IN AND OUT OF CONTEXT:

THE EFFECTS OF VISUAL EXPERIENCE AND VISUAL DISTINCTIVENESS ON MEMORY FOR SCENES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Claudia M. Gilson

August 2013

IN AND OUT OF CONTEXT:
THE EFFECTS OF VISUAL EXPERIENCE AND VISUAL DISTINCTIVENESS ON MEMORY FOR SCENES

Claudia M. Gilson, Ph.D.

Cornell University, 2013

Visual context is the distillation of visual experience that enables reasonable accuracy in interpreting the natural world. Context consists of all the constraints that limit possibilities in recognizing objects and surfaces, inferring their properties, and other cognitive processes. We contend that these visual constraints are statistically learned from experience with the world. We also contend that context, once well-established through repetition of similar views of an environment, can impede memory for images of that environment.

Two aspects of context are fundamental to scene perception. *Identity* context is co-occurrence of objects and surfaces within the scene. *Spatial* context is how objects and surfaces are laid out in the scene, and their spatial relationships. When objects and surfaces are not within observers' normal experience, either in identity or spatial context, observers are quick to notice. *Inconsistent object* effects and *spatial relationship violations* have been shown in research with real-world scenes.

Experiments presented here reproduce inconsistent object and spatial relationship violation effects using novel abstract objects, demonstrating learning from a baseline without reference to real-world objects. In delayed match to sample tasks, participants learned spatial context quickly for multiple objects within the same scene. Learned spatial

contexts markedly improved sensitivity to changes in an object’s location when it appeared outside the learned area. Participants also learned identity context for novel objects, but only when at least one object differed markedly in visual appearance from others in the scene.

Experiments with a very well-established context, natural scenery without human or animal presence, demonstrated that context can interfere with memory for individual images from the context. In memory comparisons with 100-image sets of a variety of human-oriented content, natural scenery, 1/f noise, or white noise, performance was lower for natural scenery image sets than for high variety. However, natural scenery performance was better than for white noise and 1/f noise. Poor performance was not due to inability to discriminate when exemplars are viewed side-by-side. Participants can also remember at least four exemplars of difficult-to-discriminate images, including noise.

BIOGRAPHICAL SKETCH

Claudia M. Gilson received a B.A in anthropology from Arizona State University in 1981, and an M.S. in communication from Rensselaer Polytechnic Institute (RPI) in 1987. After a career as a teacher of reading and English composition, she worked in the computer industry as a technical writer and interface designer. She returned to RPI for an additional M.S. in cognitive psychology in 2001. As a research specialist at the Lighting Research Center at RPI, she focused on human factors issues in lighting, and co-authored a number of journal articles and technical papers. She was awarded the 2004 Walsh Weston Medal for best paper from the Society for Light and Lighting. At Cornell University, she studied how visual experience forms contexts that include the structure of scenes, and how contexts affect long-term scene memory. She received a Dallenbach Research Fellowship at Cornell University, as well as the 2009 Graduate Student Teaching Award.

ACKNOWLEDGEMENTS

To my advisor, Shimon Edelman, who took a chance on me and taught me an entirely new way of thinking; to David Field, who can make the most complex ideas clear even to a humanities major, and whose support and encouragement never failed me;
to James Cutting, the epitome of the gentleman and the scholar, whose knowledge of science and art is encyclopedic beyond the dreams of search engines; to Dennis Regan, who kept me grounded in the world of research beyond perception and statistics; to each of these men, my greatest respect, gratitude, and affection.

To Andrea Knezevic, Ben Notterman, Jennifer Caglayan, Hannah Shamji, and Hua Guo, many thanks for their contributions as my research assistants. Special thanks to Anne Warlaumont, for her collaboration as well as her expertise in programming.

To my husband, John Gilson, and to my mother, Virginia Engler—because you wanted this for me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
GENERAL INTRODUCTION

1. Introduction

Visual context is the distillation of visual experience that enables us to interpret, with reasonable accuracy, what we see in the world around us. Context consists of all the constraints that limit possibilities in recognizing objects and surfaces, inferring their properties, following motion trajectories, and other cognitive processes involved in comprehending the visual world (Torralba, 2003a). We contend that these constraints are statistically learned from phenomenal experience with the world.

For example, consider the context of your office at work. Every day you travel to the same building, park in the same lot if you're driving, walk through the same entrance and traverse the same corridors, greet many of the same people along the way, unlock the same door, seat yourself in the same chair at the same desk, and take up the same kinds of tasks using the same kinds of equipment you've used, perhaps, for years. You may have had a number of different offices throughout your career, and visited many other colleagues in their offices. In each situation, however, you found commonalities: a space filled with objects and surfaces—carpets, walls, windows, desks or tables, chairs, computers, phones, papers, books, writing instruments, etc. It is highly probable that these objects and surfaces were arranged in a similar way in each space: desks were on the floor, chairs were arranged next to them, papers and computers were on top of the desk (or possibly strewn on the floor), awards and diplomas hung on the walls. Although you saw variations on this theme, your visual system noted in each instance which objects were present and where they were situated. After a certain number of repetitions and refinements, the "office" context was established. You were able to differentiate it from other contexts that contained different collections of objects, or similar objects that were spatially arranged in different ways.

What kinds of constraints were being extracted during your "office" experiences? Although context is usually discussed as a unitary concept, the constraints of which it is comprised can be subdivided. Some are purely visual, while some arise from other kinds of associations, such as memories, emotions, and learned facts. For the purposes of this thesis, it will be useful to loosely conceive of two main types of context. *Spatial*, or *configurational*, context constrains where objects and surfaces may appear in a scene, and how the entities are spatially related to each other (Oliva & Torralba, 2007). The ground plane in an outdoor natural scene is an example; most kinds of animals and plants (such as deer and tree trunks) appear on the plane, while fewer types of objects and lifeforms (birds, treetops, aircraft) can appear suspended above the plane (Biederman, 1981). *Identity,* or *semantic,* context constrains what objects and surfaces may appear in a scene (Bar, 2004; Oliva & Torralba, 2007). This type of context specifies both the likely co-occurrence of objects as well as impermissible combinations. Forks and spoons are likely candidates for object recognition when seen in conjunction with plates on a table, while pliers and screwdrivers are less likely candidates.

Other contextual constraints beyond spatial and identity can be rich sources of information in scene understanding. In fact, what many people mean when they refer to recognizing or understanding what objects they are seeing "in context" are the many kinds of constraints that arise from understanding how visual properties of a scene imply other properties. For example, the absence of leaves on trees implies that the season is autumn or winter; tall pillars and other massive architectural elements imply that a building is likely to be a bank or church, rather than a private residence. These kinds of contextual constraints may far outnumber spatial or identity constraints. However, spatial and identity context, we propose, can be learned from visual experiences alone, while other aspects of context likely require extensive interaction, both visual and non-visual, with the real world. This thesis confines itself to questions about the establishment of spatial and identity context and their effects on visual memory through statistical learning.

2. Evidence for identity context: the "inconsistent object" effect

Object co-occurrence is perhaps the most salient aspect of visual context formation, because without the ability to tabulate co-occurrence, observers could not compute whole-scene contexts at all (Biederman, 1981; Mandler & Parker, 1976; Mandler & Ritchey, 1977). The existence of such contexts is at least partially demonstrated by such well-known visual effects as rapid scene categorization (Rousselet, Joubert, & Fabre-Thorpe, 2005), and boundary extension (Intraub & Richardson, 1989), In rapid scene categorization, images can be placed in broad categories such as *city street* or *beach* within a few milliseconds by the rapid recognition of some diagnostic surfaces and objects they contain (although not necessarily by full identification of all the entities in the scene). Boundary extension, in which observers being tested may identify or draw more elements of a scene than they were actually shown, has led to speculation that observers are so familiar with the elements of such scene contexts that their memories extrapolate those elements even when they are not present.

However, the effects of identity context on scene recognition can be most clearly seen in studies showing that objects that are inconsistent with their settings draw attention, or produce reaction time (RT) differences when compared to objects that are consistent with their settings. Lack of coherence, violations of expectations, and deviations from normality often produce better performance on a variety of visual tasks (Auckland, Cave, & Donnelly, 2007; Brockmole & Henderson, 2008; Davenport, 2007; Davenport & Potter, 2004; Hollingworth & Henderson, 1998; Joubert, Fize, Rousselet, & Fabre-Thorpe, 2008; Palmer, 1975; Pezdek, Maki, Valencia-Laver, Whetstone, Stoeckert, & Dougherty, 1988; Pezdek, Whetstone, Reynolds, Askari, & Dougherty, 1989). In other words, visual input that is "out of context" often proves more recognizable or memorable; this is arguably the basis for many artistic endeavors. The sculpture together with the landscape and buildings in Figure 1-1 provides a notable example.

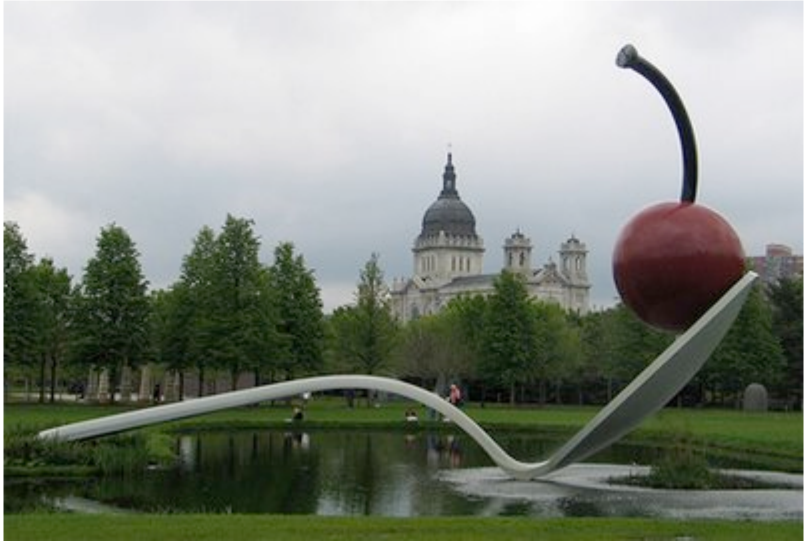


Figure 1-1: This Claes Oldenburg outdoor sculpture demonstrates the artistic motivation of visual input that is "out of context."

In other kinds of tasks, the inconsistent object effect results in longer RTs or longer first fixations when observers attend to and spend longer noting areas of scenes that violate expectations (Becker, Pashler, & Lubin, 2007; De Graef, Christiaens, & d'Ydewalle, 1990; Friedman, 1979; Loftus & Mackworth, 1978).

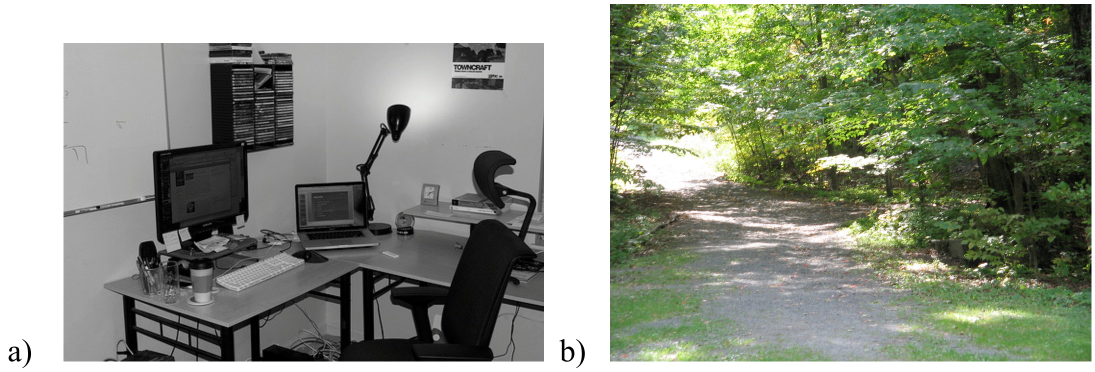
The research that brought the inconsistent object effect to light has necessarily relied on the use of objects that any observer would agree were inconsistent with their settings. The inconsistencies, however, relied at least as much on reasoning about the semantic associations of the scene as on inconsistencies that could be learned through visual experience alone. The well-known "octopus in the barnyard" drawing used by Loftus & Mackworth (1978) in their experiments illustrates the point. The drawing could have elicited such thoughts in the observer as, "the octopus is a sea creature," "sea creatures can't live out of water," "a barnyard is not an undersea environment," etc., leading to the inference that the octopus is inconsistent with the barnyard setting, perhaps at the same time as the purely visual process that returns the statistical unlikelihood of an object of that shape appearing with the other objects in the setting. To date, research that examines how identity context forms from a baseline, using novel objects and settings, has found that it forms with difficulty, and is usually overshadowed by spatial context effects (Chun & Jiang, 1999; Endo & Takeda, 2004; Jiang, Olson, & Chun, 2000; Jiang & Song, 2005). Why identity context effects are so pronounced in real-world scenes and so weak in novel stimulus displays has not been adequately explored.

3. Evidence for spatial context: the spatial violation effect

A similar inconsistency effect is produced when scenes include objects that violate normal spatial contexts, such as a sofa appearing outdoors on a sidewalk (Biederman, 1981), a Jeep hovering in the air (Neider & Zelinsky, 2006), or a television tipped over and balancing on a corner of its frame (Zimmerman, Schnier, & Lappe, 2010). Spatial context may have an even stronger influence on rapid, easy scene comprehension and almost certainly on scene memory than identity context. In fact, there is evidence that visual working memory is organized first spatially, then according to object properties (Vidal, Gauchou, Tallon-Baudry, & O'Regan, 2005; Schmidt, Vogel, Woodman, & Luck, 2002; Jiang, Olson, & Chun, 2000). According to these researchers, structural layout, or a set of locations, is first formed, and then objects and their attributes are encoded at those locations.

The reasons why spatial configuration is so important in the visual system have not been adequately explored. Because configuration changes at least slightly with every movement of the observer's eyes, head, or body, it would not seem advantageous to encode configuration very precisely. There is some evidence, however, that the episodic encoding of visual experience is viewpoint-dependent. Observers are less accurate in remembering both static scenes (Diwadkar & McNamara, 1997) and dynamic scenes (Garsoffky, Schwan, & and Hesse, 2002) if they are asked to recognize the same event shown from a different viewpoint from the original.

We probably recognize any number of canonical or stereotypical views of scenes, especially if an observer lives and travels in the same environment for an extended period of time (McNamara, Rump, & Werner, 2003). For example, the view of an office shown in Figure 1.2a is undoubtedly the same one that the inhabitant sees every time he unlocks his office door. Figure 1.2b shows a scene that would meet a walker along a forest path many times during his traversal of a well-traveled route. Both are stereotypical configurations for scenes viewed in the direction of movement from the observer's eye height. Some researchers posit that multiple stereotypical viewpoints of the same environment are encoded during prolonged interactions, leading to nearly configuration- or viewpoint-independent performance on recognition tasks (Garsoffky, Schwan, & Hesse, 2002).



Figures 1-2a and 1-2b: Stereotypical configurations of office and outdoor pathway scenes.

Besides the strong encoding produced by multiple exposures to the same viewpoint, it is possible that other mechanisms operate to make some views canonical. Learning that encodes the spatial relations among the objects in these scenes could also be associated at the same time with the proprioceptive input provided by eye and body movements, leading to an invariant response to changes in the spatial configurations caused by the observer's movements. Only if the configuration were accompanied by unusual proprioceptive cues would the invariance be broken. For example, if the same environment were viewed upside down (if one were being carried, or falling), the visual input would be associated with anomalous input from the vestibular system. In either case, repeated exposure to normal spatial contexts could sensitize the visual system to objects or areas of the scene that are markedly out of place.

Some of the most convincing evidence for the usefulness of spatial context in human vision has come from the labs of Torralba and Oliva, and their colleagues. These researchers have shown that spatial context drives the very rapid recognition of scene categories. In their view, spatial context has less to do with higher-level recognition of individual objects than with the spatial layout of coarse blobs of differing spatial frequencies and colors (Schyns & Oliva, 1994; Oliva & Schyns, 2000; Oliva & Torralba, 2007). Low-spatial-frequency maps of characteristic scene structures are usually enough to enable experimental participants to sort scenes into broad categories (street scene, living room, beach, etc.) in extremely short presentations (Bar, Kassam, Ghuman, Boshyan, Schmidt, Dale, Hamalainen, Marinkovic, Schacter, Roson, & Halgren, 2006; Greene & Oliva, 2009b; Oliva & Torralba, 2001; Torralba & Oliva, 2003; Joubert, Rousselet, Fabre-Thorpe & Fize, 2009; Walker Renninger & Malik, 2004). Acquired holistic representations of this kind are capable of directing attention to salient areas of scenes for more efficient object detection (Torralba, 2003a, 2003b).

We probably therefore learn spatial context quite readily. Like the literature on the inconsistent object effect, however, research on spatial violations has also relied on already-learned spatial relations in real-world scenes. Consequently, we know little about how easily spatial contexts are learned, and how great a degree of spatial displacement is required to produce a spatial violation effect equivalent to those found in research on real-world scenes.

