

Surprisingly, although the size and distance from fixation of an object predicted an observers' ability to detect it, the variance in performance accounted for by size and distance was lower than one might expect. It could be that other factors, such as the saliency of the object, are better predictors of its detectability in the periphery. Existing work has shown that different objects show unique eccentricity biases, such that object-selective areas overlap visual field location representations according roughly to the type of vision needed to make distinctions about such objects (e.g., face-selective regions overlap central-vision representations and building-selective regions overlap peripheral representations; (Levy, Hasson, Avidan, Hendler, & Malach, 2001). This effect cannot be explained by low-level image properties, and likely reflects preferences of high-order cognition (Yoo & Chong, 2012). It might, therefore be the case that specific objects have different "sweet spots" on the retina for which they are most detectable, resulting in a more complex relationship between ability to detect an object and its distance from fixation. Similarly, the extensive literature on visual crowding provides evidence that the relationship between object size and detectability is equally complex given that critical spacing of objects to make them identifiable in the periphery is independent of object type and size (Pelli & Tillman, 2008). Some of these factors

might be in play and account for the low correlations obtained relating object detectability to target area and to target eccentricity.

VII: Non-spatial contextual cueing of visual search

To this point, all of the contextual cues that we have considered are extracted using top-down knowledge of a given scene and have provided spatial information about the target and have facilitated search. However, not all contextual cues are spatial in nature. Top-down cues can also be non-spatial and influence neural priority maps (Serences & Yantis, 2006). Again, these task-relevant, goal-driven cues can allow the observer to search utilizing target-relevant features while ignoring irrelevant features. Researchers have summarized a long list of feature dimensions that have been studied in the visual search literature and organized them into those that have been undoubtedly been shown to guide attention (Wolfe, 2014; Wolfe & Horowitz, 2004); namely: color, motion, orientation, and size. Of note, is the size feature. How a scene might provide information about the likely size of objects has been rather unexplored. Existing work has investigated the "size" feature dimension as it pertains to guiding attention to artificial stimuli of varying sizes (Stuart, Bossomaier, & Johnson, 1993; L. G. Williams, 1966), lengths (A. Treisman & Gormican, 1988), or spatial frequencies (Moraglia, 1989; Sagi, 1988; Verghese & Nakayama, 1994). Crucially, past work has often explicitly defined the target of a search task in terms of its size and has minimally investigated the importance of real object size in the context of natural scenes. Regarding the latter, Biederman et al (1982) investigated size as one of their five relational violations (interposition, support, probability, position, and size) that detrimentally affect object recognition in line drawings of common scenes. Sherman and colleagues (2011) have shown that observers can eliminate distractor locations when given both depth and size information about possible target locations in real scenes.

Here, we directly explore how the human brain utilizes the scene information to guide the search towards likely sizes of a target object and facilitates search.

Therefore, we seek to answer two questions: Does a scene provide contextual information to guide search towards likely target object sizes? If so, is it possible that objects undergoing size violations as determined by scene context might be completely ignored by observers searching for a specific target? Our expectation is that contextual information effectively constrains object search to proponent objects of a particular expected scale/size. Although humans may be prone to ignore objects that violate size expectations, under typical circumstances, scale of objects is a useful cue to correctly recognize and classify objects. We demonstrate this by showing that a state-of-the-art object classification algorithm is not prone to missing mis-scaled objects, but in turn *is* prone to falsely detecting object categories of an incompatible scale with candidate object regions.

In this chapter, we manipulate the scale of the target object relative to other objects in a scene image and assess the impact on target detectability and eye movement guidance. Observers are shown images with targets of normal size and of drastically increased size relative to the other objects in the scene, as well as control scenes where the target object size is equated to that in the mis-scaled scene, but all other objects have been proportionally scaled-up as well. We compare target detection performance between the appropriately and inappropriately scaled objects and confirm that differences in detectability are not due to feature changes of the object upon being re-sized. Although the scale of a target is a manipulation of the "spatial" content of a scene (we are adjusting the spatial size of the target), we refer to this cue as non-spatial because the information it provides does not delineate a spatial location in which observers might search for a target, i.e., scale is relevant as a feature of an object rather than to the location of an object.

Methods

Participants

Eye-tracking and target detection response data were collected from 60 undergraduate students at the University of California, Santa Barbara who received course credit in exchange for participation. All participants were verified to have normal or corrected-to-normal vision. A second group of 106 observers performed an object labeling task that served as a control experiment. All participants were recruited and treated according to approved human subject research protocols, and provided informed written consent.

Stimuli & Design

The stimuli used were a subset of those used in past chapters, slightly modified to contain target objects that were mis-scaled relative to their surroundings. A total of 42 scenes were created in Unity 3D (Unity Technologies, Bellevue, WA, USA), each with a unique target object that would be searched for by participants in the experimental task. There were a total of 14 target objects, each repeated three times, but never identically (i.e., color, viewing angle, etc. changed across the three instances). From each scene, five images were created: (1) normal scene with target scaled proportionally to surroundings, (2) scene with target scaled 2-4x larger than normal scene, (3) zoomed-in image of scene where target is identical in size to (2), but all other objects are proportionally larger as well, (4) target absent version of (1), and (5) target absent version of (3). Examples of the five images from a sample scene are shown in Figure 28. The mis-scaled target objects were always constructed to be larger than their normal controls so that difficulty in detection cannot be attributed to the object becoming smaller and less detectable. Condition 3 serves as an additional control to ensure that the scaled-up version of the target is recognizable as the target object in normal viewing conditions.

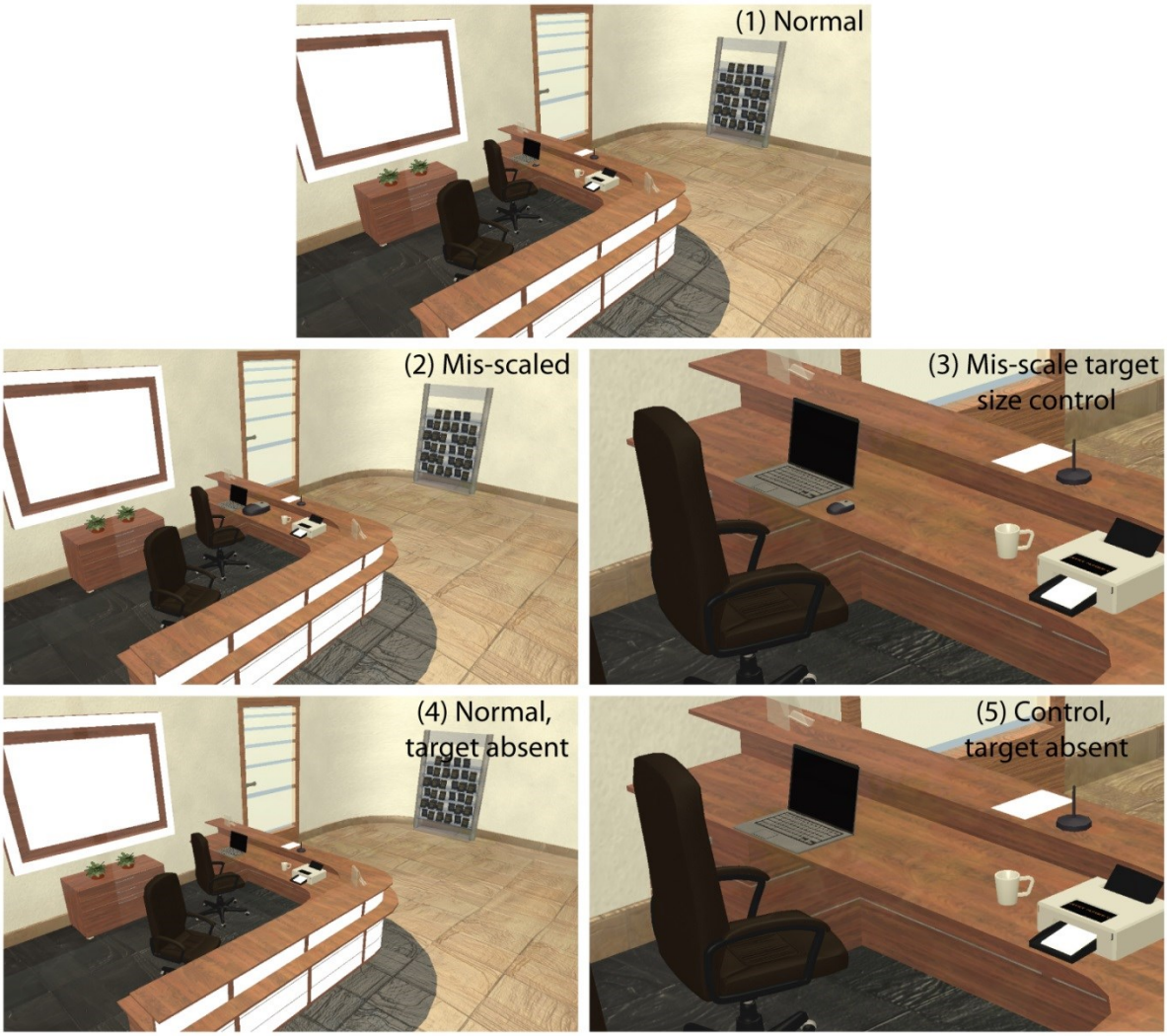

Figure 28: Sample stimuli from the target search task. The target is a computer mouse, sitting to the left of the laptop computer. (1) shows a normal sized target, (2) the target at 4x its expected size, (3) a target of the same size as (2), but with the scene context proportionally scaled up as well. (4) and (5) show the target absent versions of the first three images.

Images were divided into the five conditions using a Latin square design, resulting in 7 images per condition, with the exception of the normal target absent scenes (4, above), of which there were two groups of seven (fourteen total). This ensured there was an equal number of target present and target absent scenes. Participants were randomly assigned to view a particular stimulus set shown in randomized order. We assessed the hit and correct rejection rates of observers during the target detection task and the distribution of fixation eye movements relative to the target locations.

Apparatus

Stimuli were displayed on a 1280 x 1024 pixel resolution Barco MDRC-1119 monitor. Each pixel subtended 0.022 degrees of visual angle. Eye tracking data were recorded on an Eyelink 1000 (SR Research Ltd., Mississauga, Ontario, Canada) monitoring gaze position at 250 Hz and was calibrated and validated using a nine-point grid system. A velocity greater than 22°/s and acceleration greater than 4000°/s^2 classified an event as a saccade.

Procedure

Human Behavioral Task. Participants were instructed that they would be viewing a series of images and determining whether or not a particular object was present within them. They were told there was a 50% likelihood that the target would be present, but were not given any indication that some target objects may be sized abnormally. Each trial began with a fixation cross in the bottom-center of the screen, below the edge of where the image would appear. Participants fixated the cross and pressed a button to initiate the trial. The image appeared after the participant maintained fixation on the cross for a random interval between 0.5-1.5 seconds. The participants had 1000 ms to search for the target object while their eyes were tracked before a response screen appeared where they indicated on a ten-point scale whether the target was present and how confident they were in their response.

Object Recognition Model. We ran a Python implementation (Ren, He, Girshick, & Sun, 2015) of a deep residual learning framework (Res-Net; He, Zhang, Ren, & Sun, 2015), the current state of the art convolutional neural network (CNN) object detector as a comparison to human performance. The implementation is pre-trained on the 80 categories of objects in images from the Microsoft Common Object in Context (MSCOCO; Lin et al., 2014) database. MSCOCO is an image database chosen to contain cluttered scenes with detailed backgrounds, as opposed to more typical databases that contain images of a single object against a mostly uniform background. The model

initially proposes candidate image regions thought to contain an object and computes a probability for each object category that the region contains an object of that category. Typically, the category with the highest probability is nominated as the object within the region and non-maxima suppression is applied so that object proposal regions with significant overlap do not produce redundant object labels.

Three of the MSCOCO object categories were target objects in the visual search stimuli: toothbrush, parking meter, and computer mouse. To directly compare human and model behavior, we assessed any object proposal region that overlapped with the target objects in the images used for the visual search task. We then selected the region with the highest probability associated with the target category (in all cases, the region contained at least half of the target object) and compared the average model detection probability with observer detection hit rates. Given that our images were computer rendered, we also ran a small set of images of real scenes and assessed the detections of object categories not contained in the visual search stimuli.

Results

Target detection performance. Figure 29 shows the hit rate (for target present trials) or correct rejection rate (for target absent trials) of observers in each of the experimental conditions. There was a 0.13 drop in hit rate at detecting the mis-scaled object relative to the normally sized target object, $t(59) = -3.94$, $p<0.001$. This difference cannot be attributed to featural changes to the target object because an identical target object with contextual objects scaled proportionally to it was detected near perfectly (see the control conditions). Given that the objects in the scenes were 3D renderings of real objects, a separate group of participants completed a control task to ensure that they were able to properly identify the simulated target object as intended in the complete absence of contextual information. A total of 107 observers completed a task where they were shown an image of the mis-scaled target object in isolation on a grey background and were asked to

name the object shown. The observers were split into three separate groups (77 observers in group 1, 18 in group 2, and 12 in group 3) so that each group would only see a single instance of each target object (recall that three versions of each target were used for a total of 42 scenes), thus each group viewed a total of 14 objects. Twelve of the 42 objects were freely identified by fewer than 80% of the observers, therefore we ran a separate analysis with those objects excluded to assess whether confusion over object identity could be driving the results. With low confidence objects removed, hit rates increased by 3% overall for the normal and mis-scaled target present trials, but the difference between them remained the same ($M = -0.13$, $t(59) = -3.95$, $p<0.001$).

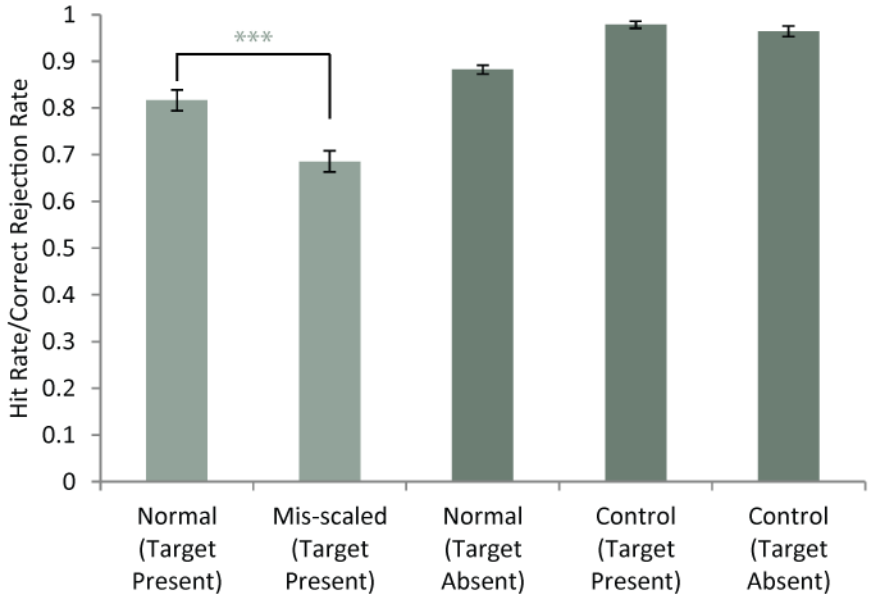

Figure 29: Hit rate (target present trials) and correct rejection rate (target absent trials) for each experimental condition in the target detection task.

Influence on eye movement guidance. Our perceptual performance results suggest that observers are likely to fail to detect an object of unexpected size. A possible explanation is that the finding is related to observers' failing to fixate the target region. Figure 30 depicts the distance of the closest fixation to the target location across all experimental conditions. On average, observers fixate 0.42° closer to the center of the target object when it is of a normal scale than when it is mis-

scaled, t(59) = -3.94, p < 0.001. Interestingly, there is a larger magnitude difference (0.82°) between the mis-scaled target present trials and target absent trials, therefore there appears to be some evidence that observers are not behaving on mis-scaled trials as though the object is entirely absent. To better understand these differences, we assessed the distance of observers' closest fixations to the target location on trials where they correctly or incorrectly detected the target object, shown in Figure 30.

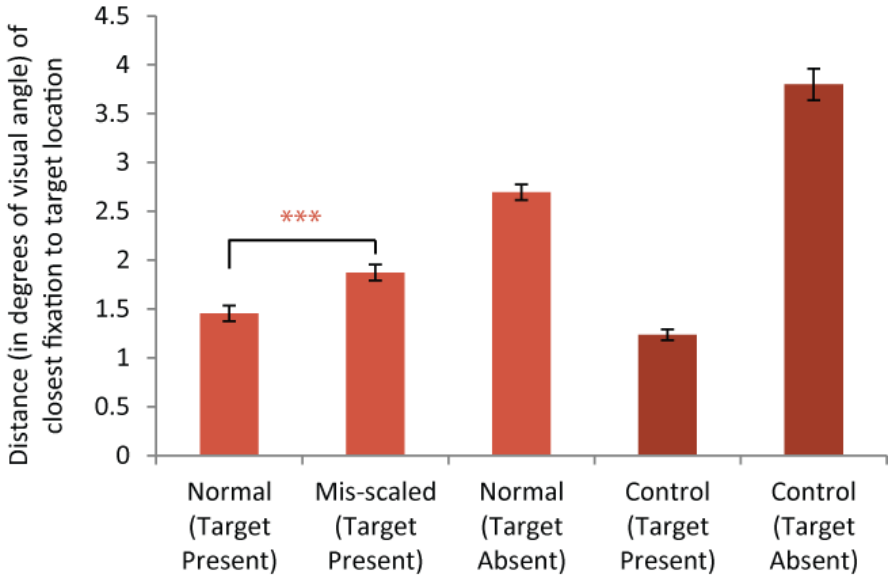

Figure 30: Average distance of the closest observer fixation to the target location on each trial for each experimental condition.