A fair critique might be that experimental results with contextual learning done in the laboratory cannot approach the complexity of learned constraints that apply to real-world scenes. Although the evidence is that natural scenes produce a greater response in the visual pathways (at least through V1) than artificial stimuli (Simoncelli & Olshausen, 2001; Vinje & Gallant, 2002), laboratory experiments can still provide explanations for why some kinds of natural scenes are more memorable than others.

4. Effects of context on scene memory

Contrary to expectations, the least memorable images might be able to tell us the most about visual long-term memory (VLTM). Visual context, as we have seen, has strong effects on online perception. Research has concentrated on these effects, finding that context probably aids in overall scene categorization and recognition of individual objects. Little attention has been paid, however, to the predictable effects of context on long-term memory for scenes or pictures.

In order to form a context, as we have seen, the observer must have encountered a sufficient number of scenes containing the same or similar objects in the same or similar locations. During an ecologically valid, real-world experience, visual input comes as a continuous stream of fixations around the environment, consisting of innumerable very similar scenes. The more numerous and similar the scenes, the stronger any contextual effects will become. Therefore, to say that the objects and surfaces in a scene are "in context" is by definition a statement about the similarity of one scene to another. Pictures of the same content that is in context, either as regards identity or spatial context, are more similar to one another than pictures that are out of that context.

The more similar pictures are, the more difficult it is to differentiate among them, and thus, the more unlikely it is that any single picture from that set will be accurately stored in visual long-term memory. Therefore, observers are more likely to make errors in recalling whether they have seen a given picture. The nature of the picture's stored representation is open to question: that is, how much actual visual detail is stored or if the content or "gist" of the scene is stored as a single semantic label (Konkle, Brady, Alvarez, & Oliva, 2010b; Oliva, 2004). If pictures differ in visual details rather than semantic content, how much different do the details have to be in order that VLTM creates a separate representation for each picture?

Consider the effects of real-world experience on the formation of visual context. Discounting exposure through television, movies, internet, and other media, the number of exposures a person has to the scenes of his own environment—his bedroom, kitchen, route to work, office, favorite bar, etc.—will be orders of magnitude greater than the number of exposures he might have had to any other environment or scene content. If an observer enters his own office 15 to 20 times a day, he will have seen approximately the same objects in the same configuration 3750 to 5000 times over the course of a year (assuming he does not change or rearrange the office's furnishings). How many times, in comparison, is the average person likely to have seen the Grand Canyon or the Eiffel Tower, unless he is a denizen of those districts?

Repeated exposure to scenes containing the same objects seen at approximately the same angles over a long period of time may cause the observer to generalize the content of a particular scene as a sort of archetype or template that "averages" among all the instances of similar details (Schacter & Addis, 2007). What form such archetypical scenes might take is unknown at present; however, a generalized representation necessarily omits actual physical details, such as precise shape or spatial location of an object. Therefore, observers are unlikely to be able to identify which of two instances of a generalized scene they have actually been exposed to. Over a large number of pictures of similar content, it follows, accuracy in recognition will be low.

In contrast, most picture memory research findings are that VLTM for detailed real-world scenes is quite good. Recognition memory in the typical study is usually above 80%

accurate, and sometimes almost at ceiling (Standing, 1973), for hundreds or thousands of pictures. A typical study uses a large set of pictures that are all different from one another in content, whether implicitly or explicitly stated by the researchers. In early studies, the ways in which the content differed were not usually described, much less quantified, although more current studies have done so.

Picture content that is easily differentiated is probably the most salient predictor of accurate performance in picture memory studies. However, there are other factors that might help account for good visual memory. First, many picture sets include content that can easily be represented by a linguistic label: for example, “boy with red balloon,” “Eiffel Tower,” or “mother and baby giraffe.” Because these pictures’ content is usually singular across the picture set, the label could substitute for more detailed visual memory and might be more easily recalled at recognition. Second, much of the content of the typical picture set is composed of people, animals, food, buildings, and all types of human artifacts. An argument could be made that this type of content is of profound interest to human observers, and as such, more memorable than any other visual content (which, by exclusion, would include only natural scenery).

5. How context is formed: unsupervised learning

“Context” is an abstraction that concerns the manner in which long-term visual memories affect and are affected by both online perception and other forms of memory. Underlying this abstraction are neurobiological mechanisms for laying down memories. Although this thesis concentrates on behavioral results that demonstrate how context is formed, we should consider how context may be implemented in the brain.

Contextual constraints are learned from the thousands of fixations an observer makes on the world every day (over 172,000 a day, at a rate of 3 fixations per second over 16 hours of waking time per day). Each fixation offers a glimpse of the *structure* of a scene; the various objects one recognizes and their spatial distribution within the scene (Edelman, 1999). In other words, a fixation gives the visual system input as to "what" is "where" in the visual field (Marr, 1982).

Scene structure is coded in the brain, in the later locations in the ventral visual stream, by the responses of ensembles of neurons that are roughly tuned to respond to complex shapes at particular locations in the visual field. These types of "what + where" individual cells and ensembles are common in the inferotemporal (IT) and prefrontal (PF) cortices (Kobatake, E. & Tanaka, K., 1994; Logothetis, Pauls, & Poggio, 1995; Op de Beeck, Wagemann, & Vogels, 2001; Rao, Rainer, & Miller, 1997). Their responses have been entrained by visual experience. However, the paradox for the visual system is that it must decide based on that experience when to dedicate an ensemble of neural responses to a given combination of object identities and spatial positions without any way of knowing in advance which combinations will recur. Computational models (Edelman, Intrator, & Jacobson, 2002; Fei-Fei, Fergus, & Perona, 2006; Fergus, Fei-Fei, Perona, & Zisserman, 2010; Li, Su, Lim, & Fei-Fei, 2010) have been shown to be capable of solving this problem through unsupervised learning over a relatively small number of trials. These models, some of which are biologically plausible, appear to indicate that an artificial or biological network can generate "what+where" units based solely on the statistics of the input it receives, without predetermined units.

If we allow that any two or more instances of input to the visual system can never be exactly the same (as is indeed the case in free natural viewing), it becomes clear that unsupervised learning in the case of unit allocation is the process of deciding how much difference in visual input is required in order to dedicate a new unit or ensemble of whatever size. At a more abstract level, the formation of visual context is the process of determining how much difference can exist among the identity attributes and positions of objects and surfaces that comprise scenes while remaining within a given context.

When the ensemble of neural responses to a set of varying inputs is substantially the same, the response is called *invariant*. Visual invariance in all its forms has been well-studied with regard to objects, but less often with regard to entire scenes and the complex interdependence among object identities and positions that give us overall scene understanding (Chua & Chun, 2003). Visual context, then, might be thought of as *whole-scene invariance*. Which of these interdependencies can vary, and how much, before our understanding of a given scene changes?

We might assume that during context formation, every presentation of visual input is treated equally by the learning network, but we do not have evidence that this is true. Such an assumption would imply that everything that can be extracted from an example is always extracted. While an entire image can be scanned by an algorithm (Fei-Fei, Fergus, & Perona, 2006; Fergus, Fei-Fei, Perona, & Zisserman, 2010; Li, Su, Lim, & Fei-Fei, 2010), in learning human visual networks, this assumption is unlikely to be true. Allocation of eye movements and attention can be influenced by both low-level aspects of the scene (Itti & Koch, 2000; Judd, Durand, & Torralba, 2011) and object- or scene-level aspects (Brockmole & Henderson, 2008; Henderson, 2003; Neider & Zelinsky, 2006), meaning that some areas of scenes are probably not examined or encoded in a presentation. We need, therefore, to understand more about how scene-invariant responses can be elicited. Does a whole-scene response depend on all the objects and surfaces being enough alike and occupying spatial positions near enough to match previous learned examples? It would be helpful for vision scientists to know if the visual system responds differently to a scene based on too great a variation in object identity alone, or too great a variation in position alone.

6. Present studies on context formation and visual long-term memory

This introduction to some issues surrounding visual context indicates that three aspects need additional research. First, the presumed importance of spatial layout in recognizing contexts implies that we should examine how easily and precisely spatial constraints are formed. Particularly because we may identify scene category by layouts of large blobs of coarse spatial frequencies (Schyns & Oliva, 1994; Oliva & Schyns, 2000), it would be desirable to find out how great a degree of spatial displacement is required to disrupt those blobs, causing a spatial violation effect. It would also lend weight to theories of statistical learning if spatial context can be learned without additional semantic knowledge about the objects or overall spatial layout. Chapter 2 focuses on the learning of such spatial constraints from a baseline using novel objects.

Second, there is a need for research on identity context that bridges between studies on real-world scenes with pre-existing contextual constraints and laboratory studies using novel stimulus arrays. The difficulty of identity context formation in the latter studies and the weakness of the effects when identity context does form are at odds with the strong effects found in the natural scene studies. There must be reasons why novel objects do not readily form contextual constraints as strong as those formed among objects in the real world. As with spatial context, we should examine whether semantic knowledge is necessary to form identity contexts. In Chapter 3, we present experimental work that examines some of the factors affecting an observer's ability to learn identity co-occurrence in novel settings.

Third, the field's inferences about the effects of context on VLTM should be explored. Since the beginnings of modern psychology, researchers have generally agreed that the brain does not store a veridical replica of all past experience that can be recalled like a video recording (Bartlett, 1932; Neisser, 1982). The vast body of literature on memory shows that this agreement has usually been supported by evidence of what aspects of experience are retained in long-term memory, as opposed to quotidian aspects that are discarded. However, it might also be useful to know what conditions, such as over-learned visual contexts, militate against the veridical representation of past experience. The experiments in Chapter 4 test assumptions about the nature of long-term visual memory representations when they are composed of strongly contextualized scenes.

CHAPTER 2
POSITION SENSITIVITY IN THE FORMATION OF SPATIAL CONTEXT

1. Introduction

The spatial layout of a scene provides strong cues about the nature of the environment. Along with the identity of objects and surfaces that make up the layout (the “whats”), the positions of those entities in space (the “wheres”) comprise the structure of a scene. Most objects and surfaces we encounter in the natural world are constrained in some ways as to where they may appear in that structure. Some of these constraints are the outcome of animal and human behaviors. A bird’s nest, for example, might appear in the top of a tree or in a tuft of grass on the ground. Its location within the scene, along with other attributes like the nest’s size and composition, is part of the context that indicates whether the nest is a hawk’s in the treetop or a prairie chicken’s on the ground. However, other constraints are more general results of obedience to physical laws; most animals and objects rest on the ground plane, rather than hovering in the air, for example (Biederman, Mezzanote, & Rabinowitz, 1982). Surface features that make up the natural landscape are likewise constrained. Mountains rise up from the ground; they never protrude horizontally across a scene. Thus, the spatial layout of a scene provides human observers with confirmation of their fundamental understanding of the physical laws all bodies in space obey. In addition, the very repetition of such visual elements as the ground plane in a scene may help provide us with the feeling of a stable physical world in which we can operate.

These contextual constraints on scene layout are undoubtedly learned through many encounters with natural and human-made environments, such that the range of spatial locations in which a given object occurs is encoded along with its identity (Li & DiCarlo, 2008). It is also possible that the spatial relationships among objects that co-occur are encoded; for example, if a desk and a computer appear in the same scene, the computer will almost always appear on (above) the desk, not below it (Conway, Goldstone, & Christiansen, 2007; Fiser & Aslin, 2005; Jiang, Olson, & Chun, 2000).

2. Violations of spatial context

Early influential work in scene processing has found that observers are aware of violations of spatial context within a single glance at a scene. There is some disagreement about how such responses arise; on one hand, they have been taken as evidence that a high-level conceptual schema or frame is activated within the first few hundred milliseconds of scene processing that provides information not only about the objects that are likely to occur in a scene, but also their permissible spatial relations with other objects and with the scene area itself (Biederman, 1981; Biederman, Mezzanotte, & Rabinowitz, 1982). On the other hand, some argue that the results could be better explained by assuming that they begin with bottom-up perceptual processes that are validated by higher-level hypotheses or scene schemas at later stages of processing (De Graef, Christiaens, & d'Ydewalle, 1990; Lleras, Rensink,& Enns, 2005). In fact, a common account of scene processing in recent vision research holds that low spatial frequency filters in the brain's early visual areas segment the scene into a coarse layout of colored blobs and boundaries that facilitates the rapid identification of the scene's main content (or "gist" or "category") (McCotter, Gosselin, Sowden, & Schyns, 2005; Oliva & Schyns, 2000; Schyns & Oliva, 1994). This initial categorization is sufficient to prime responses to the component objects and surfaces in the scene, although it does not contain the high spatial frequency information that permits full object recognition (Oliva & Torralba, 2001; Torralba & Oliva, 2003).

Similarly, typical spatial layouts may be learned as top-down schemata. Information about how many objects are present and their spatial relationships to one another, as well as the extent of the ground plane and the depth of the picture plane may be encoded (Sanocki, 2003). Such spatial layouts may be somewhat resistant to scene translation, where the pictorial viewpoint is varied (Sanocki & Epstein, 1997), but the degree of permissible object shifts has not been adequately explored. Regardless of the mechanism by which the spatial context is evaluated, however, violations of spatial context almost always influence participants' responses to scenes in experimental paradigms.

3. Position invariance

The ability to perfectly detect objects' spatial violations is, interestingly, not borne out in the research on object recognition. Both psychophysical and neurobiological investigations have indicated that, in general, mental representations of objects are not position-invariant. Translation usually results in longer RTs and lower accuracy in matching tasks (Dill & Edelman, 2001; Dill & Fahle, 1998; Newell, Sheppard, Edelman, & Shapiro, 2005). Neural imaging techniques used in some of these studies show reduced responses to translated objects (Kravitz, Kriegeskorte, & Baker, 2010; Kravitz, Vinson, & Baker, 2008).

A recent study by Juttner & Rentschler (2008), however, indicates that experience with visual patterns (i.e., learning to classify patterns by type) leads to a greater degree of translation invariance over shifts of about +/− 3°. This finding suggests that it is also important to consider how object recognition after translation differs from determining whether an object is in an unaccustomed position. The latter represents some additional amount of visual processing, beyond recognition itself, and whether it takes place concurrently with or after recognition may depend on forming higher-order representations of object attributes, such as the implicit learning of permissible locations.

4. Memory for object positions and configurations

Memory for particular instances of visual structure has proven to be relatively easy to modify through experience. In the experimental paradigm pioneered by Chun and his colleagues, for example, it has been found that repeating a particular spatial array several times throughout a sequence enables participants to detect when an object is missing from the configuration, although this effect only applies to one or two targets in the array (Brady & Chun, 2007; Chun & Jiang, 1998; Chun & Jiang, 2003; Olson, Jiang, & Moore, 2005).

The context provided by learned spatial configurations may also include object-to-object structural relationships. Jiang, Olson, & Chun (2000) presented stimulus arrays where they changed not only target probes but also the identity and location of distractor objects in a change-detection task. Such configurational disruptions prevented participants from encoding the position of the target probe. The assumption of this kind of work is that all the items in the visual array are coded in relationship to each other, organized first by spatial configuration, then by the identity of the objects at the encoded locations. The primacy of configural information is borne out by other findings; for example, Hollingworth (2006) found that participants were able to remember the locations of multiple objects in familiar real-world scenes, where participants would naturally have had considerable experience with the possible range of locations where the objects might appear. Recognition performance was better when the object was re-presented in the context of the entire scene than in a small patch of scene background. Significantly better results were also obtained for recognition of objects that were presented in the same spatial position in the test scene as in the sample scene. Previously, Hollingworth (2005) had found that subjects were significantly better able to remember the spatial position of objects that they had actually seen in the sample scene than to guess where the objects were likely to be in the context of a scene that did not actually contain the target. This converging body of findings indicates that additional research is needed on how spatial contexts are formed and how much invariance, or tolerance, for object locations should be expected.

5. Experiment 1

Previous research shows that spatial configuration or layout is easy to learn; but no research has yet investigated how easy spatial *constraints* are to learn. It would be particularly valuable if we could show that scene structure that violates newly learned spatial constraints produces the same kind of responses that earlier work on real-world photographs and line drawings does. It should be noted, however, that we are interested in spatial constraints not as a form of viewpoint dependence, but as a distinct item of information that could aid in scene understanding and spatial context formation.

In this experiment, we tested how implicit learning of a novel scene structure where some objects could occupy a range of locations could induce stronger responses when the objects subsequently appeared outside of the implied spatial constraint. Two objects, each with a bounded area where it could appear, were simultaneously presented during training in the experiment. The first bounded area was relatively large, occupying the lower left quadrant of the scene area. The second bounded area was a smaller square area in the upper right quadrant, occupying about one-eighth of the scene area.

Note that in the following description of this experiment and throughout this thesis, for brevity we use the terms *translated* and *moved* for the situation in which an object in an original sample scene appears in a different position in a later test scene. We do not mean to imply that the object itself moves during scene presentation.

5.1 Participants

Eleven Cornell University undergraduates participated in the first experiment. The experiment lasted for about 90 minutes. All participants had normal or corrected-to-normal vision. They received course credit for their participation. All participants gave their informed consent.