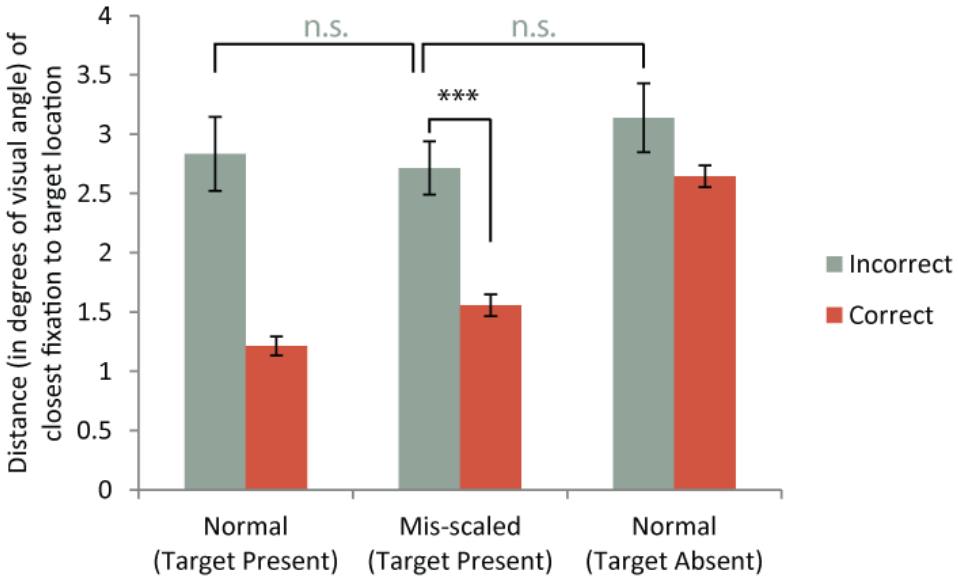

Figure 31: Average distance of the closest observer fixation to the target location on each trial for the non-control experimental conditions, divided between trials where the observer correctly or incorrectly detected the presence of the object.

We computed four t-tests to assess the difference between the correct and incorrect mis-scaled trials, the normal and mis-scaled conditions, and the incorrect mis-scaled trials and incorrect target absent trials. First, looking solely at mis-scaled trials, observers fixate significantly closer to the target region on mis-scaled trials where they correctly detected the target than on trials where they missed the target ($t(53) = 4.77$, $p < 0.001$). Next, comparing fixations between the normal and mis-scaled targets, observers' proximity of fixations to the target location are significantly closer on trials where they correctly detected the target when the target is normally scaled than when it is mis-scaled, but this difference is small (0.34°; Normal v Mis-scale, correct: $t(59) = -3.23$, $p = 0.002$). More importantly, the proximity of fixations to the target location on trials when observers missed the target was not significantly different (0.12° closer when mis-scaled; Normal v Mis-scale incorrect: $t(35) = 1.48$, $p = 0.15$). Therefore, when observers are failing to detect the mis-scaled target, there is no indication that it is due to differences in their eye movement guidance. In mis-scaled trials where observers fail to detect the target object, they behave similarly to when the target is absent from the scene ($t(39) = -1.05$, $p = .30$). Overall, this suggests that many times when

observers fail to detect mis-scaled objects, it is because they are failing to fixate the target region, rather than fixating the target region but failing to perceive the object once there. However eye movement guidance does not appear to be affected greatly by whether a target is mis-scaled relative to its surroundings. So what accounts for the drop in performance between the normal and mis-scaled conditions?

In Figure 32, we have plotted the proportion of times that observers missed the target when they fixated within two degrees of the target region. Observers were more likely to miss the target upon fixating it when it was mis-scaled than when it was normally scaled, $t(59) = 3.49$, $p < 0.001$, a total difference of 0.12, which is similar to the overall difference in hit rate across all trials between normal and mis-scaled trials (0.13). Observers fixated the target region on 335 out of 420 trials when it was normally scaled and on 299 trials when it was mis-scaled. Of those times, they failed to detect the target 38 times and 71 times, respectively.

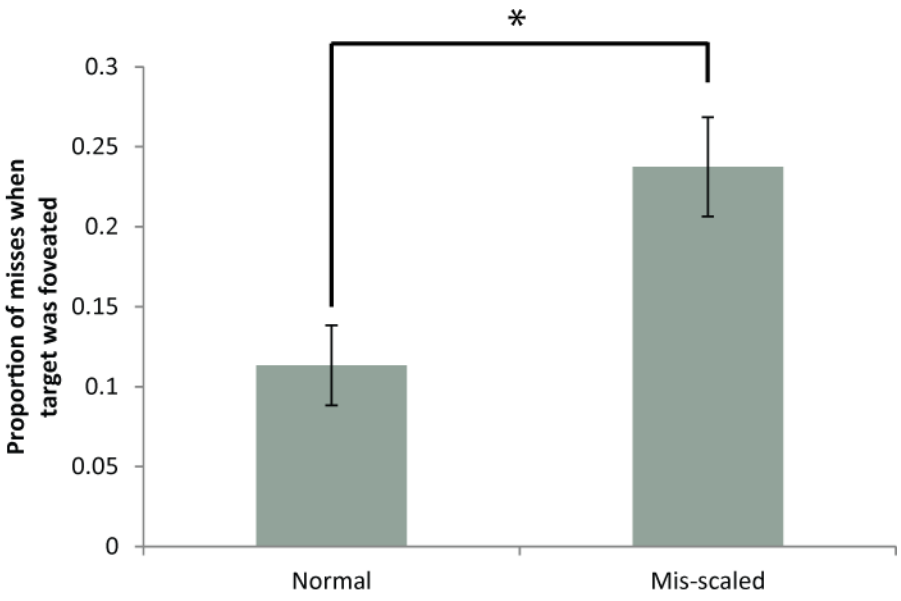

Figure 32: The proportion of times that observers missed the target on trials when they foveated the target region.

One aspect of the experiment that might influence the effect of target size inconsistency is that as the experiment progresses the observers might learn that targets occasionally appear at

sizes that are inconsistent with the scene. This learning process might lead observers to change strategies and guide their search across a variety of spatial scales (consistent and inconsistent with the scene) and diminish the influence of the mis-scaled target on search performance. To evaluate this possible influence of learning on our results we separated our performance measures across experimental blocks. Figure 33 depicts the average hit rate across seven-trial blocks for both the normal and mis-scaled targets. Not until the final seven trials of the experiment (the sixth block) did the hit rate for mis-scaled trials show a clear indication of increasing to the level of performance with the normal images. We compared the average hit rate on the first five blocks to the final block hit rate on mis-scaled trials using an independent samples t-test with unequal variance and found that observers' hit rates increased significantly by 0.12, $t(50) = 1.99$, $p = 0.03$.

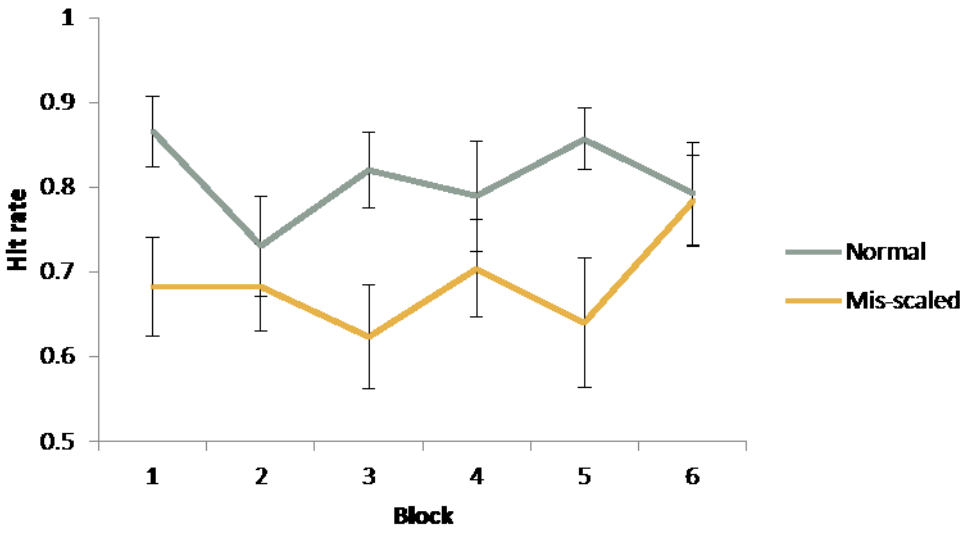

Figure 33: Average hit rate across blocks of seven trials (in chronological order) for the normal and mis-scaled conditions.

Comparison to object detection model. We have shown so far that humans are susceptible to failing to detect mis-scaled targets. In most cases, reliance on scale information is a useful heuristic in guiding visual search and can likely constrain the possibility of confusing larger objects for smaller objects and vice versa. Computer vision models of object detection do not directly rely on scale information to inform object classification. In Figure 34 we show the average probability of

target category detection according to a state-of-the art object detector in the images viewed by observers in the visual search task for both normal, mis-scaled, and control target images. For comparison are the hit rates of human observers detecting the objects for the same conditions. Whereas human observers fail to detect mis-scaled objects compared to normal objects (hit rate decreases by 0.35, $t(8) = 4.54$, $p < 0.001$), the object detection model detects normal and mis-scaled objects with the same confidence (probability of detection difference is 0.00, $t(8) = 0.01$, $p = 0.5$).