5.2 Stimuli

Artificial scenes were constructed using grayscale photographs of diatoms (microscopic unicellular forms of algae with silicon skeletons) randomly placed on a white background. Each scene consisted of five diatoms.

The scenes were displayed on a 17-in. (43.2 cm) Dell computer monitor with a resolution of 1024 x 768 pixels. The monitor was viewed from a distance of approximately 25 in. (63.5 cm) in a normally lighted office. Each diatom had maximum dimensions of 100 x 100 pixels and subtended approximately 2.9º of visual angle vertically and horizontally, while the entire scene subtended approximately 24.3º of height and 32.3º of width. The experiment was controlled by MATLAB 7.0 with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997).

Three of the five diatoms were potential memory probes. Two probe diatoms were constrained to appear in the bounded areas of the screen during the training phase; each probe was confined to its own bounded area. Probe 1 was confined to an area occupying the lower left quadrant of the scene area, which subtended approximately 14.67º horizontally and approximately 11.00º vertically. Probe 2 could occupy a smaller region measuring 300 x 300 pixels in the upper right quadrant, subtending approximately 8.6º horizontally and vertically. The objects could be translated only within their individual bounds during training. A third diatom was chosen as the control object; it could appear and be translated anywhere within the scene, including within the boundaries for the spatially constrained objects. The other two objects were randomly positioned in the scene area, including within the bounded areas; however, they never overlapped with the other objects.

Scenes were presented for 8 s, followed by a 4-s mask consisting of a medium gray screen.

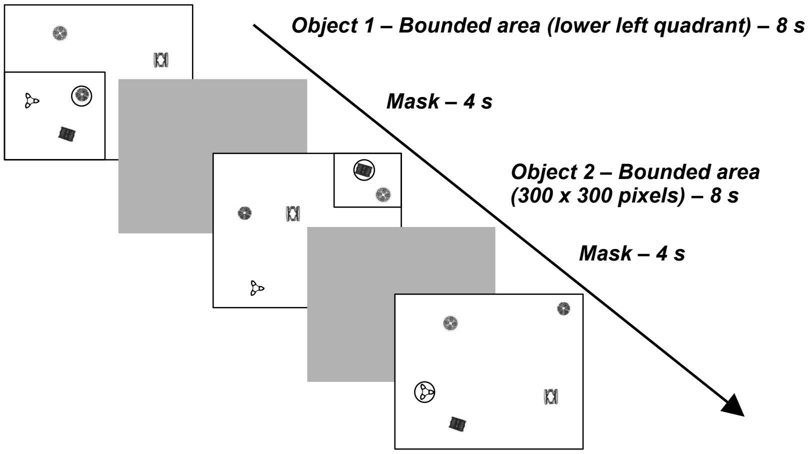


Figure 2-1: Sample scenes showing object boundaries from Experiment 1.

In the training phase of the experiment, a probe object that was translated could only appear inside its bounded area from sample to test. In the testing phase, the probe object could appear inside or outside its bounded area. A control probe object could appear anywhere randomly within the scene. The boundaries are shown for illustrative purposes only, and the actual order of scenes containing the probe objects was randomized, not the order shown here.

5.3 Procedure

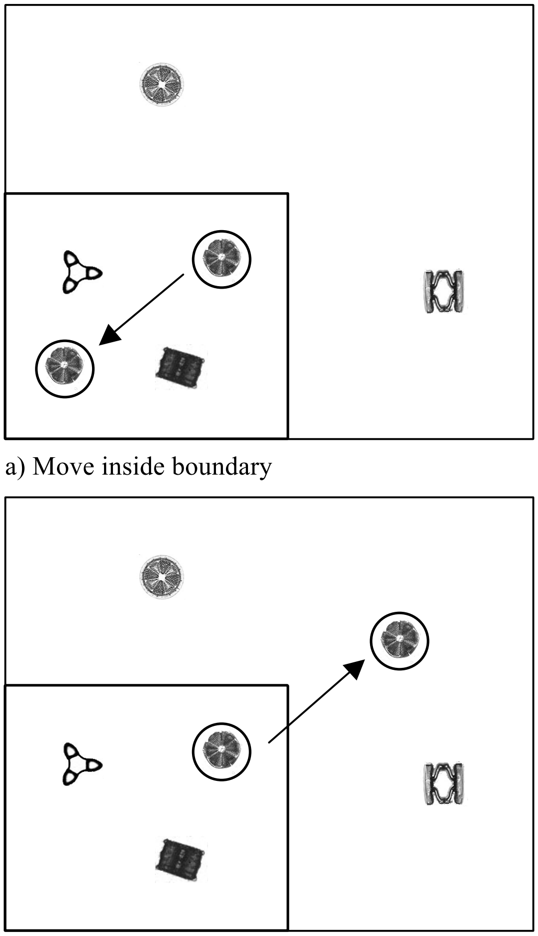
The experiment consisted of 360 trials with a delayed match to sample task. The scenes were presented with one intervening scene between each sample and test pair, so that each intervening scene became the sample for the next test scene. Thus, participants always had to remember at least one additional scene before they responded to a test scene. In test scenes, the probe diatom was circled in red.

The response measure was a forced choice to the question, "have you seen *that* object in *that* location?" The probe question was deliberately unclear about the number of previous scenes that the participant was asked to remember. The ambiguity was intended to force participants to rely on their implicit learning of the contextual regularities being presented rather than their episodic memory for previously presented scenes.

Participants read instructions for the experiment on screen, then completed a short practice session before beginning the main experiment. The practice session included one trial of each of the six conditions. Participants indicated their yes-no response with a keypress. They were permitted to view the test scene as long as they wanted before responding.

The experiment began with a training phase consisting of 180 scenes. The two constrained objects as well as a third, unconstrained object, served equally often as probes. In the training phase, the constrained objects were translated only within their bounded areas from sample scene to test scene. The third, unconstrained object could be translated in any direction; scenes were structured so that these moves were of approximately the same physical extent as the smallest and greatest distances the two constrained objects could cover. The scenes were equally divided into 'move' and 'no-move' trials.

A brief rest period followed the training, after which participants were told to continue with the experiment. In the test phase, the probe objects could be translated outside their bounded areas, although participants were not informed of this fact. The test phase consisted of an additional 180 trials, also divided equally between ‘move’ and ‘no-move’ trials. Within ‘move’ trials, there were equal numbers of moves within bounds and outside bounds for each probe object. (The unconstrained probe’s moves were equally often of smaller and larger extent.)



a) Move inside boundary

b) Move outside boundary

Figures 2-2a and 2-2b: Sample scenes showing moves inside (a) or outside (b) object boundaries from Experiment 1.

(Boundaries and arrows are shown for illustrative purposes only.)

5.4 Results

We calculated *d′* for each condition in the experiment. If a participant answered “no” to the probe question “Have you seen *that* object in *that* location before?” and the trial was a move trial, the response was counted as a hit. A “yes” answer to the question for the same condition was counted as a miss. A “no” answer to a no-move trial was thus counted as a false alarm, and a “yes” answer as a correct rejection. The correctness of answers referred to the move/no-move condition of the one-back sample scene. Higher *d′*s indicate increases in participants’ sensitivity to changes in the scene between sample and test scenes. The *d’* values were submitted to a two-way repeated-measures analysis of variance, with move type (*inside boundary, outside boundary*) and object training (*object 1 [larger area], object 2 [smaller area], object 3 [no constraint]*) as factors.

When the two probe objects were translated within their trained boundaries, participants remembered their locations quite poorly. However, when these objects were translated outside their boundaries, participants were much more sensitive to the changed locations. The main effect of move type was statistically significant (F (1, 10) = 34.63, $p < 0.0001$, $d = 3.72$). The effect was due to the training for spatial context induced in the first phase of the experiment, as shown by the main effect of object training (F (2, 9) = 11.97, $p < 0.003$). The unconstrained control probe (object 3), which was translated to an equal extent, did not evoke the same sensitivity to its changed position as the constrained probes (objects 1 and 2). There was no interaction between move type and object training.

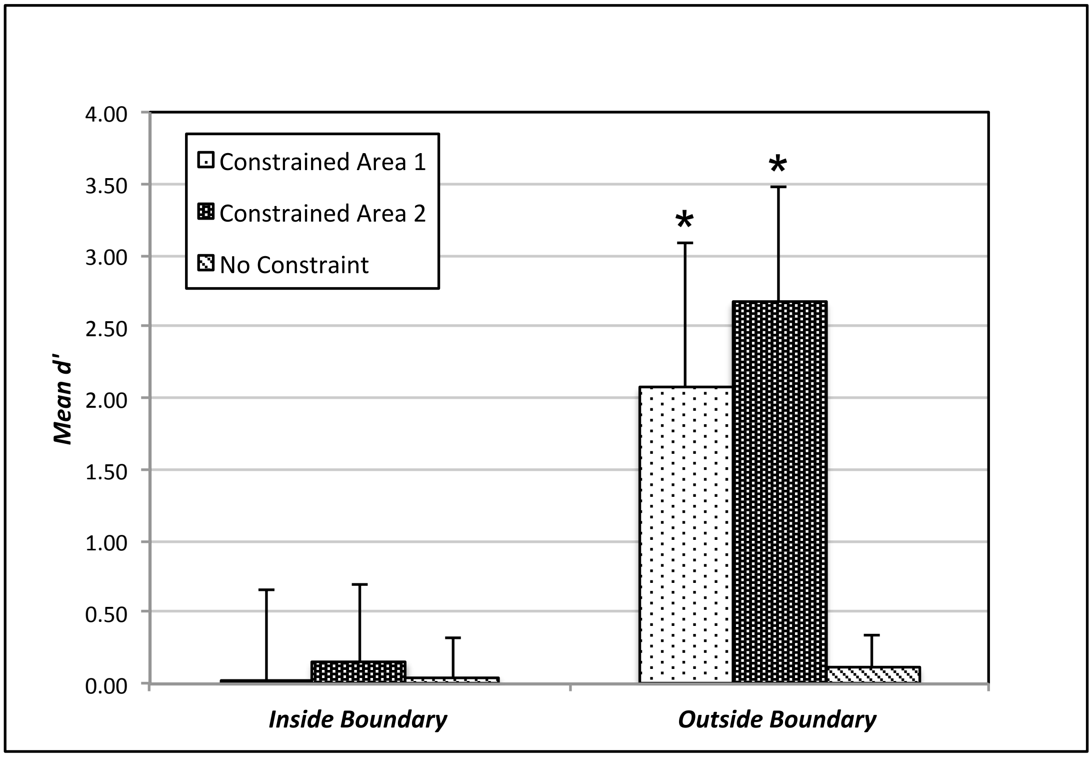


Figure 2-3: Performance results from Experiment 1.

After training for spatial constraint, participants were more sensitive (had higher *d'*s) to translations outside a trained object's bounded area. For the object with no training (object 3), there was no sensitivity difference for moves of approximately the same physical extent as the smallest and largest moves the trained objects could make. Asterisks indicate statistically significant effects. Error bars represent 95% confidence interval for the mean.

5.5 Discussion

The results of this experiment show that observers can quickly learn a spatial context for novel objects. The increases in sensitivity to violations of that spatial context were marked. This is especially notable because, in this experimental paradigm, participants were required to remember the structure of scenes over several seconds, in a delayed match to sample task that included an intervening re-structuring of the same objects. Remembering the spatial position of any of the objects was apparently very difficult, as shown by the low *d'* values for the untrained object, as well as the values for each of the trained objects when they were translated within their bounded areas. With a training period where participants saw the objects in their context only 30 times per object, however, a spatial context was induced that sharply improved memory for these objects' previous positions. The effect of spatial violations was not due to the physical extent of the translation. For a set of inside-boundary and outside-boundary translations matched for physical extent (pooled for both bounded areas), a paired-samples t-test showed a highly statistically significant difference in performance ($t(7) = -9.9$, $p < 0.0001$).

6. Experiment 2

Having demonstrated that training for expected location establishes a spatial context for relatively large areas of the visual field, we next focused on investigating whether observers could learn spatial contexts with more subtle restrictions. This experiment examined observers' ability to learn contexts for two objects, one whose changes were confined to vertical translations and one whose changes were confined to horizontal translations. The motivation for investigating this type of spatial constraint was based on human observers' better performance on perceptual tasks involving stimuli that are oriented horizontally or vertically than to oblique stimuli—the "oblique effect" (Appelle, 1972) and on the known preponderance of contours with vertical and horizontal orientations in real-world scenes, both indoor and outdoor (Coppola, Purves, McCoy, & Purves, 1998), as well as the fact that there appear to be more cells in V1 that respond to horizontal and vertical orientations than to other orientations (Li, Peterson, & Freeman, 2003).

Additionally we were motivated by the notion that observers could potentially be more sensitive to object translations in the vertical dimension than the horizontal because in natural visual experience, objects more often move about on the ground plane than move vertically (excluding changes in retinal position caused by the observer's eye, head, and body movements). Psychophysical evidence using natural scenes, as opposed to gratings, has shown poorer responses to scene patches with horizontal orientations than vertical or oblique orientations—a “horizontal effect” (Hansen & Essock, 2004).

6.1 Participants

Twelve participants took part in the second experiment, which included both training and testing phases, for about 40 minutes. All participants had normal or corrected-to-normal vision. They received course credit for their participation. All participants gave their informed consent.

6.2 Stimuli

The scenes in Experiment 2 were composed of five objects on a white background as in Experiment 1. The objects were abstract shapes created with an algorithm designed by Op de Beeck, Wageman, & Vogels (2001). These objects were presented on a white square 600 x 600 pixels centered on the computer screen, with a medium gray surround. Each object subtended approximately 1.4° of visual angle, and had maximum dimensions of 50 x 50 pixels. The physical apparatus and computer monitor were the same as for Experiment 1.

In this experiment, participants were to learn a spatial context for one object that was constrained to appear in an area comprising a vertical “stripe” on the right side of the scene area, with coordinates x1=400, x2=450, y1=50, y2=550 within that square. This stripe subtended 14.3° vertically and 1.4° horizontally. A second object was constrained to appear in a horizontal stripe in the lower half of the scene area, with coordinates x1=50, x2=550, y1=400, y2=450. A third object served as a control object; its initial positions were not constrained. The other two objects were randomly positioned in the scene area; they never overlapped with the other objects, although they were permitted to appear in either the vertical or horizontal bounded stripes.

Scenes were presented for 2 s, followed by a 1-s mask consisting of a scrambled photograph of a natural scene. A medium gray screen appeared for 0.5 s before the start of a new trial.

6.3 Procedure

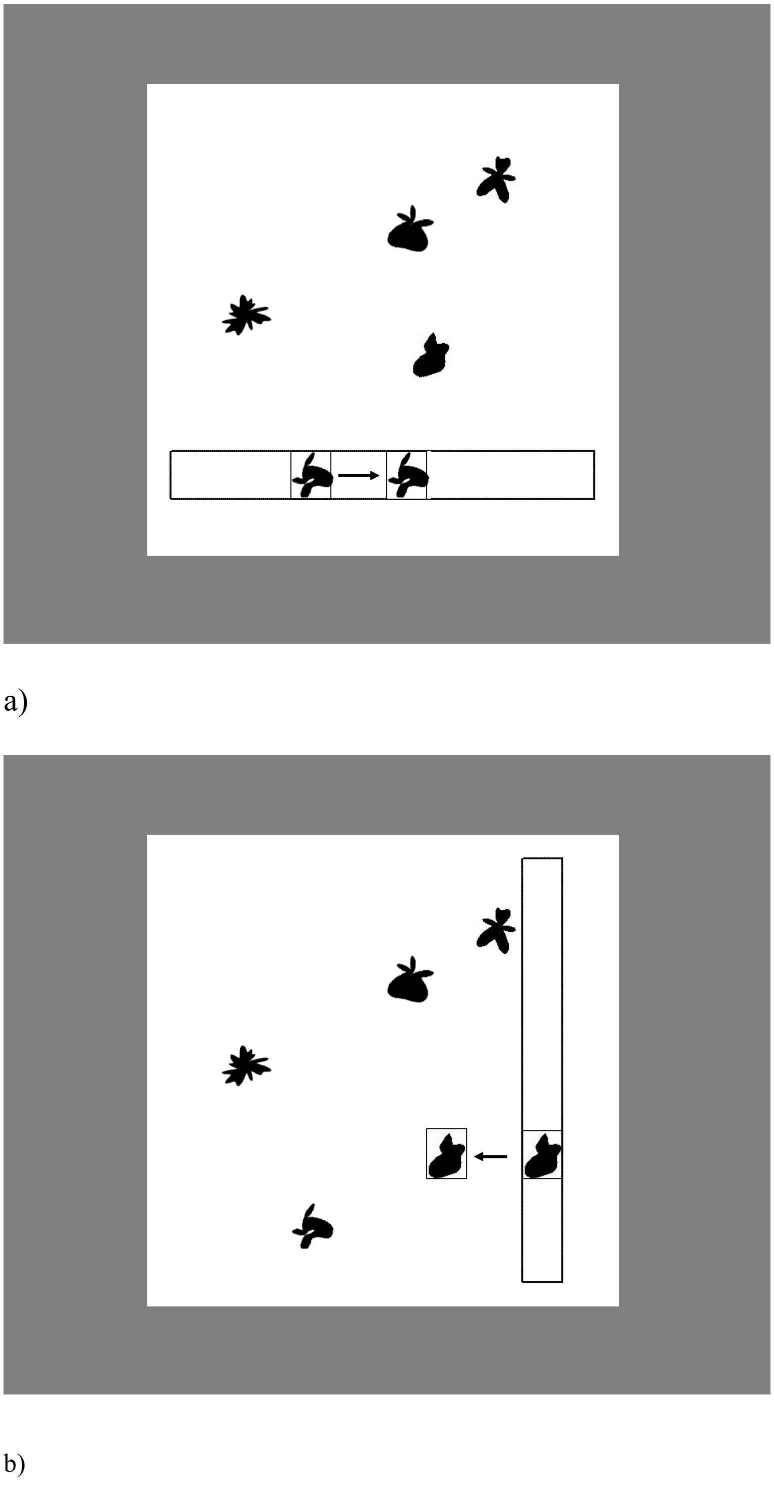
The conduct of this experiment was similar to that of Experiment 1. It began with instructions read on the computer screen, then a short practice session consisting of one trial of each of the 12 conditions. The training phase followed. Participants were unaware of the training, as they responded to each test scene in the same way as for Experiment 1.

In the training phase, the constrained objects were translated only within their individual bounded areas. In 50% of the scenes, the probe object subsequently appeared in a different position from sample to test. The third, unconstrained object could also be translated either vertically or horizontally, but its original position was not constrained to any portion of the scene area. The translated positions for the constrained objects were chosen from a uniform distribution of random positions anywhere within the 500-pixel constrained vertical or horizontal stripes. The extent of the translation, therefore, was up to 450 pixels (so that the 50-pixel object was always confined within the 500-pixel stripe). The vertical and horizontal translations for the unconstrained object were chosen from a uniform distribution of random positions up to 450 pixels from the original position. The training phase consisted of 192 scenes, distributed as shown in Table 1 below:

Table 1: Distribution of trials for the training phase of Experiment 2.

	Vertical Object	**Horizontal Object**	**Unconstrained Object**			
	Vertical Move	*Horizontal Move*	*Horizontal Move*	*Vertical Move*		
moves	32	32	16	16	96	Total Moves
no move	32	32	32		96	Total No Moves
					192	***Total Training Trials***

In the testing phase, the constrained objects could be translated outside their bounded areas. In 50% of the trials, the object was translated from sample to test in one of two ways: (1) It was translated vertically (or horizontally) within its bounded area either 50 pixels (*in-context–small moves*) or 100 pixels (*in-context–large moves*); or (2) it was translate either 50 pixels outside the bounded area (*out-of-context–small moves*) or 100 pixels (*out of context–large moves*). The in-context moves were equally balanced between up and down moves for the vertical object and left and right moves for the horizontal object. The out-of-context moves were equally balanced on the left and right sides of the bounded area for the vertical object, and above and below the bounded area for the horizontal object.



a)

b)

Figures 2-4a and 2-4b: Sample scenes showing translations inside (a) or outside (b) the horizontal and vertical object boundaries from Experiment 2.

(Boundaries and arrows are shown for illustrative purposes only.)

For comparison, in 50% of the trials for the unconstrained control object, it was translated either 50 or 100 pixels; equal numbers of trials contained displacements in vertical and horizontal directions. The testing phase consisted of 144 trials, as shown in Table 2 below:

Table 2: Distribution of trials for the testing phase of Experiment 2.

	Vertical Object		**Horizontal Object**		**Unconstrained Object**			
	Vertical Move	*Horizontal Move*	*Horizontal Move*	*Vertical Move*	*Horizontal Move*	*Vertical Move*		
small moves	6	6	6	6	6	6	48	*Total Small Moves*
large moves	6	6	6	6	6	6	48	*Total Large Moves*
no moves	24		24		24		48	*Total No Moves*
							144	***Total Testing Trials***

6.4 Results

We calculated *d´* for each condition in the testing phase in the same manner as for Experiment 1. A repeated-measures analysis of variance was performed on these data, with object training (*object 1* [*vertically trained], object 2 [horizontally trained], object 3 [unconstrained]*); context (*in-context, out-of-context*); and move size (*small [50-pixel]*, *large [100-pixel*]) as factors.

There were statistically significant main effects of both context (*F* (1, 126) = 23.11, $p < 0.0001$, $d = 0.85$) and move size (*F* (1, 126) = 6.06, $p = 0.015$, $d = 0.43$). The interaction of context with move size also approached significance (*F* (1, 126) = 2.52, $p = 0.085$).

After training, participants were much more sensitive to changes in position when objects were translated *outside* their trained spatial contexts than when they were translated *within* their contexts. This effect was especially notable for the large (100-pixel) move size. When objects made large moves out of context, sensitivity improved markedly for both objects (mean *d′* for the vertically trained object = 1.22; for the horizontally trained object = 1.25). Even small moves out of context improved performance (mean *d′* for the vertically trained object = 0.52, for the horizontally trained object = 0.36). Sensitivity for large moves within the trained context, however, was no better than chance for both the vertically trained object (mean *d′* = -0.07) and the horizontally trained object (mean *d′* = 0.00). Small moves (50 pixels) within the trained spatial contexts produced equally poor performance for both the vertically and horizontally trained objects (both, mean *d'* = -0.03).

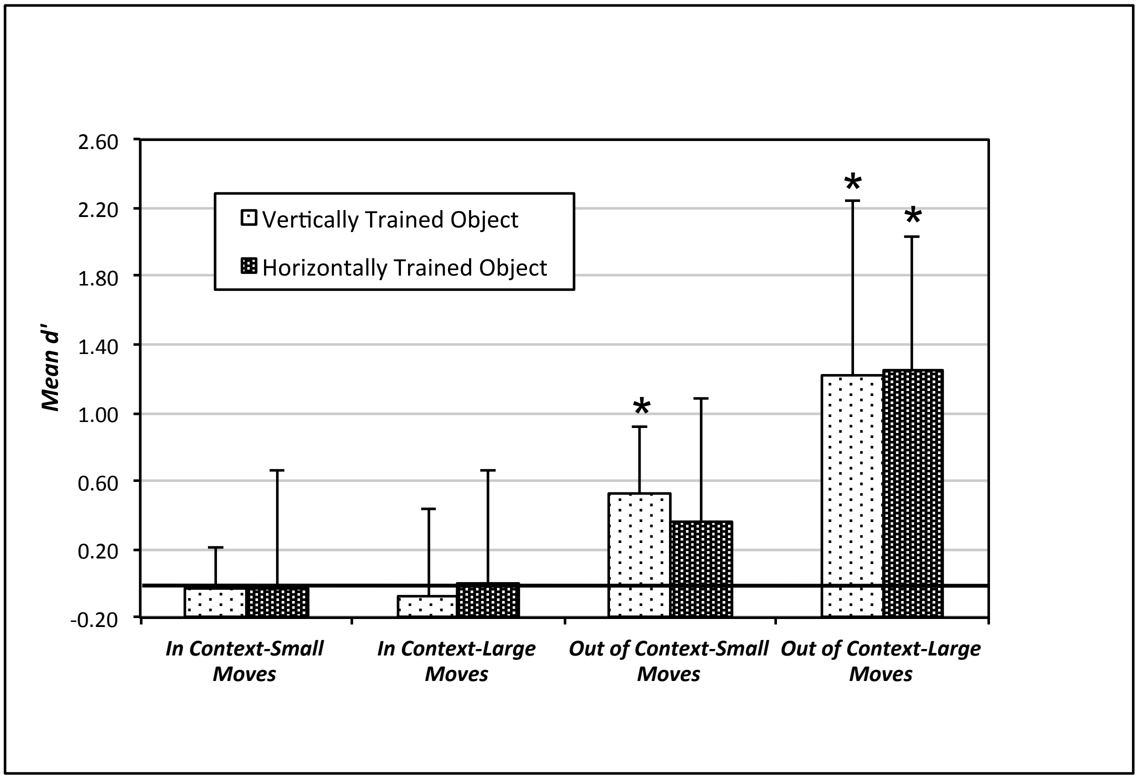


Figure 2-5: Performance results for the trained objects from Experiment 2.

After training for spatial constraint, participants were more sensitive to translations outside an object's spatial constraints. Asterisks indicate statistically significant results. Error bars represent the 95% confidence interval around the mean.

By contrast, sensitivity to the translations undergone by the untrained object was uniformly poor across all the conditions. This object had no spatial constraints but could be translated vertically or horizontally in equal extents to the trained objects. Only large translations in the horizontal plane produced a *d'* greater than chance ($t(11) = 2.39$, $p = 0.036$).

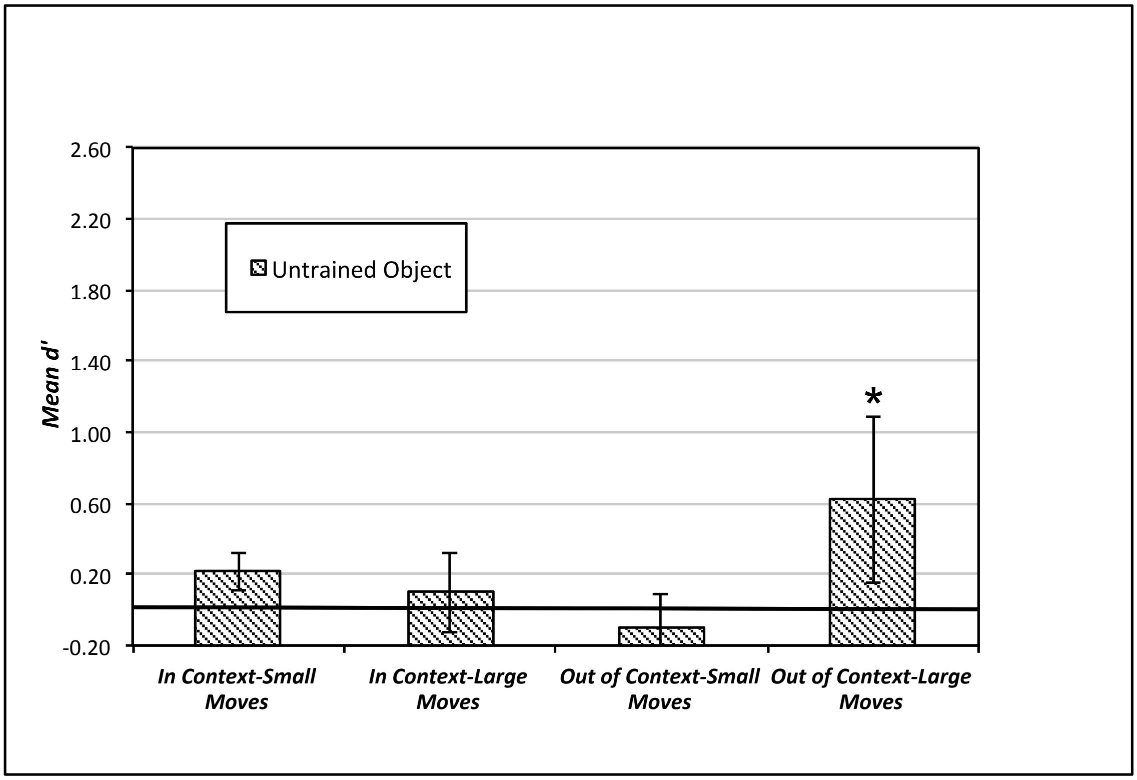


Figure 2-6: Performance results for the untrained object from Experiment 2.

This object's translations were of the same small and large extents in the vertical and horizontal dimensions as the trained objects, but the translations were not confined to a spatial context. Sensitivity was no better than chance to any translations except large moves in the horizontal dimension. Error bars represent the 95% confidence interval around the mean.

The pattern of results for all three probe objects was similar: best for large moves, especially out-of-context moves in the case of the spatially constrained objects, and worst for the small moves. There was no main effect of the type of object training (vertical, horizontal, or untrained) ($F(1, 126) = 0.01$, $p = 0.931$), and no interactions.

Likewise, there was no statistically significant difference between performance for the vertically trained object's small moves and the horizontally trained object's small moves ($t(22) = 0.44$, $p = 0.686$). A separate test of performance for the vertically trained object's small moves was significantly better than chance ($t(11) = 2.89$, $p = 0.015$). When one considers that the small move size places this object just barely outside its spatial context, so that the object's side would almost touch the context's implicit borders, this sensitivity seems remarkable.

6.5 Discussion

The results of this experiment indicate that observers are capable of learning spatial contexts that are more restricted than those of our first experiment, where the spatial constraints were relatively large in area. Translations outside the stripes to which the probe objects were confined in this experiment were, we should note, perfectly predictable in the opposite dimension to the previously trained dimension. That is, the observer needed only to learn to note whether the object appeared above or below (or to the right or the left of) the implicit straight line formed by its previous presentations. Nevertheless, the translations within the stripes were less predictable, as the objects' initial positions could be anywhere within the stripe and the translations were displacements within the trained dimension from this unpredictable position.

We were interested to note that participants were slightly more sensitive to the small translations outside the spatial context of the vertically trained object than the horizontally trained object. Because the translation involved a change in the horizontal dimension to take the object *out* of its context, it is difficult to interpret this result as sensitivity to translations in the *vertical* dimension; it may be the result of greater attention to the vertically trained object's translations in general.

7. General Discussion

As far as we are able to tell, the results of these two experiments are the first to reproduce in a controlled laboratory experiment the "position violations" effect that previous researchers have shown using familiar real-world scenes (Biederman, 1981; Biederman, Mezzanotte, & Rabinowitz, 1982). They show that observers readily learn spatial contexts that sensitize them to violations of objects' spatial relationships to the rest of the scene.

The experiments also support the notion that detecting position violations is probably a bottom-up process that does not require the activation of a conceptual scene schema (De Graef, Christiaens, & d'Ydewalle, 1990; Lleras, Rensink,& Enns, 2005). The statistical learning that takes place during the training phase depends only on the repeated encounters with a particular object in a range of positions in the scene, and not on knowledge about the function of the object or the physical constraints of the natural world.

The changes in sensitivity are not due solely to translations that cross visual hemifields or quadrants as the results of Experiment 2 show. Approximately equal sensitivities were produced to translations that stayed entirely in one hemifield as to translations that crossed the midline.

Sensitivity to deviations from the horizontal or vertical stripe probably partially share in the explanation for vernier hyperacuity, the phenomenon by which observers can detect offsets between two lines or edges that are much finer than the field of a single photoreceptor (Poggio, Fahle, & Edelman, 1992). The disproportionate number of V1 neurons that are selective for vertical and horizontal orientations (Li, Peterson, & Freeman, 2003) make it possible for a large ensemble of finely graded neural responses to represent very narrow gaps between lines. Although the neural representations of the positions of the objects in Experiment 2 are never all active at once, the training probably induces differential sensitivities of these ensembles to positions within the spatial contexts and those outside the context that make fine distinctions along the given dimension possible.

CHAPTER 3
VISUAL DISTINCTIVENESS IN THE FORMATION OF IDENTITY CONTEXT

1. Introduction

The experiments in Chapter 2 indicated that spatial contexts are, as we had reason to suppose, easily formed. We would also expect that visual experience of the natural world allows us to learn which objects and events tend to co-occur in predictable ways, and that this information serves as an even more powerful top-down cue in recognizing and interpreting scene contexts. For example, we learn that some kinds of objects occur more frequently in some settings than others—offices frequently contain books and computers, while garages contain lawnmowers and automobiles (Biederman, 1981; Mandler & Parker, 1976; Mandler & Ritchey, 1977). This statistical learning of the co-occurrence of objects in predictable sets results in the formation of what we are terming *identity contexts*.

The effects of identity contexts are readily seen in behavioral experiments using real-world objects in line drawings and photographs. Identity context enables observers to identify objects more rapidly when they are consistent with their surroundings than when they are inconsistent (Davenport, 2007; Davenport and Potter, 2004; Pedzek *et al.,* 1988, 1989). However, the effects of context also cause observers to attend to and spend longer noting areas of scenes that violate some expectancy formed by experiencing many instances of normal co-occurrence (Becker, Pashler, & Lubin, 2007; De Graef, Christiaens, & d'Ydewalle, 1990; Friedman, 1979; Hollingworth & Henderson, 1998; Loftus & Mackworth, 1978; Palmer, 1977). Contextual learning is often detailed enough to allow observers to judge when objects are inappropriately positioned or sized within a scene

(Biederman, Mezzanotte, & Rabinowitz, 1982), and to very rapidly decide whether scenes contain animals or other classes of objects (Rousselet, Joubert, & Fabre-Thorpe, 2005).

We can infer from such results that identity contexts are formed from many interactions with the natural world. Little attention, however, has been paid to identity context formation as a topic of interest in itself. As far as we are aware, there are no published laboratory experiments using those terms. However, there are studies under the heading of statistical learning that report learned association among collections of novel objects (Conway, Goldstone, & Christiansen, 2007; Fiser & Aslin, 2002; Fiser & Aslin, 2005). Likewise, the literature on visual working memory contains a number of studies where identity context is formed from a novel baseline.