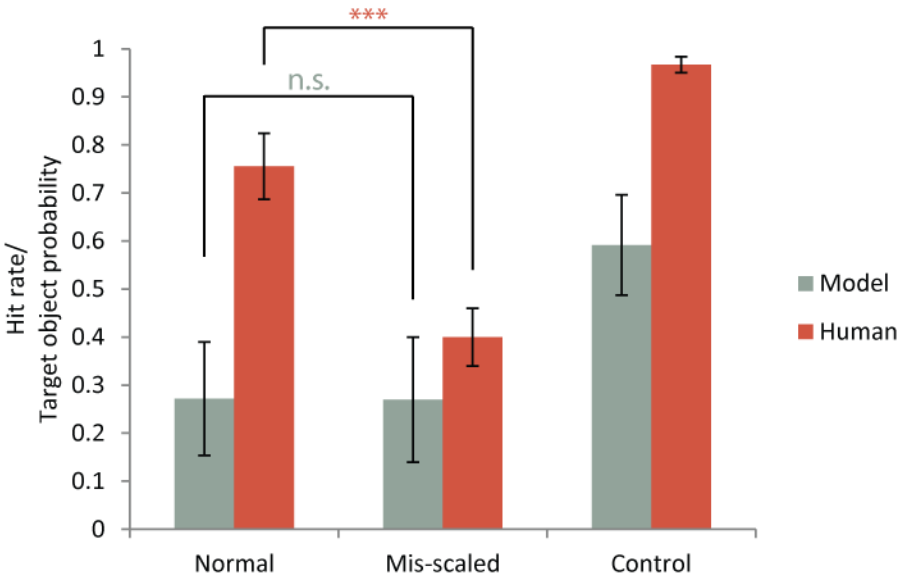

Figure 34: Difference in target detectability between humans and a state-of-the-art object detector. The target object probability is the output of the RCNN and should not directly be compared to hit rate.

The deleterious effects of failing to take scale information into account are demonstrated in Figure 35. The top row illustrates instances where the computer vision model falsely detects with high probability a particular instance of an object category, but the mistaken category is of improper scale relative to the proposed object region. For comparison, the bottom row illustrates instances where the model correctly detects instances of the same category.

Figure 35: Top row – instances where the object detector mis-classifies with high probability an object category of inconsistent scale with the actual object region. Bottom row – correct classifications of the same object categories for comparison.

Discussion

In this chapter, we were interested in assessing the extent to which contextual surroundings in scenes provide information about expected scales of objects and whether these expectations can interfere with performance on a mis-scaled target detection task. We found a strong effect such that observers were worse at detecting objects that were significantly larger (2-4x) than the typical object sizes consistent with their scene context (normal sized objects). It is important to note that these results cannot be explained on the basis of the objects becoming less conspicuous upon being mis-scaled, because they were all increased in size when made inconsistent in size with their scene. We also ensured that the mis-scaled object was detectable in isolation and at the mis-scaled size but consistent in size with its surrounding scene (up-scaling of entire image). Therefore, these results cannot be attributed to differences in low level properties of the target between the normal and mis-scaled conditions.

Interestingly, on trials where participants failed to detect the mis-scaled object, their eye movements were similarly proximal to the target region as in trials when observers failed to detect the normally scaled image and trials where the target was absent. In trials when observers correctly detected the mis-scaled object, their eye movements were similarly proximal to the target region as

in trials when observers correctly detected a normal sized target. Overall, it appears that although target scale has an effect on task performance, it does not have an effect on eye movement guidance associated with target search. The associated drop in performance is likely explained due to observers being more likely to miss the target if it is mis-scaled upon foveating the target region.

Despite instances of detecting mis-scaled targets upon fixation, observers' hit rates on mis-scaled target trials did not increase until the final sixth of trials in the experiment. This suggests that observers did not start monitoring different target spatial scales upon their first encounter with a mis-scaled target. We have also demonstrated that state-of-the-art object detectors fail to take scale information into account and detect with equal probability objects of consistent and inconsistent scale. During classification of cluttered scenes, this can lead models to falsely suggest the presence of objects that would be mis-scaled relative to the scene given the size of the detector's proposed object region.

VIII: General Discussion

The role of scene context in guiding search and its contribution to successful search has been long recognized. Howewer, scene context has been broadly defined and there have been few attempts to clarify how scene context contributes to visual search guidance. Although the presented partitioning of contextual cues is not exhaustive, it is a starting point and is based on the types of cues already frequently discussed in the literature. In our experimental manipulation we changed the background of the scene, thus altering the global features of the image and the semantic category of the scene background. We also manipulated objects within the scene. Other objects that are highly predictive of the location of a searched-for target have been shown to guide search (Mack & Eckstein, 2011). The configuration of multiple objects has been given less attention in the eye movement literature but several studies have investigated its neural basis (Inhoff & Ranganath, 2015).

Beginning with a definition of scene context that stipulates the condition that scene context must provide information that aids in a particular perceptual task, we verified that our chosen manipulations did in fact provide information concerning where a target object might be located in a visual search task. Scenes containing background category or multiple object configuration cues were selected to be more informative than scenes without those cues 96% and 84% of the time. When using expected target location judgments within images containing a single cue to predict target location judgments within images containing all cues, the co-occurring object cue specifically was found to have the strongest explanatory power of the target location expectations of observers viewing scenes containing all cues. We take this as evidence that, indeed, our manipulations are providing contextual information to aid visual search and subsequent target localization and detection.

Table 5 summarizes the results from each chapter to guide the general discussion of the overall findings from this collection of work. The cells in the table are color coded based on the strength of the effect of each cue (columns) on the different metrics (rows) considered in each experiment. Red indicates the strongest effect, yellow the second strongest, and green the weakest effect. Grey cells indicate a tie (based on statistical significance of the results).

	Metric	O	M	B
Target detection performance	d' (Ch. IV)			
	d' (Ch. V)			

	Informativeness of target location			
	Bias (Ch. IV)			
	Bias (Ch. V)			
Eye movement guidance	Distance of closest fixation to target location (target present trials; Ch. IV)			
	Distance of closest fixation to target location (target present trials; Ch. V)			
	Distance of closest fixation to expected target location (target absent trials; Ch. IV)			
	Distance of closest fixation to expected target location (target absent trials; Ch. V)			
	Time to foveate target (Ch. IV)			
Temporal extractability	Timing of peak cue usage (fixation number, Ch. V)	2nd	1st	?
	Increase in cue usage relative to upper bound across fixations (Ch. V)			
Extractability across the visual field	Average detectability of cue across periphery (Ch. VI)			
	Drop in detectability of scene from 0 to 16 degrees in periphery (Ch. VI)			

Table 5: Summary of the results from each chapter to guide the general discussion of the overall findings from this collection of work. The cells in the table are color coded based on the strength of the effect of each cue (columns) on the different metrics (rows) considered in each experiment. Red indicates the strongest effect, yellow the second strongest, and green the weakest effect. Grey cells indicate a tie (based on statistical significance of the results).

The relative influences of spatial contextual cues

A consistent finding across the studies in this work is that object information more than background information facilitates target localization, detection, and eye movement guidance. Object-based cues were more informative of target locations than background-based cues as well. This increase in information about expected target locations by a cue can be related to the guidance of eye movements to target regions during visual search. Most strikingly, object information consistenly facilitates search behavior on many metrics: target detection sensitivity (d'), detection response bias, proximity of eye movements to target region, and time to fixate the target region. More specifically,the co-occurring object is more spatially informative of the target object location,

provides better guidance toward the target location, and reduces conservative detection bias more than the configuration of multiple objects. The reverse trend (though not significant) was observed for target detection sensitivity (d'), which is increased more by multiple object configurations than by object co-ocurrence. This possibly suggests that, although observers are more sensitive at detecting a target when provided with multiple object configuration information than with co-occurring object information, they are also biased toward reporting that the target is present. We believe this is likely the result of observers having to search a larger expected region when presented with multiple object configuration information (relative to the expected region suggested by the co-occurring object) in order to locate the target object, during which time they are biased to report that the target is absent, until they have localized it. Observers who localize a co-ocurring object are then less uncertain about possible target regions and more inclined to make a target present judgment.

What, then, is the use of background category information in visual search tasks? Although its strength of influence on behavior is less than that of object information, we still have evidence suggesting that its influence is significant. In Chapter IV, the combined effect of the background category cue on target detection sensitivity and the guidance of eye movements to the target location on target present trials was also significant. However, the likely key role of background information is to help us make better sense of object information in scenes, as was evident in our results concerning the extraction of multiple object configurations in the periphery. Without background information and co-occurring object information, observers were less able to detect the jumbling of multiple objects. Likely, the background information contributed to helping participants assess the structure of a scene and whether the objects within it adhered to what is typical (e.g., it's difficult to assess whether a painting belongs where it does without a wall as reference). Also, the effect of eliminating bias to respond that a target is absent was reduced most significantly by the co-

occurring object, but adding background category information to that cue greatly strengthened the reduction of bias. These results are consistent with the deluge of past work that has shown that objects that are consistent with their surrounding impacts are more easily recognizable (e.g., Biederman, Mezzanotte, & Rabinowitz, 1982).

Finally, whereas there are clear differences in the strength of utilization of each cue, how is cue information combined when multiple cues are availble? Our findings are consistent with the conclusion that observers are optimally linearly combining the cue information when assessing their target detection sensitivity. This suggests that observers are not prone to suboptimally utilizing multiple sources of contextual information. More interestingly, on the contrary, they also do not appear to make super additive use of contextual information. One could imagine that various cues interact such that the presence of one cue allows a participant to make even better use of another cue. In fact, we did find evidence of this when investigating observers' ability to extract cue information in the periphery. Participants were better able to differentiate jumbled multiple object configurations when the background category and co-occurring object were intact. Thus, we may have expected observers to make suboptimal combined use of cues when multiple object configuration information was available without another supporting cue, or superoptimal combined use of cues when all were available. However, our analysis specifically concerning the combination of cues on target detection sensitivity does not demonstrate an interaction between cues.

The time course of extraction of contextual cues

Our results suggest that target detection sensitivity is not differentially impacted over time by contextual information and that across fixations, observers consistently demonstrate optimal linear combination of various contextual cues. However, eye movement guidance does rely on varying contextual cues as scene viewing unfolds. We found that observers initially utilize multiple

object configurations to obtain coarse localization of possible target regions, and then use object co-occurrence information to make finer localizations to the target location. In the first three fixations of a scene, background category information consistently fails to significantly influence eye movement guidance and target detection performance. We have also shown that, relative to the upper bound of information that each individual cue can provide for localizing the target, the object co-occurrence cue is the only cue to be increasingly utilized as search unfolds. Considering these results in combination with the finding that response bias is significantly more liberal upon the first fixation when object co-occurrence information is present suggests that observers are likely detecting the co-occurring object in the periphery prior to the first fixation (consistent with results from Chapter VI that show the co-occurring object can be detected above chance up to 20° into the periphery), using that information to optimally adjust their bias, but failing to utilize that information for target localization until the second fixation. This is consistent with research that has demonstrated that observers often store information collected before an initial saccade for use during a second, later saccade (Caspi, Beutter, & Eckstein, 2004).

Are cues constrained by their extractability in the periphery?

We have found that observers are able to extract each type of contextual cue above chance from images presented at least up to sixteen degrees into the visual periphery. The background category is most easily extracted in the periphery. Background category information consistently fails to facilitate visual search to the extent that object information does, but not because it is unable to be extracted according to this result. This further reinforces the conclusion that background category information is not informative for visual search guidance, consistent with the lack of information provided by the background about expected target locations in Chapter III.

A surprising result is that observers are more able to detect single objects than they are able to detect whether multiple object configurations have been jumbled in the periphery. Multiple object configuration is the only cue that is differentially detectable depending on whether the two other cues were also present in the image, suggesting that gleaning a perceptual sense of the structure of a scene is highly dependent on multiple cue types being present.

Given that we have found evidence that some objects are detectable above chance even at 20 degrees into the visual periphery, additional credence is lent to the notion that scene background and object information could interact very early on in visual perception. Even in the initial glimpse of a scene, during which we have traditionally assumed only global information can be resolved, humans are capable of resolving information about object identity when attending to them. Even more contrary to our standard conception of contextual-based object information is the result that specific object identities are more detectable in the periphery than spatial configurations of objects overall. It should be noted however that observers' tasks concerning spatial arrangement was to determine if it was jumbled or typical of orderly everyday scenes. We typically do not make such explicit judgments in everyday tasks, and regardless of our ability to distinguish the orderliness of a spatial arrangement, we may still be able to utilize the arrangement we perceive in conjunction with identifying object identities to guide our eye movements.

Cortical scene processing implications

Where in the brain might these spatial context cues be processed? Functional magnetic resonance imaging studies provide some hints. Within the cortex, the parahippocampal place area (PPA) has been identified to encode important components of scenes, specifically layout and general geographical features of space (Epstein, Graham, & Downing, 2003; Epstein & Kanwisher, 1998). In addition, the section anterior to the PPA (Aminoff, Gronau, & Bar, 2007; Bar, 2004) represents

semantic and spatial associations between co-occurring objects (but see Ward, MacEvoy, & Epstein, 2010). The contextual location of a target within a scene is represented across various areas including the lateral occipital complex, the intra-parietal sulcus and the frontal eye fields. The information these areas contain about the expected target location within a scene correlates with the inherent amount of contextual guidance in a scene (Preston et al., 2013). How the information about scene background, object co-occurrence and multiple object configuration is integrated by the brain to generate a prediction about the likely target location is still not known. Our results provide preliminary evidence that the regions supporting the extraction of the spatial layout of a scene and the objects within it likely support the representation of the location of a target within a scene. Brain regions thought to extract the background of a scene could still support or facilitate the processing of layout and objects without contributing directly to the internal representation of a likely target location map.

Past work has also demonstrated general medial temporal lobe (MTL) involvement in tasks mediated by contextual cueing. The hippocampus, a portion of the MTL, has typically been associated with declarative (or explicit) memory processes in humans (Squire, 1992). However, observers with MTL damage and otherwise unimpaired implicit perceptual learning were impaired on contextual learning tasks (Chun & Phelps, 1999; Giesbrecht, Sy, & Guerin, 2013). Facilitation of target detection in contextual learning tasks is known to result from implicit learning of spatial associations between background and target elements (Chun & Jiang, 1998). The observed deficits in implicit processing are surprising considering the wealth of results relating MTL function to exclusively declarative memory processes. Furthermore, MTL damage has sometimes conflictingly been shown not to impair contextual cueing facilitation (Preston & Gabrieli, 2008; Squire, Shimamura, & Amaral, 1989). Studies conducted to make sense of these opposing results have suggested that the MTL is critical for processing spatial associations and providing online feedback

to the visual system to guide attentional allocation (Giesbrecht et al., 2013; Kasper, Grafton, Eckstein, & Giesbrecht, 2015). This stimulus set serves as a way to compare contextual guidance in normal and amnesic patients using realistic stimuli to further our understanding of the involvement of hippocampal regions in the possible integration of scene information and modulation of brain areas underlying visual processing. Given that the facilitation of target detection by contextual cues observed in our scenes presumably arises from spatial associations learned in real environments throughout observers' lives, contextual cueing in these scenes is not of the traditional, implicit variety elicited by artificial displays (á la Chun & Jiang, 1998). We might still expect to see recruitment of the MTL and subsequent modulation of visual area activity (especially the ventral stream; Westerberg, Miller, Reber, Cohen, & Paller, 2011) underlying the effects observed in this work.

Not all cues are spatial

Though our efforts were mostly focused on improving our understanding of spatial contextual cues given their prevalence in the literature, we have also demonstrated that not all cues are spatial in nature. Expectations about the features of particular cues can also have an impact on observers' performance on visual search tasks. Overall, it appears that non-spatial cues are effective in causing an observer to ignore candidate target regions that might otherwise be explored if a target of expected scale were present there. Specifically, observers are worse at detecting mis-scaled objects than objects of normal scale, despite the mis-scaled objects being more conspicuous and easily detectable when scaled properly relative to their surroundings or in isolation. However, we did not observe a strong effect on observers' eye movement behavior on mis-scaled trials relative to normal trials. Although there was a tendency for observers to fixate more closely to the target region on normal trials than on mis-scaled target trials, further inspection of hit versus miss

trials revealed that eye movement guidance is highly similar between normal and mis-scaled trials. When observers miss a target in both cases, it is because they did not fixate the target region, and show fixation proximity to the target region more similar to that of target absent trials. This suggests that observers' eye movements are simply not guided to the location of the mis-scaled target. Observers consistently fail to detect the mis-scaled target until the final several trials of the experiment, suggesting this effect is robust and persistent despite some prevalence of correctly identifying mis-scaled objects on some trials and thus conscious awareness that targets may be mis-scaled. Although humans perform poorly at detecting mis-scaled objects, in natural environments this cue helps them eliminate false object detection in ways that state-of-the-art computer vision models are unable to, clarifying a factor in the continued divide between human and computer object detection.

Overall conclusions and suggestions for future directions

Together, our results improve the current understanding of what specific portions of scene images are facilitating observers' detection of objects in natural environments. These results have implications in informing biologically inspired models of human vision and computer vision object detection models. It is likely the case that object detection and understanding will be a more useful component of successful scene understanding in machines than background or global image processing, unless global image processing can capture structural information provided by object information.

This work aligns with the recent goal of the scene understanding literature to improve its methods by empirically quantifying the image statistics that relate to a particular cue or to verify that experimenter scene manipulations are accurate operationalizations of a particular cue. Currently, this aim is being motivated by observations that humans are poor estimators of the statistics of object and scene properties in their environments (e.g., Greene, 2016), and therefore

should not manipulate images based on intuition. Although it is important to be cautious of biases when creating experiments that are premised on what is or is not common in the real world, exhaustive data about the state of the world is not available in the quantity necessary to inform rigorous, data-driven stimulus production. Our work proposes an alternative approach to improving contextual cue methodology, which is to measure the informativeness of a cue to a particular task, and relate that measure to the influence that altering the cue has on other behavioral measures.

Finally, our review of the literature emphasized the problematic use of the descriptor “scene gist” for any perceptual impression rapidly obtained from a scene. We have avoided explicitly relating any of our own cues to scene gist. The stimuli used in the spatial cue experiments could provide insight into what aspects of a scene contribute to various definitions of scene gist and clarify its future use in the literature.

References

Aminoff, E., Gronau, N., & Bar, M. (2007). The parahippocampal cortex mediates spatial and nonspatial associations. *Cerebral Cortex*, *17*(7), 1493–1503.

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology*, *103*(1), 62–70. http://doi.org/10.1037/h0036799

Antes, J. R., Penland, J. G., & Metzger, R. L. (1981). Processing global information in briefly presented pictures. *Psychological Research*, *43*(3), 277–292.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629. http://doi.org/10.1038/nrn1476

Bar, M., Ullman, S., & others. (1996). Spatial context in recognition. *PERCEPTION-LONDON-*, *25*, 343–352.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*(43), 77–80.