In fact, much of the work on implicit learning from visual experience has actually been done on the formation of spatial context rather than identity context. One reason for this research gap may be that in free viewing, object identity and spatial position are inextricably confounded. But this confound replicates how identity context formation would actually occur in real-world observation; therefore, these results have considerable ecological validity for the purposes of this study. We find that much of the work on implicit learning from visual experience has shown that spatial contexts are more easily learned than the co-occurrence of object identities. It has been speculated that spatial context learning is actually independent of object identity, producing an overall invariance to scene changes such as object color, shape, luminance and other perceptual features (Jiang, Olson, & Chun, 2000). If this is so, it presents problems for the object inventory hypothesis by suggesting that the visual system's registration and tabulation of object occurrences may not always occur automatically. If object identity is sometimes ignored in favor of spatial layout in scene perception, we cannot count on statistical learning to form visual contexts as readily as we might have supposed.

But an important ecological observation has gone unaddressed in the controlled laboratory experiments on statistical learning of object co-occurrence. That is, the objects and surfaces present in any natural scene are usually quite varied in size, shape, color, texture, and other visual attributes. The inconsistent object effect in the real-world photographs of the previously cited studies usually occurs for objects that are quite different from each other, like the octopus and tractor used in Loftus & Mackworth's (1978) experiment, or the football player and priest in Davenport & Potter's (2004) study. In contrast, most working memory studies use novel objects collected or generated in some way that gives them a similar appearance or style, such as the abstract shapes used by Endo & Takeda (2004). These researchers have good reasons for controlling for size and visual similarity in their experiments. Yet extrapolating from their results to learning object co-occurrence from real world scenes is akin to presenting participants with many photographs of parking lots and concluding that object identity is difficult to learn because the participants did not register the different details among many similarly sized, shaped, and colored cars. There are fewer situations in life in which the objects present to view share attributes than situations in which the object assemblage is very heterogeneous in identity attributes.

A few studies have looked at statistical learning of co-occurrence of object identities in repeated spatial displays (Chun & Jiang, 1999; Endo & Takeda, 2004; Jiang & Song, 2005). Observers were capable of associating a particular collection of distractor objects with particular targets, resulting in faster reaction times, although the spatial configuration changed across repeated displays of the same distractor identities (Chun & Jiang, 1999). In this study, the collection of distractor object identities was the only cue to the target location. Jiang & Song (2005), in contrast, kept spatial configuration constant, but varied which of two object shapes could occupy the positions of the configuration. They found that observers disregarded the objects' identities when they were trained on one shape but the object shape was switched at test. Only when the training phase included both object shapes did observers show a reaction time advantage for shapes that matched the configurations they had learned.

Endo & Takeda (2004) have looked most carefully at the interaction of object identity and configuration. They constructed displays of novel objects where both types of contextual information were repeated within a single experiment: spatial configuration only, object identity only, and combined spatial configuration and object identity. In their first experiment, after training on displays that combined configuration and object identity, observers were then tested on displays of object identity only, spatial configuration only, and combined configuration and identity. Reaction times were faster for the two conditions that contained a repetition of the configuration, but not for the condition that contained only a repetition of object identities. Their second experiment trained observers to associate the object identities with the target location; here, observers showed faster reaction times to conditions that contained repetitions of the object identities, although the configuration changed on each repetition. In their third experiment, observers were trained and then tested on all three conditions. Because both identity and configuration had been predictive of target location, they found that reaction times were fastest for the combined condition, and somewhat faster for the condition that contained a repetition of object identities (but not configuration, in this case). Their final experiment decoupled configuration from identity; the repetition of configuration condition predicted the target object's identity but not its location, while the repetition of identity condition predicted the target object's location but not its identity. Here, because the only condition that contained a cue to the target object's location was that which contained the repetition of object identities, the effect of identity produced reliably faster reaction times but configuration did not. Endo and Takeda suggest that people can learn both types of contextual information, but are most likely to learn only whatever type is most useful in a given situation. In cases where both types are equally useful, the effects of configuration learning and identity learning appear to be additive.

The general theme of this kind of research is that the effects of configurational or spatial cuing predominate over the effects of object identity cuing, except in situations contrived to force statistical learning to be contingent on object identities. If this is true, it raises some interesting questions about how spatial configuration and object identity interact in viewing natural scenes. If it is indeed easier for observers to learn spatial configuration without attending to object identities, they may be more likely to confuse object identities within the configuration. Such errors might account for some examples of change blindness (Levin, Simons, Angelone, & Chabris, 2002; Simons & Levin, 1998).

The goal of the present study is to investigate whether the encoding of object identities along with spatial configuration can become more automatic if the novel objects in the scene are visually different from one another, as would be true in more realistic scenes. The objects in abstract arrays such as those used by Jiang & Song (2005), Endo & Takeda (2004) and Chun & Jiang (1999), as well as the experiments in this study, often bear a general resemblance to one another. They may be perfectly discriminable in shape, but they are almost always very similar in size, color, average spatial frequency, and other low-level visual attributes. It is possible that experiments using such similar objects bias participants toward reliance on spatial configuration at the expense of object identity exactly because the objects are easy to mistake for one another. We combined objects of very different visual appearance in one experiment to test whether more distinctive appearance results in better object identity encoding.

In addition, we investigate whether the advantage for configuration holds when the task is made less resource-intensive, and changed from visual search to memory for location. In Endo and Takeda's 2004 study, observers learned a relatively large number of configurations (28) composed of a relatively large number of object identities (91). Each display contained 10 of those objects in one of the repetition conditions. Although observers showed a statistically significant effect of object identities in the experiment where identity alone reliably cued the target location, they did not learn identities when that information was redundant with the configuration in cuing target location. It is possible that the burden imposed by learning the large number of both configurations and object identities caused observers to rely on the more robust information contained in the configuration. It might be instructive to make the task less memory-intensive by limiting the number of learned configurations and ensuring that each configuration is composed of a unique set of object identities.

The visual search task in the contextual cuing paradigm imposes its own effects: reduced reaction times for conditions where the identity or configuration of objects in the scene cues the location of a search item could be partially or wholly attributed to focused attention on the proper location. When search is initiated, the observer's memory might drive his or her attention to the region where the target appeared. But within that region, the search can be quickly completed because the target is actually present (on target-present trials, at least). Additionally, it is open to question how much of the configuration is actually remembered, because the entire configuration is re-presented at search.

However, if observers are asked instead to indicate the location of an object from a previously viewed display, they might benefit less from focused attention. Instead, their attention might spread out over locations near the remembered object, in the case where spatial memory is good, or over the entire display if they are having trouble remembering the location. Reaction times could be longer, but more information about the precision of both absolute location and the amount of configuration information retained in memory would be a welcome substitute.

2. Experiment 3

This experiment was designed to test the hypothesis that the objects in a context must differ visually from one another in some way in order to become salient to the construction of the context. To test this, we combined three different types of objects in one experiment. Each type of object appeared in three different spatial configurations. In the training phase, only objects of a particular type appeared in the three configurations. At test, however, the target object could be either from the same type (consistent) or from one of the other types (inconsistent). Thus the spatial configuration remained the same, but all possible combinations of other object types could appear as a target object within the configuration.

2.1 Participants

In this experiment, 14 Cornell University undergraduates took part for course credit. Their age range was 18 to 27 years. All participants had normal or corrected-to-normal vision. They received course credit for their participation. All participants gave their informed consent.

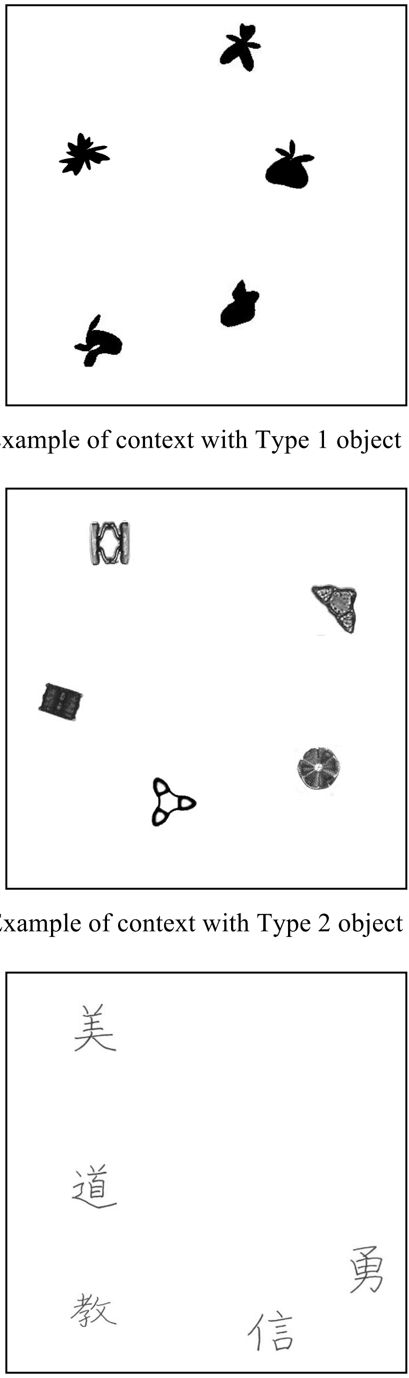
2.2 Stimuli

Artificial scenes were constructed using black shapes randomly placed on a white background. Each scene consisted of five objects. Each object was 50 x 50 pixels in size, subtending approximately 1.4° of visual angle.

The scenes were laid out by an algorithm created with MATLAB 7.0 with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). The objects were placed on a 600 x 600-pixel white square centered on the monitor screen and surrounded by a medium gray background extending to the edges of the screen. The generation algorithm ensured that the objects did not touch or overlap; they could appear within 4 pixels of each other, although in practice this did not occur.

We chose three types of objects as stimuli, and each type had multiple exemplars. The types were intentionally very different from each other: Type 1 comprised abstract objects generated by an algorithm created and used by Op de Beeck, Wagemans, & Vogel (2001). Type 2 comprised grayscale photographs of freshwater diatoms. Type 3 comprised characters from the Japanese *hiragana* syllabary. (All participants in this experiment were also screened to exclude those who spoke or read Japanese in any form.)

The object-type "families" are shown in Figures 3.1a, 3.1b, and 3.1c. Within each family, exemplars are clearly distinguishable from one another, yet they resemble each other more than they resemble exemplars from the other families. Each family has a distinctive visual style, or overall appearance.



a) Example of context with Type 1 object family – abstract objects created by algorithm

b) Example of context with Type 2 object family – freshwater diatoms

c) Example of context with Type 3 object family – *hiragana* characters

Figures 3-1a, 3-1b, and 3-1c: Examples of contexts with the three different object-type families.

The scenes were displayed on a 17-in. (43.2 cm) Dell computer monitor with a resolution of 1024 x 768 pixels. The monitor was viewed from a distance of approximately 25 in. (63.5 cm) in a normally lighted office, so that the entire viewing area from screen edge to edge subtended approximately 24.3º of height and 32.3º of width. The experimental protocol was controlled using MATLAB 7.0.

2.3 Procedure

After signing the informed consent and reading the experimental instructions on-screen, the participants began with a practice session consisting of 9 trials, including one representative trial each of the experimental conditions for each context (described below). The practice trials were not used in the actual experiment.

For each trial, the sample scene appeared for 2.0 s, followed by a 1.0-s mask consisting of a scrambled image of a natural scene, followed by the response screen (Figure 3.2). Participants were instructed to click with the computer mouse as close as possible to the center of the location where they believed the probe object had appeared in the sample scene. They were permitted to view the response screen as long as they wanted before responding. The trial presentation order was randomized for each participant.

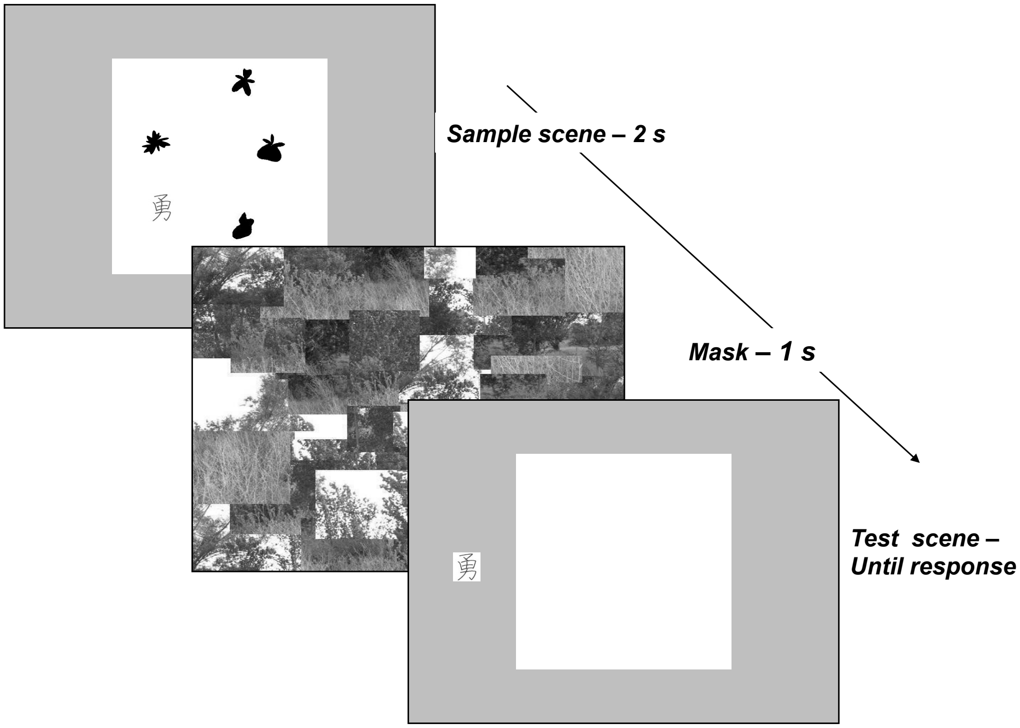


Figure 3-2: Trial sequence for Experiment 3.
Participants clicked on the remembered location of the probe object from the sample scene.

This experiment was blocked; there were 15 training blocks, with one trial each of the three different object types in one of three different spatial configurations for each object type. Therefore, each training block contained 9 trials, and the total training phase consisted of 135 trials. For each configuration within each object type, there were two possible probes chosen from the five objects in the configuration, while the other three obje cts could appear in any of the other positions in the configuration. Thus, the position of the probe was cued solely by the type of object and the overall configuration, yet participants could not rely on learning the absolute position of just one object per configuration. There were no distractor trials in this experiment.

In the training phase, the probe objects in the test scenes were always of the same object type as the other objects that made up the configuration. However, in the single 72-trial block of the transfer phase of the experiment, the same configurations as were learned in training appeared as sample scenes, while the objects were varied in the following combinations from sample to test:

- ***Type1 Objects/Type 1 Probe***: (12 trials)
- ***Type 1 Objects/Type 2 Probe***: (6 trials)
- ***Type 1 Objects/Type 3 Probe***: (6 trials)
- ***Type 2 Objects/Type 2 Probe***: (12 trials)
- ***Type 2 Objects/Type 1 Probe***: (6 trials)
- ***Type 2 Objects/Type 3 Probe***: (6 trials)
- ***Type 3 Objects/Type 3 Probe***: (12 trials)
- ***Type 3 Objects/Type 1 Probe***: (6 trials)
- ***Type 3 Objects/Type 2 Probe***: (6 trials)

Unlike many other experimental procedures of this kind, at test, the participants were presented with an entirely blank 600 x 600 pixel white square surrounded by a medium gray background as a response screen instead of the previous configuration with one object missing. A copy of the probe object was shown at the left side of the blank response square. The participants were asked to indicate the remembered location of the probe object within the blank square. This was done so that the presence of the distractor objects would not help cue the response.

Participants were not told that the configurations or the shapes that comprised them were being altered in the transfer phase. The entire experiment took approximately 25 minutes.

2.4 Results

The dependent measures were the spatial error (in pixels) between the center of the probe object in the sample scene and the coordinates of the mouse click the participant made on the response screen, and reaction time. Mixed-model regression analyses were performed on the spatial error and RT data for this experiment, with *participant* as a random factor and *object type* (*Type 1, Type 2, and Type 3*) and *consistency* (*consistent, inconsistent*) as fixed factors. Approximately 3% of the trials with RTs greater than two standard deviations from the mean were removed from these analyses.