Biederman, I. (1981). *On the semantics of a glance at a scene*.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision Research*, *49*(24), 2992–3000. http://doi.org/10.1016/j.visres.2009.09.014

Boucart, M., Moroni, C., Szaffarczyk, S., & Tran, T. H. C. (2013). Implicit Processing of Scene Context in Macular Degeneration. *Investigative Opthalmology & Visual Science*, *54*(3), 1950. http://doi.org/10.1167/iovs.12-9680

Boucart, M., Moroni, C., Thibaut, M., Szaffarczyk, S., & Greene, M. (2013). Scene categorization at large visual eccentricities. *Vision Research*, *86*, 35–42. http://doi.org/10.1016/j.visres.2013.04.006

Boyce, S. J., & Pollatsek, A. (1992). An exploration of the effects of scene context on object identification. In *Eye movements and visual cognition* (pp. 227–242). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4612-2852-3_13

Brockmole, J. R., Castelhano, M. S., & Henderson, J. M. (2006). Contextual cueing in naturalistic scenes: Global and local contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 699–706. http://doi.org/10.1037/0278-7393.32.4.699

Burgess, A. (1985). Visual signal detection. III. On Bayesian use of prior knowledge and cross correlation. *JOSA A*, *2*(9), 1498–1507.

Calvo, M. G., Nummenmaa, L., & Hyönä, J. (2008). Emotional scenes in peripheral vision: Selective orienting and gist processing, but not content identification. *Emotion*, *8*(1), 68.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525. http://doi.org/10.1016/j.visres.2011.04.012

Caspi, A., Beutter, B. R., & Eckstein, M. P. (2004). The time course of visual information accrual guiding eye movement decisions. *Proceedings of the National Academy of*

Sciences of the United States of America, *101*(35), 13086–13090. http://doi.org/10.1073/pnas.0305329101

Castelhano, M. S., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. *Attention, Perception, & Psychophysics*, *72*(5), 1283–1297.

Castelhano, M. S., & Heaven, C. (2011). Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. *Psychonomic Bulletin & Review*, *18*(5), 890–896. http://doi.org/10.3758/s13423-011-0107-8

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*(4), 753.

Castelhano, M. S., & Henderson, J. M. (2008). The influence of color on the perception of scene gist. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 660.

Castelhano, M. S., & Henderson, J. M. (2010). Flashing scenes and moving windows: an effect of initial scene gist on eye movements. *Journal of Vision*, *3*(9), 67–67. http://doi.org/10.1167/3.9.67

Chen, X., & Zelinsky, G. J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research*, *46*(24), 4118–4133. http://doi.org/10.1016/j.visres.2006.08.008

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, *4*(5), 170–178.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*(1), 28–71.

Chun, M. M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, *10*(4), 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. http://doi.org/10.1038/12222

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*(8), 559–564.

De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, *52*(4), 317–329.

Divvala, S. K., Hoiem, D., Hays, J. H., Efros, A. A., & Hebert, M. (2009). An empirical study of context in object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009* (pp. 1271–1278). http://doi.org/10.1109/CVPR.2009.5206532

Dosher, B. A., & Lu, Z. L. (2000). Mechanisms of perceptual attention in precuing of location. *Vision Research*, *40*(10–12), 1269–1292.

Droll, J. A., Abbey, C. K., & Eckstein, M. P. (2009). Learning cue validity through performance feedback. *Journal of Vision*, *9*(2), 18.1-23. http://doi.org/10.1167/9.2.18

Droll, J., & Eckstein, M. (2010). Expected object position of two hundred fifty observers predicts first fixations of seventy seven separate observers during search. *Journal of Vision*, *8*(6), 320–320. http://doi.org/10.1167/8.6.320

Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science*, *9*(2), 111.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, *11*(5). Retrieved from http://www.journalofvision.org/content/11/5/14.short

Eckstein, M. P., Beutter, B. R., Pham, B. T., Shimozaki, S. S., & Stone, L. S. (2007). Similar neural representations of the target for saccades and perception during search. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *27*(6), 1266–1270. http://doi.org/10.1523/JNEUROSCI.3975-06.2007

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychological Science*, *17*(11), 973–980.

Eckstein, M. P., Mack, S. C., Liston, D. B., Bogush, L., Menzel, R., & Krauzlis, R. J. (2013). Rethinking human visual attention: spatial cueing effects and optimality of decisions by honeybees, monkeys and humans. *Vision Research*, *85*, 5–19. http://doi.org/10.1016/j.visres.2012.12.011

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 1–26.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, *17*(6–7), 945–978. http://doi.org/10.1080/13506280902834720

Epstein, R., Graham, K. S., & Downing, P. E. (2003). Viewpoint-specific scene representations in human parahippocampal cortex. *Neuron*, *37*(5), 865–876.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. http://doi.org/10.1038/33402

Ernst, M. O. (2006). A Bayesian view on multimodal cue integration. *Human Body Perception from the inside out*, *131*, 105–131.

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), 10–10. http://doi.org/10.1167/7.1.10

Giesbrecht, B., Sy, J. L., & Guerin, S. A. (2013). Both memory and attention systems contribute to visual search for targets cued by implicitly learned context. *Vision Research*, *85*, 80–89.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1974). Wiley New York. Retrieved from http://andrei.gorea.free.fr/Teaching_fichiers/SDT%20and%20Psytchophysics.pdf

Greene, M. R. (2013). Statistics of high-level scene context. *Frontiers in Psychology*, *4*. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3810604/

Greene, M. R. (2016). Estimations of object frequency are frequently overestimated. *Cognition*, *149*, 6–10.

Grill-Spector, K., Kushnir, T., Hendler, T., & Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. *Nature Neuroscience*, *3*(8), 837–843. http://doi.org/10.1038/77754

Groen, I. I. A., Ghebreab, S., Prins, H., Lamme, V. A. F., & Scholte, H. S. (2013). From Image Statistics to Scene Gist: Evoked Neural Activity Reveals Transition from Low-Level Natural Image Structure to Scene Category. *The Journal of Neuroscience*, *33*(48), 18814–18824. http://doi.org/10.1523/JNEUROSCI.3128-13.2013

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv Preprint arXiv:1512.03385*. Retrieved from http://arxiv.org/abs/1512.03385

Henderson, J. M., & Hollingworth, A. (1998). Eye movements during scene viewing: An overview. *Eye Guidance in Reading and Scene Perception*, *11*, 269–293.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, *50*(1), 243–271.

Henderson, J. M., Mcclure, K. K., Pierce, S., & Schrock, G. (1997). Object identification without foveal vision: Evidence from an artificial scotoma paradigm. *Perception & Psychophysics*, *59*(3), 323–346. http://doi.org/10.3758/BF03211901

Henderson, J. M., Weeks Jr, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(1), 210.

Hoffman, J. E. (1979). A two-stage model of visual search. *Perception & Psychophysics*, *25*(4), 319–327.

Hollingworth, A. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, *127*(4), 398.

Hwang, A. D., Wang, H.-C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research*, *51*(10), 1192–1205.

Inhoff, M. C., & Ranganath, C. (2015). Significance of objects in the perirhinal cortex. *Trends in Cognitive Sciences*, *19*(6), 302–303.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10–12), 1489–1506. http://doi.org/10.1016/S0042-6989(99)00163-7

Jiang, Y., & Wagner, L. C. (2004). What is learned in spatial contextual cuing—configuration or individual locations? *Perception & Psychophysics*, *66*(3), 454–463. http://doi.org/10.3758/BF03194893

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision*, *3*(7), 4–4. http://doi.org/10.1167/3.7.4

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research*, *47*(26), 3286–3297.

Kasper, R. W., Grafton, S. T., Eckstein, M. P., & Giesbrecht, B. (2015). Multimodal neuroimaging evidence linking memory and attention systems during visual search cued by context. *Annals of the New York Academy of Sciences*, *1339*(1), 176–189.

Keysers, C., Xiao, D., Földiák, P., & Perrett, D. I. (2001). The speed of sight. *Cognitive Neuroscience, Journal of*, *13*(1), 90–101.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, *4*(4), 219–27.

Koopman, B. O. (1956). The theory of search. I. Kinematic bases. *Operations Research*, *4*(3), 324–346.

Koopman, B. O. (1956). The theory of search. II. Target detection. *Operations Research*, *4*(5), 503–531.

Koopman, B. O. (1957). The theory of search III. The optimum distribution of searching effort. *Operations Research*, *5*(5), 613–626.

Larson, A. M., Freeman, T. E., Ringer, R. V., & Loschky, L. C. (2014). The spatiotemporal dynamics of scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 471.

Larson, A. M., & Loschky, L. C. (2009). The Contributions of Central Versus Peripheral Vision to Scene Gist Recognition. *Journal of Vision*, *9*(10). http://doi.org/10.1167/9.10.6

Levi, D. M. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, *48*(5), 635–654. http://doi.org/10.1016/j.visres.2007.12.009

Levy, I., Hasson, U., Avidan, G., Hendler, T., & Malach, R. (2001). Center–periphery organization of human object areas. *Nature Neuroscience*, *4*(5), 533–539.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, *99*(14), 9596–9601. http://doi.org/10.1073/pnas.092277599

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014* (pp. 740–755). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-10602-1_48

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

Loschky, L., Boucart, M., Szaffarczyk, S., Beugnet, C., Johnson, A., & Tang, J. L. (2015). The contributions of central and peripheral vision to scene gist recognition with a 180° visual field. *Journal of Vision*, *15*(12), 570. http://doi.org/10.1167/15.12.570

Loschky, L. C., & Larson, A. M. (2010). The natural/man-made distinction is made before basic-level distinctions in scene gist processing. *Visual Cognition*, *18*(4), 513–536. http://doi.org/10.1080/13506280902937606

Luck, S. J., Hillyard, S. A., Mouloua, M., & Hawkins, H. L. (1996). Mechanisms of visual-spatial attention: resource allocation or uncertainty reduction? *Journal of Experimental Psychology. Human Perception and Performance*, *22*(3), 725–737.