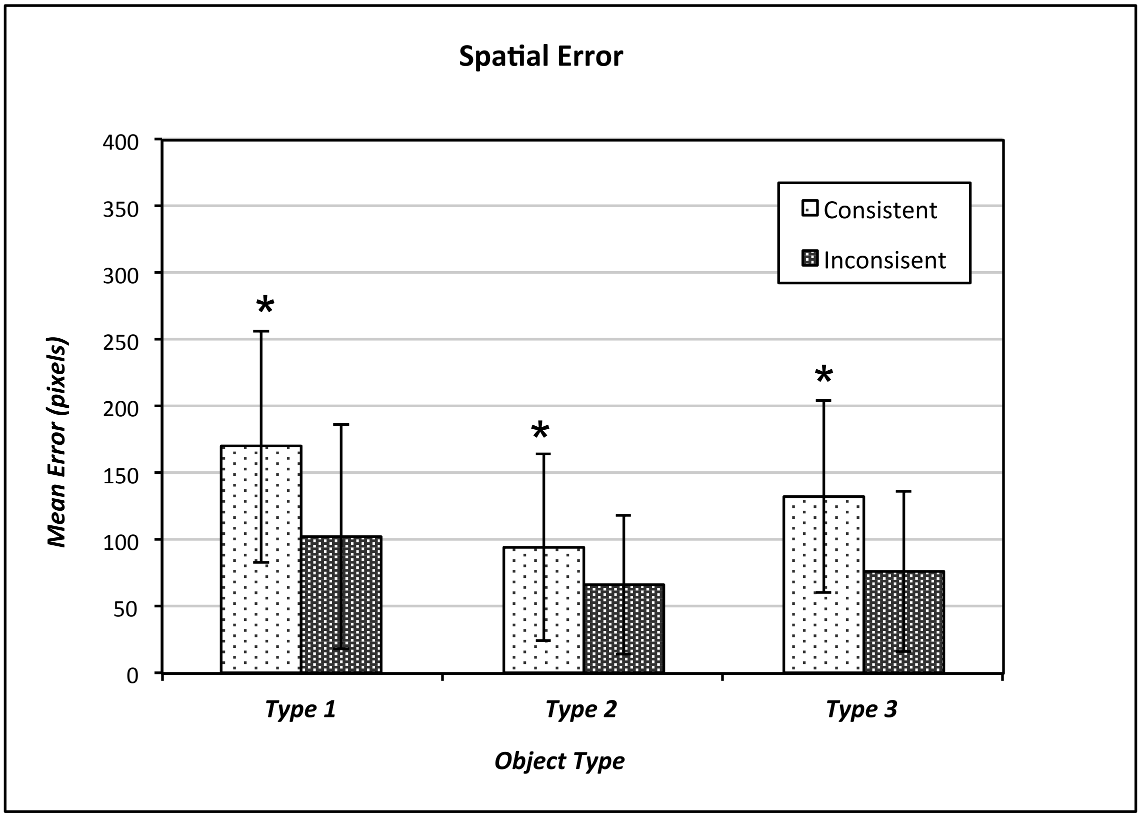


Figure 3-3: Spatial error data from Experiment 3.

Contextual consistency (probe objects appearing with other objects from the same object family) caused significantly greater spatial errors in remembering the probe object's position than contextual inconsistency (probe objects appearing with objects from dissimilar families). Asterisks indicate statistically significant results. Error bars represent the 95% confidence interval around the mean.

For the spatial error data, there was a strong main effect of consistency (*F* (1, 960) = 54.930, *p* < 0.0001, *d* = 0.48). When the probe object was inconsistent with the other type of objects with which it appeared, the participants remembered its location much more accurately than when it was of the same object type. There was also a strong main effect of object type; the Type 2 (diatom) family configurations were much easier for participants to remember than the other object families (*F* (2, 960) = 22.790, *p* < 0.0001). Individual comparisons of the contextual conditions were performed (all Fisher's LSD with Bonferroni correction applied for the number of comparisons made) that indicated the following:

When a probe object appeared with other objects that were of its same family (contextual consistency), participants made larger errors in remembering its position than when it appeared with objects of either dissimilar family (contextual inconsistency). This was true for Type 1 objects paired with Type 2 or with Type 3 objects (both comparisons *df* = 955 , *p* < 0.0001). Type 3 objects paired with either Type 1 or Type 2 objects were also significantly worse (both comparisons *df* = 955, *p* < 0.0001). Only Type 2 objects paired with either Type 1 or Type 3 objects were not statistically significantly worse (Type 2/Type1, *df* = 955, *p* = 1.000, Type 2/Type3, *df* = 955, *p* = 0.998), although the consistent pairing did have somewhat larger errors than both inconsistent pairings. It did not matter which of the inconsistent families a probe object appeared with: the error difference between Type 1/Type 2 and Type 1/Type 3 was not significant (*df* = 955, *p* = 1.000), nor were the error differences between Type 2/Type 1 and Type 2/Type 3 (*df* = 955, *p* = 1.000) or Type 3/Type 1 and Type 3/Type2 (*df* = 955, *p* = 1.000).

The reaction time data also showed a statistically significant main effect of object type (*F* (2, 960) = 5.575, *p* = 0.004). Performance was slightly faster overall for Type 2 objects and slowest overall for Type 1 objects.

The significant effect of contextual consistency was very clear: when the probe object was inconsistent with the other objects in the configuration, participants remembered its location more quickly than when it was consistent (*F* (1, 960) = 154.567, *p* < 0.0001, *d* = 0.80).

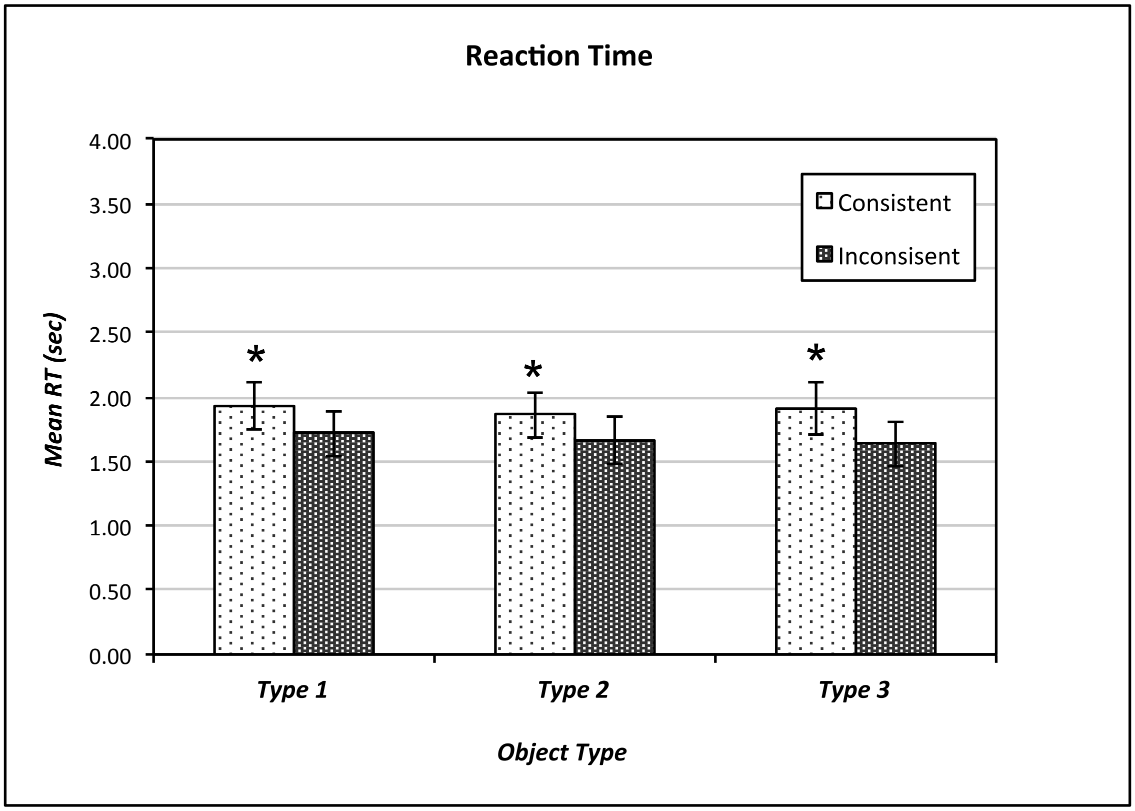


Figure 3-4: Reaction time data from Experiment 3.

Contextual consistency (probe objects appearing with other objects from the same object family) caused significantly longer RTs in remembering the probe object's position than contextual inconsistency (probe objects appearing with objects from dissimilar families). Error bars represent the 95% confidence interval around the mean.

Again, the comparison of probe objects paired with other objects from an inconsistent family was significant for all three object families: for Type 1 paired with Type 2 or with Type 3 objects (both comparisons, *df* = 955, p < 0.0001); for Type 2 paired with Type 1 objects (*df* = 955, p = 0.004) or with Type 3 (*df* = 955, p < 0.0001); and for Type 3 paired with Type 1 or Type 2 (both comparisons, *df* = 955, p < 0.0001). It did not matter which of the inconsistent families a probe object was paired with. RTs were not statistically significantly different between a Type 1/Type 2 pairing and a Type1/Type 3 pairing, between a Type 2/Type1 pairing and a Type 2/Type3 pairing, or between a Type 3/Type 1 pairing and a Type3/Type 2 pairing (all comparisons, *df* = 955, p = 1.000).

2.5 Discussion

These results imply that object identity is indeed attended to in circumstances such as those in this experiment. It is likely that the spatial configuration did not cue the remembered location as much as the identity of the probe object. However, it must be noted that the identity of all of the individual objects that make up the configuration is probably not being encoded even in this special case. Instead, the appearance of probe objects that are markedly different from the other objects also present in the sample scene is drawing participants' attention to these objects. It is far easier to remember the location of a single anomalous object than the features and locations of a number of similar objects, just as in visual search paradigms, it is easier to locate a feature singleton because it pops out (Treisman & Gelade, 1980). A more interesting interpretation of these results, however, is that objects that are visually similar probably do not make the "cut" in drawing visual attention, and thus may not be registered by the visual system as individual instances of object co-occurrence.

3. Experiment 4

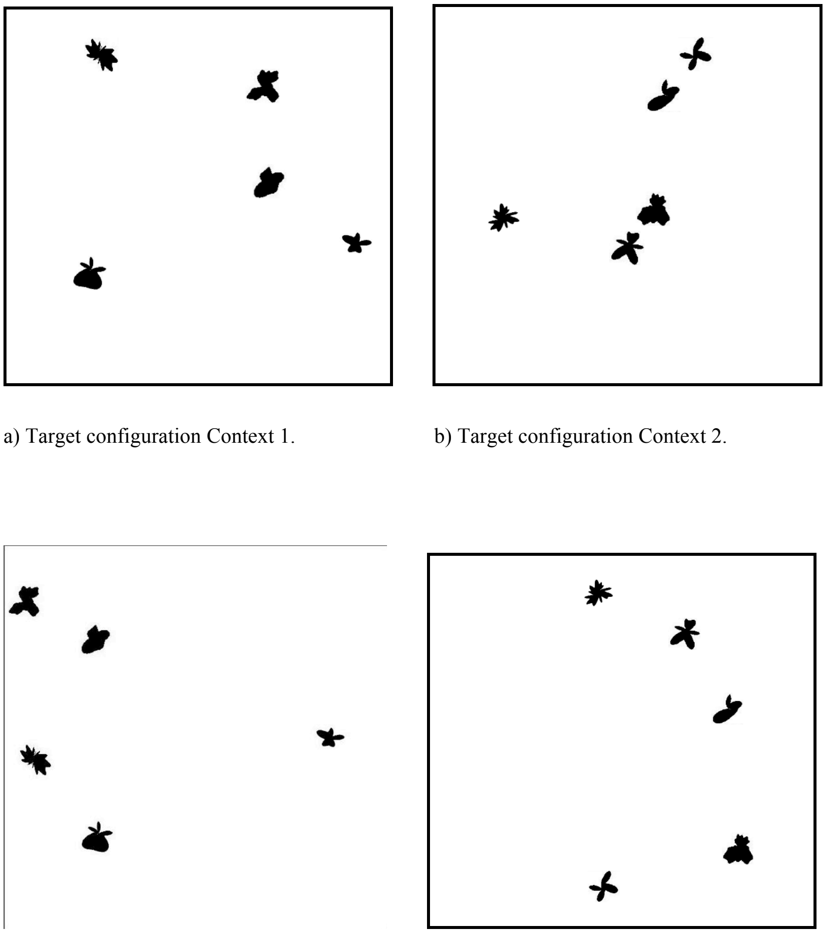
The purpose of Experiment 4 was to ascertain that the types of objects and an experimental protocol similar to what we used in Experiment 3 would replicate some of the findings of previous researchers. Thus, we could more confidently attribute the results of Experiment 3 to a real "inconsistent object" effect produced from a novel baseline.

3.1 Participants

Twenty-two Cornell University undergraduates participated in this experiment. All participants had normal or corrected-to-normal vision. They received course credit for their participation. All participants gave their informed consent.

3.2 Stimuli

The artificial scenes for this experiment consisted of just one object family: the black abstract shapes called Type 1 from the first experiment. Four target scene configurations were generated, referred to here as Contexts 1, 2, 3, and 4. Each context consisted of a given set of object shapes in a given overall configuration (Figures 3.5a, 3.5b, 3.5c, and 3.5d). The configurations were grouped in pairs (Contexts 1 and 2, or Contexts 3 and 4), and each pair was learned by a randomly assigned group of one-half of the participants. Within a pair, the two contexts did not share any object shapes; the same set of object shapes was, however, used for both pairs of contexts. Distractor scenes were also generated, containing a different set of object shapes from either of the two contexts within a pair. Each distractor scene contained a different configuration of the distractor objects.



a) Target configuration Context 1.

b) Target configuration Context 2.

c) Target configuration Context 3.

d) Target configuration Context 4.

Figures 3-5a, 3-5b, 3-5c, and 3-5d: Target configurations for Contexts 1 through 4.

3.3 Procedure

The participants began with a practice session consisting of 10 trials, including one representative trial each of the three experimental conditions for each context (described below), plus 4 distractor trials. The practice trials were not used in the actual experiment.

Each participant learned two target contexts, either Contexts 1 and 2, or Contexts 3 and 4. During the learning phase of the experiment, each of the target contexts was repeated 30 times, plus 60 distractor trials, for a total of 120 trials.

For each trial, the sample scene appeared for 2.0 s, followed by a 1.0-s mask consisting of a medium gray screen, followed by the test scene (Figure 3.6). In test scenes, the scene configuration reappeared, but a single probe object that had appeared in the sample scene was missing. The probe object was the same for each context (i.e., probe object A for Contexts 1 and 3; probe object B for Contexts 2 and 4).

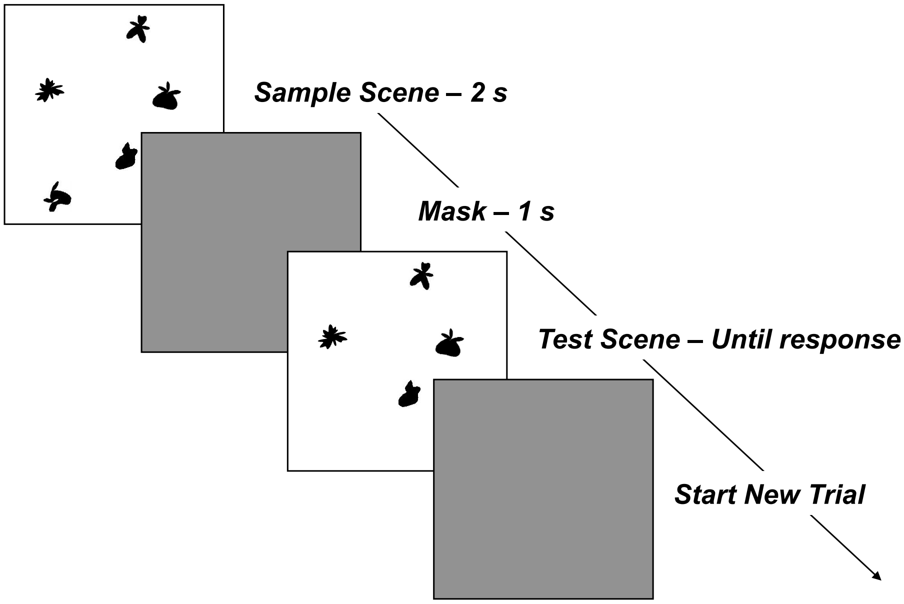


Figure 3-6: Trial sequence for Experiment 4.

From sample to test scene, one probe object was removed from the configuration. Participants clicked on the remembered location of the probe object from the sample scene.

For the transfer phase of the experiment, the learned context scenes appeared as the sample scenes, but the test scenes were altered according to one of three transfer conditions:

- ***Same configuration/same objects condition***: the same configuration was occupied by the same objects as in the sample scene. In other words, sample and test scenes were identical.
- ***Same configuration/different objects condition***: the same configuration as in the sample scene was occupied by different objects, chosen randomly for each trial from among the set of distractor objects and the objects for the opposite context.

- ***Different configuration/same objects condition***: the same objects from the sample scene were arranged in a new configuration.

We did not include a *different configuration/different objects* transfer condition because we judged that these might appear more like distractor scenes to the participants than part of the learning conditions. The transfer phase consisted of 20 trials of each transfer condition for each context, plus an equal number of distractor trials constructed in the same manner as for the learning phase, for a total of 240 trials. Participants were not told that the configurations or the shapes that comprised them were being altered in this phase.