Mack, S. C., & Eckstein, M. P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of Vision*, *11*(9), 9–9. http://doi.org/10.1167/11.9.9

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics*, *2*(11), 547–552. http://doi.org/10.3758/BF03210264

Malcolm, G. L., & Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: evidence from eye movements. *Journal of Vision*, *9*(11), 8.1-13. http://doi.org/10.1167/9.11.8

Maunsell, J. H. R., & Cook, E. P. (2002). The role of attention in visual processing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *357*(1424), 1063–1072. http://doi.org/10.1098/rstb.2002.1107

Metzger, R. L., & Antes, J. R. (1983). The nature of processing early in picture perception. *Psychological Research*, *45*(3), 267–274.

Moores, E., Laiti, L., & Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. *Nature Neuroscience*, *6*(2), 182–189. http://doi.org/10.1038/nn996

Moraglia, G. (1989). Visual search: Spatial frequency and orientation. *Perceptual and Motor Skills*, *69*(2), 675–689.

Munneke, J., Brentari, V., & Peelen, M. V. (2013). The influence of scene context on object recognition is independent of attentional focus. *Frontiers in Psychology*, *4*. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3748376/

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*(5), 614–621.

Oliva, A. (2005). Gist of the scene. *Neurobiology of Attention*, *696*(64), 251–258.

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, *42*(3), 145–175. http://doi.org/10.1023/A:1011139631724

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

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

Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on* (Vol. 1, p. I–253).

Olson, I. R., & Chun, M. M. (2002). Perceptual constraints on implicit learning of spatial context. *Visual Cognition*, *9*(3), 273–302.

Palmer, tephen E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, *3*(5), 519–526. http://doi.org/10.3758/BF03197524

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123. http://doi.org/10.1016/S0042-6989(01)00250-4

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, *11*(10), 1129–1135. http://doi.org/10.1038/nn.2187

Pereira, E. J., & Castelhano, M. S. (2014). Peripheral guidance in scenes: The interaction of scene context and object content. Retrieved from http://psycnet.apa.org/psycinfo/2014-31982-001/

Peterson, M. S., & Kramer, A. F. (2001). Attentional guidance of the eyes by contextual information and abrupt onsets. *Perception & Psychophysics*, *63*(7), 1239–1249.

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*(2), 160–174. http://doi.org/10.1037/0096-3445.109.2.160

Potter, M. C. (1975). Meaning in visual search. *Science (New York, N.Y.)*, *187*(4180), 965–966.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 509.

Preston, A. R., & Gabrieli, J. D. E. (2008). Dissociation between Explicit Memory and Configural Memory in the Human Medial Temporal Lobe. *Cerebral Cortex*, *18*(9), 2192–2207. http://doi.org/10.1093/cercor/bhm245

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research*, *42*(11), 1447–1463.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems* (pp. 91–99). Retrieved from http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks

Rousselet, G., Joubert, O., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*, *12*(6), 852–877. http://doi.org/10.1080/13506280444000553

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*(1–3), 157–173.

Sagi, D. (1988). The combination of spatial frequency and orientation is effortlessly perceived. *Perception & Psychophysics*, *43*(6), 601–603. http://doi.org/10.3758/BF03207749

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science*, *5*(4), 195–200.

Serences, J. T., & Yantis, S. (2006). Selective visual attention and perceptual coherence. *Trends in Cognitive Sciences*, *10*(1), 38–45. http://doi.org/10.1016/j.tics.2005.11.008

Shimozaki, S. S., Eckstein, M. P., & Abbey, C. K. (2003). An ideal observer with channels versus feature-independent processing of spatial frequency and orientation in visual search performance. *Journal of the Optical Society of America A*, *20*(12), 2197. http://doi.org/10.1364/JOSAA.20.002197

Shimozaki, S. S., Eckstein, M. P., & Abbey, C. K. (2003). Comparison of two weighted integration models for the cueing task: Linear and likelihood. *Journal of Vision*, *3*(3). Retrieved from http://w.journalofvision.org/content/3/3/3.short

Smith, P. L. (2000). Attention and luminance detection: effects of cues, masks, and pedestals. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(4), 1401.

Spotorno, S., Malcolm, G. L., & Tatler, B. W. (2014). How context information and target information guide the eyes from the first epoch of search in real-world scenes. *Journal of Vision*, *14*(2), 7–7. http://doi.org/10.1167/14.2.7

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*(2), 195–231. http://doi.org/http://dx.doi.org/10.1037/0033-295X.99.2.195

Squire, L. R., Shimamura, A. P., & Amaral, D. G. (1989). Memory and the hippocampus. *Neural Models of Plasticity: Experimental and Theoretical Approaches*, 208–239.

Strasburger, H., Rentschler, I., & Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision*, *11*(5). Retrieved from http://ww.w.journalofvision.org/content/11/5/13.short

Stuart, G. W., Bossomaier, T. R. J., & Johnson, S. (1993). Preattentive processing of object size: implications for theories of size perception. *Perception*, *22*(10), 1175–1193. http://doi.org/10.1068/p221175

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5).

Thorpe, S., Fize, D., Marlot, C., & others. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522.

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M., & Bülthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience*, *14*(5), 869–876. http://doi.org/10.1046/j.0953-816x.2001.01717.x

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*(3), 391–412.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, *95*(1), 15–48. http://doi.org/http://dx.doi.org/10.1037/0033-295X.95.1.15

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Trommershauser, J., Kording, K., & Landy, M. S. (2011). *Sensory cue integration*. Oxford University Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=M41pAgAAQBAJ&oi=fnd&pg=PP1&dq=landy+cue+combination&ots=afosw5upGZ&sig=hHK_BCkrtTnxeSENN3lFdvtsQbI

Verghese, P., & Nakayama, K. (1994). Stimulus discriminability in visual search. *Vision Research*, *34*(18), 2453–2467. http://doi.org/10.1016/0042-6989(94)90289-5

Võ, M. L.-H., & Henderson, J. M. (2011). Object–scene inconsistencies do not capture gaze: evidence from the flash-preview moving-window paradigm. *Attention, Perception, & Psychophysics*, *73*(6), 1742–1753.

Võ, M. L.-H., & Schneider, W. X. (2010). A glimpse is not a glimpse: Differential processing of flashed scene previews leads to differential target search benefits. *Visual Cognition*, *18*(2), 171–200. http://doi.org/10.1080/13506280802547901

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*(9), 1395–1407.

Walthew, C., & Gilchrist, I. D. (2006). Target location probability effects in visual search: an effect of sequential dependencies. *Journal of Experimental Psychology. Human Perception and Performance*, *32*(5), 1294–1301. http://doi.org/10.1037/0096-1523.32.5.1294

Ward, E. J., MacEvoy, S. P., & Epstein, R. A. (2010). Eye-centered encoding of visual space in scene-selective regions. *Journal of Vision*, *10*(14), 6–6. http://doi.org/10.1167/10.14.6

Wasserman, E. A., Teng, Y., & Castro, L. (2014). Pigeons exhibit contextual cueing to both simple and complex backgrounds. *Behavioural Processes*, *104*, 44–52. http://doi.org/10.1016/j.beproc.2014.01.021

Westerberg, C. E., Miller, B. B., Reber, P. J., Cohen, N. J., & Paller, K. A. (2011). Neural correlates of contextual cueing are modulated by explicit learning. *Neuropsychologia*, *49*(12), 3439–3447.

Whitney, D., & Levi, D. M. (2011). Visual crowding: a fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, *15*(4), 160–168. http://doi.org/10.1016/j.tics.2011.02.005

Williams, L. G. (1966). Target conspicuity and visual search. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *8*(1), 80–92.

Williams, L. G. (1966). The effect of target specification on objects fixated during visual search. *Perception & Psychophysics*, *1*(9), 315–318. http://doi.org/10.3758/BF03215795

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238.

Wolfe, J. M. (1998). Visual search. *Attention*, *1*, 13–73.

Wolfe, J. M. (2014). Approaches to visual search: Feature integration theory and guided search. In A. C. Nobre & S. Kastner (Eds.), *Oxford Handbook of Attention*. New York: Oxford University Press.

Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., & Sherman, A. M. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception, & Psychophysics*, *73*(6), 1650–1671.

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.

Wu, C.-C., Wang, H.-C., & Pomplun, M. (2014). The roles of scene gist and spatial dependency among objects in the semantic guidance of attention in real-world scenes. *Vision Research*, *105*, 10–20. http://doi.org/10.1016/j.visres.2014.08.019

Wu, C.-C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology*, *5*. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3915098/

Yoo, S.-A., & Chong, S. C. (2012). Eccentricity biases of object categories are evident in visual working memory. *Visual Cognition*, *20*(3), 233–243. http://doi.org/10.1080/13506285.2012.663416

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*(4), 787.