3.4 Results

The dependent measures were again spatial error in pixels and RT. Each measure was analyzed in a four-factor mixed-model regression, where *participant* was entered as a random factor, and the three experimental condition factors were entered as fixed factors (*context–1, 2, 3, 4*; *positions–same, different;* and *object identities–same, different).* Trials with RTs greater than two standard deviations from the mean for all trials within a group of participants (those learning either Contexts 1 and 2, or 3 and 4) were discarded; approximately 3.7% of trials were discarded for the Context 1 and 2 pair, and 3.2% for the Context 3 and 4 pair. Distractor trials were also excluded from the analyses. Because we had decided in advance that a transfer condition of *different configuration/different objects* would be too disconcerting for participants, a fully ranked regression to estimate the interaction of configuration and object identities was not possible.

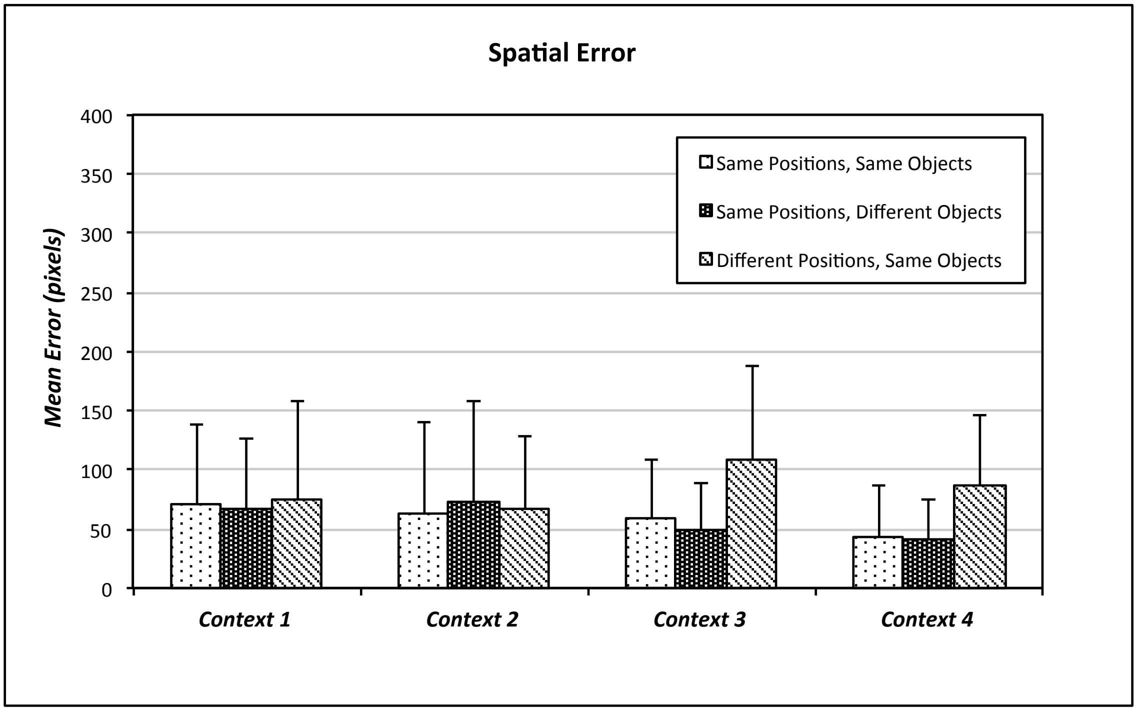


Figure 3-7: Spatial error data from Experiment 4.
Error bars represent the 95% confidence interval around the mean.

The spatial error data showed statistically significant main effects of context (F (3, 56) = 11.42, p = 0.005) and positions (F (1, 2533) = 71.61, p < 0.0001, d = 0.34). The effect of object identities, however, was not at all significant (F (1, 2533) = 0.185, p = 0.667). A second mixed-model regression was performed, with *participant* as a random factor, and *context* and a combined experimental condition factor (*same positions/same objects, same positions/different objects,* and *different positions/same objects*) as fixed factors. This regression also showed a statistically significant main effect of context and an interaction between context and condition (F (6, 2527) = 16.466, p < 0.0001). Pairwise comparisons between the conditions (not obtainable in the previous regression analysis, because it was not fully ranked) across all contexts showed that while there was a statistically significant difference between the *different positions/same objects* transfer condition and each of the other two conditions, there was no such significant difference between the *same positions/same objects* condition and the *same positions/different objects* condition (Fisher's LSD with Bonferroni correction, *df* = 2527, *p* = 1.000).

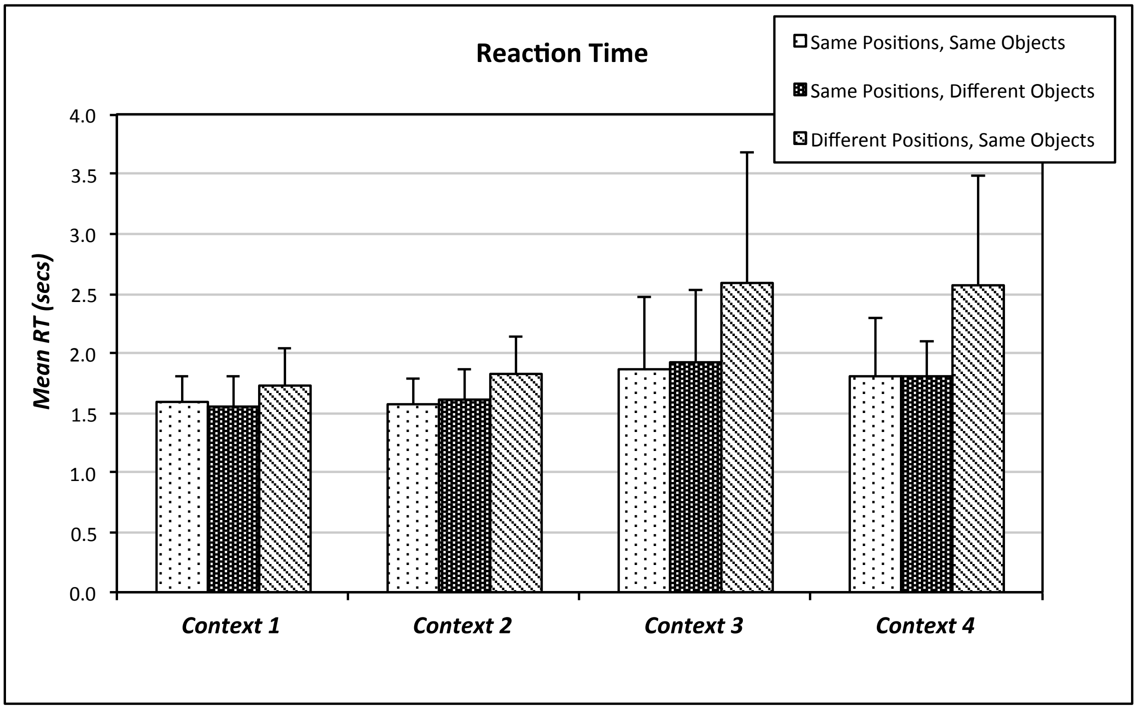


Figure 3-8: Reaction time data for Experiment 4.
Error bars represent the 95% confidence interval around the mean.

For the reaction time data, there was again a main effect of context (*F* (3, 59) = 4.68, *p* = 0.005), and a main effect of positions (*F* (1, 2533) = 135.17, *p* < 0.0001, *d* = 0.46). The effect of object identities, again, was not significant (*F* (1, 2533) = 0.242, *p* = 0.623). A mixed-model regression with *participant* as a random factor, and *context* and a combined experimental condition factor (*same positions/same objects, same positions/different objects,* and *different positions/same objects*) showed a main effect of condition (*F* (2, 2528) = 72.31, *p* < 0.0001) as well as a main effect of context. A statistically significant interaction between context and condition was also found (*F* (6, 2528) = 8.06, $p < 0.0001$). Reaction times for Contexts 3 and 4 were longer than for Contexts 1 and 2, especially in the *different positions/same objects* condition. Pairwise comparison of RTs for the *different positions/same objects* transfer condition with the other two conditions across all contexts were each statistically significant (for *different positions/same objects* with *same positions/same objects*, *df* = 2529, $p < 0.0001$; for *different positions/same objects* with *same positions/different objects*, *df* = 2529, $p < 0.0001$). But comparisons showed that the difference between *same positions/same objects* and *same positions/different objects* was not significant (*df* = 2527, $p = 1.00$). All comparisons were made using Fisher's LSD with Bonferroni correction.

3.5 Discussion

These results are similar to the results achieved by Endo & Takeda (2004) and Jiang & Song (2005). These studies, including the present experiment, found that changing the configuration of a scene's objects resulted in poorer performance, but that changing object identities within the same configuration had no significant effects. The lack of an effect on performance when one object was substituted for another probably indicates that participants were not very sensitive to the shape information of each individual object—as long as the objects were of the same size and a similar enough shape to be taken for one another. More importantly, the results of Experiment 4 suggest that the "inconsistent object" effect found in Experiment 3 probably arises from using a training set composed of several very different groups of objects. In that situation, participants become aware that scenes can contain multiple object styles or appearances; they may remain more aware of object shape, texture, and style. When participants learn only scenes composed of similarly shaped objects, like those used in Experiment 4, they likely find other sources of information than object shape inherent in the scenes to aid their task performance.

The positional cuing effect of the repeated configurations was observed for two learning situations (Contexts 1 and 2, and Contexts 3 and 4). Participants quickly learned that the appearance of the probe object in the sample scene always cued its location in the test scene, regardless of the identities of the objects associated with it in the configuration. However, it was also possible for them to learn the probe object's location without reference to the configuration at all. These results argue that participants in this experiment paid less attention to either the configuration or the identities of the objects that comprise it, but rather to the identity of only one object per context and its absolute location maintained across all conditions of the other objects' identities and locations. The finding that reaction time was longer for the transfer condition where the other objects appeared in a new configuration suggests that the probe object's location could be encoded both in absolute coordinates in the scene area, and as the "missing" member of a sample scene configuration (regardless of the identities of the other objects in the configuration). The absolute location may be more difficult to recall from long-term memory than the "missing" object's coordinates from a sample scene just present in visual working memory.

Like other studies where the formation of a context including object identities is attempted from a novel baseline, the results of Experiment 4 suggest that the contextual cuing effect of spatial position predominates over the effect of object identity, but perhaps only in situations where the configuration is easier to remember than the identities. Such situations might occur where visual working memory is overtaxed, or where the identity of other objects in the configuration is not required for good performance on the task. Neither interpretation indicates that object identity information is discarded in these memory tasks, only that the examined tasks do not rule out such an interpretation. We acknowledge that this experiment only touches on the many ways in which the visual properties of controlled stimulus arrays differ from those of the natural world.

4. General Discussion

A human visual system viewing the natural world can never learn to associate collections of objects and surfaces with each other without at the same time learning their relative spatial positions. Although this fact makes the design and interpretation of laboratory experiments using novel objects more difficult, the disadvantage is more than outweighed by the increased ecological validity of the learning situation. Rather than concluding that learning identity context is secondary to learning spatial context, we ask how conditions in the laboratory differ from natural viewing, and whether those conditions work against the formation of identity context. In Experiment 3, we showed that simply by training participants on scenes composed of objects that were very visually different from each other, we were able to produce an "inconsistent object" effect similar to the results in experiments using line drawings and photographs of scenes with which participants already have a great deal of real-world experience. These results are the first we are aware of that reproduce the inconsistent object effect in a laboratory setting from a novel baseline.

The effect of spatial context does appear to be profound; spatial configurations are learned more readily under more difficult conditions than object identities. Our results in Experiment 4 found an advantage for spatial configuration resembling the findings of other researchers (Endo & Takeda, 2004; Jiang & Song, 2005) even though we simplified the learning task. It is also widely held that scene recognition, or at least categorization, relies heavily on the low spatial frequencies of a scene's overall layout (Torralba, 2003). These factors pose certain problems for the view that an inventory of object co-occurrences is automatically extracted through experience with similar scenes. Statistical learning of object associations must take place, however, because inconsistencies in scene content are so easy to notice, as Figure 3.9 demonstrates. Such scenes are striking, because most people's visual experience does not include alligators in bathtubs.

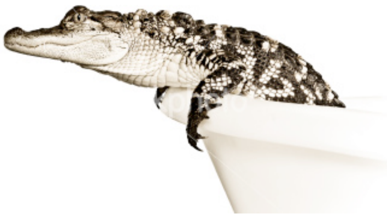


Figure 3-9: Objects out of context usually attract immediate attention.

To counteract the predominance of spatial context learning, we must ask what conditions best facilitate the learning of identity context. An object's ability to draw and/or focus attention is clearly one of these factors. Enhanced ability to discriminate among objects in scenes is probably another.

On the other hand, a second interpretation of these combined results is that observers do register the identities of objects in particular spatial configurations as they learn from repeated presentations, and that when they are probed for the remembered location of a given object, they are simply not distracted by the identities of other objects in the configuration. This interpretation is reinforced by the results of Experiment 3, where faster

RTs and more accurate memory for position of the out-of-context objects were probably due primarily to the objects' identities relative to objects from the other families.

Whether we accept the view that the spatial layout of scenes is a more salient memory for observers than the identities of the object that comprise it, or that object co-occurrence is learned but is merely not well tested by these kinds of experiments, the results of this study and others show that using visually similar objects in artificial arrays leads to equivocal findings.

There is evidence that observers can learn many kinds of contextual associations, including associations among collections of objects presented sequentially rather than spatially (Fiser & Aslin, 2002), non-visual semantic characteristics of object collections (Goujon, Didierjean, & Marmeche, 2007), and likely movements of objects along learned trajectories (Chun & Jiang, 1999). However, display conditions under which this learning takes place in such controlled experiments have been manipulated to downplay the cue importance of spatial configuration. Most such research uses displays of objects that are of similar size, and usually somewhat controlled for similarity of shape. Yet in free viewing of the natural world, most scenes offer a multiplicity of objects of varied sizes and visual appearances. It is not just that objects are easily discriminable by shape: shape may differ quite radically, but so also do attributes like size, color, and texture or pattern. Relatively few scenes in our experience offer collections of objects that are visually homogeneous. Examples might include the sight of people in crowds, herds of animals, flocks of birds, automobiles in parking lots, or collections of similar artifacts. In such scenes, the visual system might not bother to extract the unique shape or other visual attributes of each member object, but only the configuration formed by the collection of objects.

This picture also accords well with research that shows that the earliest stages of scene recognition or categorization are driven by low spatial frequencies (Oliva & Torralba, 2001; Torralba, 2003; Torralba & Oliva, 2003). Scenes like those used here in Experiments 3 and 4 have almost the same power spectrum profile in the lowest spatial frequencies, because the objects that comprise the scenes are the same size and have similar amounts of filled contour. Meanwhile, the higher spatial frequencies that delineate the different shapes of the constituent objects are processed later in the visual stream (Grill-Spector & Kanwisher, 2005).

The results of Experiment 3 occur partly as a result of visual pop-out (Treisman & Gelade, 1980; Franconeri, Hollingworth, & Simons, 2005; Maljkovic & Nakayama, 1994), where a target can be more quickly located among distractors if it has one or more unique perceptual features, such as a different color, shape, or orientation, or if its appearance creates luminance transients in the scene. It is important to note that visual pop-out occurs whether or not participants have been trained to expect the kinds of elements in the display; elements pop out precisely because of perceptual features that are pre-attentively evident at relatively early stages of visual processing, before object recognition can be assumed to have taken place (Wolfe & Bennett, 1996; Wolfe, 2003. But see Joseph, Chun, & Nakayama, 1997). In the case of Experiment 3, however, both visual pop-out and expectations induced by familiarity with the elements of the display are at work. This may be the very definition of "out of context." The features that cause the object to pop out of the display drive the observer's attention to it, followed by recognition of the object and awareness that it is different from other objects. An object that attracts little attention when it appears with other objects of like kind attracts immediate notice when it appears with elements that are dissimilar to it.

The results of this study indicate that contextual learning of object co-occurrence can occur if some of the presented objects are visually distinctive from one another. They do not, of course, tell us how different from each other the objects must be, or whether this criterion would fail if all the objects in a scene were visually distinctive. They merely indicate that using at least one object that is visually distinctive appears to attract observers' attention when it appears with other objects that are relatively visually homogeneous. However, the ability to draw attention and remain in observers' memories is very likely a prerequisite to the registration of object co-occurrence, and thus to the kinds of implicit learning that result in scene schemas and visual context in general.

CHAPTER 4
MEMORY FOR WELL-ESTABLISHED CONTEXTS

1. Introduction

So far we have examined the two aspects of visual context formation that together form the necessary components of scene structure that can be learned solely from visual experience—the whats (identity context) and the wheres (spatial context). In doing so, it has become apparent that visual contexts are collections of long-term memory traces that affect how the structure of all future scenes is interpreted. We have seen how these memory traces affect scenes with either identity or spatial elements that are “out of context”—observers exhibit behavioral cues implying that they notice abnormal elements and interpret the scenes differently. But what happens over time as observers encounter more and more scenes that are “in context”? What happens as the same scenes that formed a context originally are repeatedly encountered over time? In other words, how is long-term memory for scenes affected once a context is formed? The nature of scene memory representations is sufficiently unclear to make these questions difficult to answer. To even begin, we must consider whether existing research on long-term memory for natural scenes can provide any insight, whether alternative accounts of visual memory are warranted, and most importantly, how representations of very similar scenes could be discriminated.