Zelinsky, G., Zhang, W., Yu, B., Chen, X., & Samaras, D. (2005). The role of top-down and bottom-up processes in guiding eye movements during visual search. In *Advances in neural information processing systems* (pp. 1569–1576). Retrieved from http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2005_727.pdf

Appendix

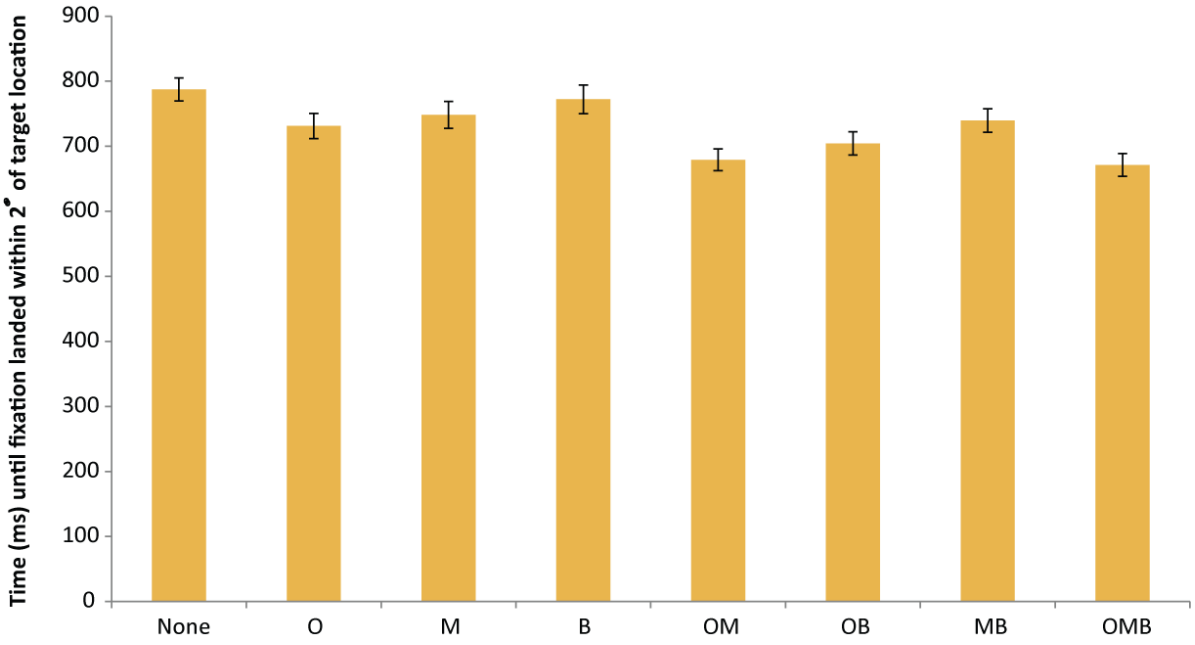

Figure S1: Average time until a fixation landed within two degrees of the target location across all contextual information conditions.

Cue combination derivation

Here, we will demonstrate the derivation of the isolated cue sensitivities used in our assessment of the additivity of contextual cue information, premised on a signal detection theory framework, similar to that presented in Eckstein (1998). We assume that each scene will elicit an internal response in the observer, x_{cue}, dependent upon whether the target and various cues are present in the image and subject to internal noise. The internal response of an observer to a target absent (noise) and target present (signal) scene can be represented by Gaussian probability distributions. The observers' ability to discriminate between target present and target absent images is represented by the index of detectability, d', and is equivalent to the difference in mean internal responses to the signal and noise response distributions, weighted by the variance of the signal, assuming equal signal and noise distribution variance (Equation A.1).

$$d'_x = \frac{\langle x_s \rangle - \langle x_n \rangle}{\sigma_x} \qquad \text{[Equation A.1]}$$

We assume that the mean internal response to a target absent image containing a given cue, $\langle x_{cue}, n \rangle$, is 0 and therefore that the mean internal response to a target present image containing the same cue, $\langle x_{cue}, s \rangle$, is equal to d' for that cue, d'_{cue}. When cues are combined within an image, we are testing the hypothesis that the combined internal response (y) is a weighted linear sum of the

internal responses for the individually cued images (Equation A.2), and seek to derive the optimal combination of d'_{cue} in such a case.

$$y = w_1 x_{cue_1} + w_2 x_{cue_2} + w_3 x_{cue_3} \quad \text{[Equation A.2]}$$

Using Equation A.1, we can derive d'_y given $\langle y_s \rangle$, $\langle y_n \rangle$, and σ_y. We can calculate the expected value of the signal and noise distributions of y using Equations A.3 and A.4. We assume that $\langle x_{cue,n} \rangle = 0$, therefore $\langle y_n \rangle = 0$. Similarly, we have defined $\langle x_{cue,s} \rangle = d'_{cue}$ and know that the optimal weighting of a linear combination of multiple independent cues should be proportional to the information provided by those cues (represented by the index of detectability for that cue; Green & Swets, 1966; Shimozaki, Eckstein, & Abbey, 2003), therefore $w_{cue} = d'_{cue}$ and $\langle y_s \rangle = d'^2_{cue1} + d'^2_{cue2} + d'^2_{cue3}$.

$$\langle y_n \rangle = w_1 \langle x_{cue1,n} \rangle + w_2 \langle x_{cue2,n} \rangle + w_3 \langle x_{cue3,n} \rangle = 0 \quad \text{[Equation A.3]}$$

$$\langle y_s \rangle = w_1 \langle x_{cue1,s} \rangle + w_2 \langle x_{cue2,s} \rangle + w_3 \langle x_{cue3,s} \rangle \quad \text{[Equation A.4]}$$

Finally, the standard deviation of y is derived using Equation A.5 and solved in A.6, noting that we assume unit variance.

$$\sigma_y^2 = \left[\frac{\partial y}{\partial x_{cue1}}\right]^2 \sigma_{x_{cue1}}^2 + \left[\frac{\partial y}{\partial x_{cue2}}\right]^2 \sigma_{x_{cue2}}^2 + \left[\frac{\partial y}{\partial x_{cue2}}\right]^2 \sigma_{x_{cue2}}^2 \quad \text{[Equation A.5]}$$

$$\sigma_y = \sqrt{d'^2_{cue_1} + d'^2_{cue_2} + d'^2_{cue_3}} \quad \text{[Equation A.6]}$$

Returning to Equation A.1, using Equations A.3, A.4, and Equation A.6, we can derive the optimal linear additive combination of d',

$$d'_y = \sqrt{d'^2_{cue_1} + d'^2_{cue_2} + d'^2_{cue_3}} \quad \text{[Equation A.7]}$$

Using the result in Equation A.7, we can also derive an equation to isolate the effect on d' due to a single cue, relative to the no cue condition. We assume that when no cues are present, there is a baseline sensitivity, d'_{none}, and when a single cue is present, d' for that image, $d'_{cue,none}$, is the optimal combination (according to Equation A.7) of d'_{cue} and d'_{none}, as shown in Equation A.8. Therefore we solve for d'_{cue} to derive the isolated effect on d' of a given cue (Equation A.9).

$$d'_{cue,none} = \sqrt{d'^2_{none} + d'^2_{cue}} \quad \text{[Equation A.8]}$$

$$d'_{cue} = \sqrt{d'^2_{cue,none} - d'^2_{none}} \quad \text{[Equation A.9]}$$

Tables of linear regression results

The tables below present the results of using the mode of closest fixations' to target and expected target locations x- and y-coordinates in the conditions with a single cue to predict the same for the conditions with images containing all cues. See the main text for highlights of the results concerning the partial correlations, but refer below to assess the zero-order correlations and model coefficients.

*significant at the 0.05 level
**significant at the 0.01 level
***significant at the 0.001 level

One fixation allowance; using singly cued fixations to predict fixations in images with all cues:

	Zero-order *r*						β		pr		*b*	
	M		B		OMB							
	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
O	0.25*	0.42**	0.62***	0.54***	0.58***	0.35**	0.48	0.01	0.46	0.01	0.50**	0.01
M			0.14	0.30*	0.51***	0.73***	0.39	0.70	0.46	0.68	0.61**	0.73***
B					0.36**	0.31*	0.01	0.10	0.01	0.12	0.01	0.09
				$R^{2:}$	0.48***	0.54***			Intercept:		-73.84	59.04

Two fixation allowance; using singly cued fixations to predict fixations in images with all cues:

	Zero-order *r*						β		pr		*b*	
	M		B		OMB							
	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
O	0.56***	0.38**	0.50***	0.62***	0.80***	0.66***	0.54	0.64	0.67	0.56	0.63***	0.50***
M			0.55***	0.43**	0.72***	0.39**	0.34	0.17	0.48	0.21	0.55***	0.17
B					0.60***	0.39**	0.15	-0.08	0.24	-0.08	0.27	-0.07
				$R^{2:}$	0.77***	0.46***			Intercept:		-214.60	141.85

Three fixation allowance; using singly cued fixations to predict fixations in images with all cues:

	Zero-order *r*						β		pr		*b*	
	M		B		OMB							
	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
O	0.84***	0.56***	0.64***	0.75***	0.86***	0.80***	0.44	0.66	0.48	0.59	0.52**	0.65***
M			0.70***	0.54***	0.85***	0.60***	0.36	0.23	0.39	0.31	0.46*	0.24*
B					0.71***	0.64***	0.18	0.02	0.28	0.03	0.22	0.02
				$R^{2:}$	0.81***	0.67***			Intercept:		-119.29	24.27