Consider that the best-established visual contexts any observer can probably have are those presented by her home environment. The rooms in her home and her workplace, the stores and public places she frequents, the objects she owns and uses every day, and her family members, friends, and pets have already appeared to her far more often, probably by orders of magnitude, than any other scene content she could have encountered. These visual contexts are already well established; as long as she continues to inhabit the same environment, she will accumulate more and more instances of almost exactly the same scene content from almost exactly the same viewpoints. Therefore, the observer's long-term visual memory contains either hundreds of thousands of extremely similar visual memories of the scenes of her environment, which seems altogether unlikely, or some more generalized representations of each type of scene. Which case is true gives us important information about the actual nature of visual context. If a context is a large group of very similar episodic memory traces, how can an observer discriminate among those traces? If a context is a generalized memory representation, which elements or aspects of scene structure are actually encoded and which are lost?

Although visual contexts are quite plastic, as previous research and the preceding chapters have shown, there is no reason to suppose that a scene context such as "my bedroom" would be materially altered by the accumulation of another year's worth of glimpses of it from the same viewpoint of its doorway.

1.1 VLTM capacity: many different elements vs. many similar elements

Considerable research has been done in an effort to find the absolute capacity of long-term visual memory. Yet the intuition that we probably cannot remember large numbers of very similar scenes flies in the face of the findings of the picture memory research. This literature holds that indeed we can store extremely large numbers of visual memories. Studies testing the limits of our ability to remember complex visual scenes began decades ago, with the number of images used ratcheting upward from several hundred (Nickerson, 1965; Shepard, 1967) to several thousand (Standing, Conezio, and Haber (1970), culminating in Standing's well-known 1973 study, where participants were presented with

10,000 pictures over several days. The consensus of these studies was that humans can remember scenes they have seen only once with remarkable accuracy and that this capacity may be almost unlimited. After reaching this consensus, interest in testing picture memory's limits waned. However, vision research has recently seen a resurgence of interest in the characteristics and capacity of this type of memory. A number of influential papers have reexamined this issue in light of new understanding about how visual long-term memories are formed, indicating that many purely visual details of scenes are encoded (Brady, Konkle, Alvarez, and Oliva, 2008; Henderson & Hollingworth, 2002; Hollingworth, 2005; Konkle, Brady, Alvarez, & Oliva, 2010a; Vogt & Magnusson, 2007).

It must be kept in mind that this type of research was not designed to answer precisely the question of the capacity of long-term memory for a given visual context. Most studies in this literature are optimized to find the limitations of visual long-term memory for the opposite case—for as many different contexts as they can devise. For example, in most picture memory studies it is either explicitly stated or implicitly assumed that the pictures chosen as stimuli are both subjectively memorable and very different from each other. In effect, researchers have been looking for the limits of memory under optimal conditions of memorability. A difficulty arises when we assume that optimal limits equate to the practical limits of visual memory for scenes with well-established. In order to be remembered, pictures must be discriminable from one another. Yet only a small number of studies have paid attention to the criteria on which pictures are discriminable.

The nature of picture memory experiments is also very different from the kind of everyday visual experience that would result in context formation. During most picture memory experiments, participants encounter in a relatively short time period a wide range of visual content, with far more diversity than they would ever see in the same length of time in everyday life. We cannot, therefore, assume that the picture memory research results so far obtained are a good predictor of how many scenes a person could actually remember if the same number of highly diverse images was presented over a much longer time.

1.2 Picture selection biases

Likewise, the choice of stimuli for many well-known picture memory experiments may be biased in ways that preclude their application to questions about memory for a single quotidian context. One source of bias is *photographic.* The very act of taking a photograph implies that the photographer considers that glimpse of his visual world to be more worth recording than other ephemeral glimpses: he takes a photograph because he wants to remember that scene. Most photographers tend to center a single focal object in their images. A bias to present a single focal object per photograph may mean that less attention is required to encode the scene, and thus that its contents can make their way into memory more easily. Moreover, the choice to use photographs collected from multiple sources and photographers, as in the well-known Corel collection of stock photographs, is also biased; the images are noticeably prettier and more distinctive than the average amateur photographer's collection, precisely because the photographers are presenting the best examples of their talent. Their images are more artfully composed, better lighted, and, if taken in color, more vivid than the average amateur's snapshots.

Another source of bias is *selection*: In compiling a collection of a very large number of pictures, the individual images are already selected with a number of biases, including experimenter selection and source biases. As images are considered for inclusion, experimenters may consciously or unconsciously develop criteria that affect how different the images are in any number of ways. Images taken from news and advertising sources, such as have been used in several studies (Standing, Conezio, and Haber, 1970; Standing, 1973) might exhibit a different group of biases. The subject matter may differ in unknown ways from images that people encounter in everyday life. Odd juxtapositions of content that advertisers intend to increase the image's memorability are common.

A final source of bias is *anthropocentrism*. That is, the images used in picture memory studies have almost always featured visual content that is of intrinsic interest to humans. Pictures of animals, buildings, artifacts, and especially of other humans naturally attract attention, and are probably more memorable than pictures that do not feature such content. Given that a number of visual brain areas respond preferentially to anthropocentric content, including faces, human-made objects, and buildings (Aguirre, Zarahn, & D'Esposito, 1998; Chao, Haxby, & Martin, 1999; Epstein & Kanwisher, 1998; Grill-Spector, 2003; Martin, Wiggs, Underleider, & Haxby, 1996), the memorability of this content is not surprising.

These limitations on the applicability of picture memory research, however, do not detract from their central message: the capacity of long-term visual memory is great if it is filled with the right kinds of memories. What remains to be seen is whether and how many instances of a given context that do not display most of these types of bias fit the criteria.

1.3 Detailed vs. conceptual scene representations

It is unclear why we should have such a high-capacity memory for scenes, especially because the capacity of visual working memory is so small and because its contents require so much attention to capture and maintain. Memory for other kinds of visual stimuli is quite limited. For example, memory for quantitative differences in continuous perceptual dimensions such as length or brightness is limited to a relatively small number of divisions (the famous “7 plus or minus 2” of George Miller’s much-cited paper [1956]), and memory for isolated objects in stimulus arrays is limited to just four or five (Irwin & Zelinksy, 2002; Vogel, Woodman, & Luck, 2001; Wolfe, 2003). However, other lines of research indicate that working memory probably plays only a small part in the overall representation of a scene because memory for individual objects coheres in a somewhat durable memory representation as multiple fixations are made over a scene (Gajewski & Henderson, 2005; Hollingworth & Henderson, 2002; Hollingworth, 2005; Melcher, 2006). The notion of a long-term memory of a scene as a composite taken from multiple glimpses offers a tempting explanatory mechanism for the formation of visual contexts. It cannot be accepted, however, without considering how much detail is incorporated into the composite.

Recent research has shown that observers retain a great deal of visual detail, sufficient to discriminate among different exemplars of objects in the same category, and even of different states of the same object. Vogt and Magnusson (2007) examined memory for pictures of exactly the same content: front doors of houses. They found that after viewing 400 images of doors, participants could determine whether a given image had been edited to remove some of its distinctive details with 63.6% accuracy 30 min after presentation. After nine days, recognition memory for altered images was still approximately 59%.

Brady *et al.* (2008) presented participants with 2500 images of various household and naturally occurring objects. They then tested recognition memory in three conditions: previously viewed objects paired with new objects; previously viewed objects paired with new exemplars from the same category of objects; and previously viewed objects paired with the same object in a different configuration or “state.” In this condition, for example, a toaster oven with a loaf of bread inside it might be paired with the same oven where the bread was positioned on the open door of the oven. Memory performance was quite high for all three conditions (92.5% for the novel objects condition, 87.6% for the different exemplar condition, and surprisingly, 87.2% for the different state condition).

Konkle *et al.* (2010a) showed participants as many as 64 exemplars within a single semantic category in an experiment that used 128 distinct semantic categories of scenes, with a total of 2912 images over 5.5 hours. Participants’ performance on the 64-exemplar categories was 76%, compared with 84% on categories with four exemplars. Performance dropped only 1.8% with each doubling of the number of exemplars, evidence of the considerable amount of visual detail encoded in order to discriminate among large numbers of exemplars.

Thus it appears that memory for large numbers of pictures can include a great deal of visual detail, but in all the studies reviewed so far, the kind of visual detail and the degree of variety in the picture sets has been optimal for human observers in several ways, as the discussion of biases in the previous section showed. Additionally, although the overall number of images used in these studies was fairly large relative to picture memory experiments in general, the number within each category, or context, was only a fraction of the fixations on a scene that might accrue in a single hour spent in the same physical environment, such as one’s office.

1.4 Semantic subdivision of picture memory

The question naturally arises whether the ability to remember large numbers of diverse scenes is due to visual memory alone. Some influential picture memory research has explored whether the retained information is purely visual in nature, or is remembered only as a conceptual label (Bransford & Franks, 1971; Friedman, 1979; Gentner & Loftus, 1979; Potter, 1976; Potter & Levy, 1969; Potter, Staub, & O'Connor, 2004). For a time the idea that memory was stored entirely as semantic, conceptual codes and not as specific perceptual details was current. False memory for visual details could be elicited if participants viewed pictures that were consistent with a particular schema conveyed through a story (Friedman, 1979; Mandler & Johnson, 1976). This conception vied with the notion of the "picture superiority effect," whereby it was assumed that pictures conveyed more information than sentences about the same topic, because pictures are richer in perceptual detail. But early research along this line failed to find much evidence for differences among images that varied in detail level (Pezdek *et al.*, 1988; Pezdek, *et al.*, 1989; Viera & Homa, 1991), leading researchers to believe that pictorial detail is encoded purely as part of conceptual or schematic constructs, and that increased detail does not lead to greater picture memory.

A more useful potential explanation might be that VLTM relies on both semantic memory to subdivide the images and purely visual memory within a semantic category to encode the details that permit discrimination among similar images. This explanation might help explain how the visual system encodes many different contexts with many different examples, although it would be of only marginal use if the number of contexts an observer must remember is limited, but the examples are unlimited. This kind of situation is more likely what obtained in the evolutionary course of the human visual system: contexts were fewer because hominids had a limited range of physical environments and almost no artifacts.

Naturally, the more accessible detail in scene memory, the greater an observer's ability to discriminate among similar scenes. We can only speculate, however, on how the detail is stored, whether it might itself be conceptual in nature (i.e., "brown rock in lower left corner") or purely visual. Nor can we do more than speculate on which details are critical for discrimination in a given context or how different they must be, as was shown by the results of the experiments in Chapters 2 and 3.

1.5 Information and memory for scenes

Without indulging in too much speculation, on the other hand, it may be possible to determine if a given approach might be fruitful in answering such questions. Instead of focusing on the differences between conceptual and visual memories, we can examine current research exploring the mappings between the organization and function of the visual brain and the task requirements of human operation in the natural world. This line of research has concentrated on the efficient coding of natural scenes by the early visual system (e.g., Field, 1987, 1994; Simoncelli & Olshausen, 2001; Graham, Chandler and Field, 2008). This work has argued that the parameters of neurons at least in the early stages of the visual pathway are very well suited to the statistical properties of natural scenes. Our visual system efficiently encodes the natural world by taking advantage of the power spectra (Field, 1987), sparse structure (Graham, Chandler, & Field, 2008), contour structure (Field, Hayes, & Hess, 2003; Tversky, Geisler, & Perry, 2004), and other statistical properties of images.

Given the connection between neural organization and scene properties, the inevitable next step has been an attempt to find some set of metrical properties of images that can predict human observers' performance on a wide variety of visual tasks, such as detecting, identifying, or classifying objects, or estimating their distance or motion (see Geisler, 2008 for a detailed review). Preliminary visual tasks such as eye movements and attention to areas of a scene can be inferred using metrics such as contrast (Reinagel & Zador, 1999), visual salience (Bruce & Tsotsos, 2009; Itti and Koch, 2000; Parkhurst, Law, & Niebur, 2002; Torralba, Oliva, Castelhano, & Henderson, 2006), or visual surprise (Itti & Baldi, 2006) to predict fixation patterns in viewing natural scenes.

Later visual processes that are likely to be involved in forming pictorial memories are also influenced by scene statistics. Scenes can be categorized at a high level (that is, into broad categories such as *beach*, *mountain*, or *forest*) within 19-67 ms, more rapidly than the individual objects within the scene could be attended and recognized (Greene & Oliva, 2009a). Scenes can be matched to a verbal description given a pre-cue, such as *picnic,* in a single glance, within a rapid serial presentation stream of about six images per second (Potter, 1976; Potter, Staub, & O'Connor, 2004). Observers can decide whether a scene contains an animal or not in the merest fraction of a second with remarkable accuracy (Fei-Fei, VanRullen, Koch, & Perona, 2005; Joubert, Rousselet, Fabre-Thorpe, & Fize, 2009; Kirchner & Thorpe, 2006). However, higher-level visual processes may predominate over processes based on lower-level scene statistics if the higher-level process directs attention to an object or area of the scene that is more behaviorally relevant than the lower-level process does. Such evidence also suggests that the results of the higher-level process are more likely to make their way into memory (Stirk & Underwood, 2007). Work to discover what metrical properties the visual system is using to make such rapid decisions is ongoing (Greene & Oliva, 2009b).

These findings lead us to speculate on what relationships might be found between the metrical properties of images and the very large capacity of long-term memory for natural scenes. In other words, does visual memory make use of all the information that can be measured from an image? To date, the relationship between images' statistical properties and visual memory is not clear. It might seem worthwhile to approach the problem of scene memory using classic information theory (Shannon, 1948). Information theory has proved valuable in understanding how a perceptual signal with known characteristics is transmitted through a channel such as the human perceptual system with some known but many unknown characteristics. A complete discussion of information theory as applied to images is beyond the scope of this paper. However, *entropy*, as Shannon defined it, may be directly related to the visual system's ability to discriminate among scenes.

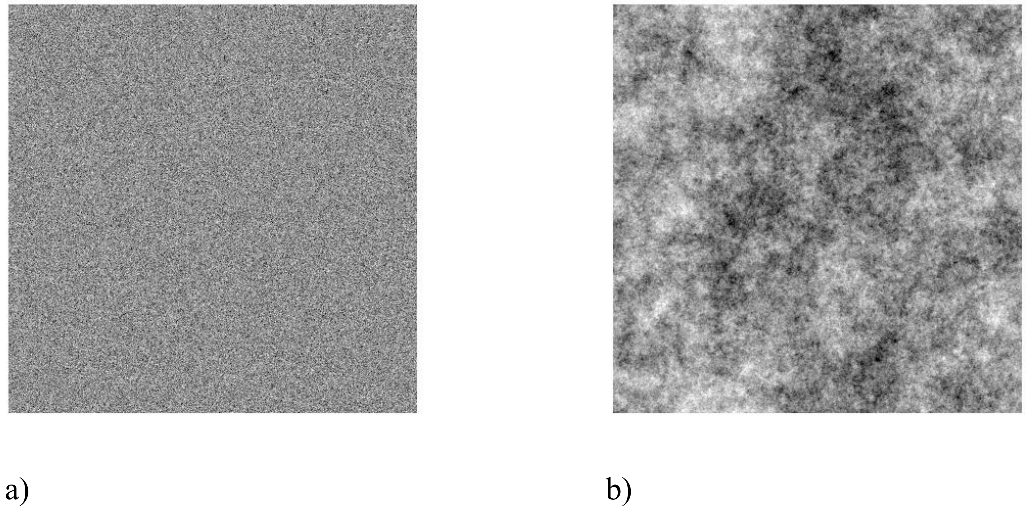
As a justification for using entropy as a metric for memorability, consider the high-dimensional space of all possible grayscale images of a given size with 256 levels of gray, where the axes of the space represent the intensity of each pixel in the image. The number of possible 256 x 256-pixel images in such a space is $2^{(256 \text{ x } 256 \text{ x } 8)}$ or $2^{524,288}$ images, a quantity much larger than the number of particles in the universe. Within this space, the vast majority of images will appear to be noise. A very small number will appear to have recognizable features (consisting of objects and surfaces) that are recognizable to human observers. Within that tiny group, a smaller number will be discriminable to human observers, and a yet smaller number will be sufficiently distinctive to be remembered as unique. But within this state space of possible images, *any* particular image can be simply described by a list, or vector, of its pixel intensities, just as a digital camera does. Such a vector comprises a complete and unique descriptor of any possible scene within this space.

If the visual system had access to information in the form of this vector, it could accurately discriminate between scenes, no matter how similar they might appear to be.

To explore the ramifications of these observations, the present study compares recognition memory performance for images that are distributed throughout the image space in different ways—in other words, for images that contain differing amounts of information. Comparing categories of images that have somewhat well understood informational metrics with the category we are most interested in—those of a single type of context but with widely varying details—might either show a promising trend or allow us to rule out information theory as an approach to measuring detail.

'White noise' images are defined as those where the intensity of each pixel is selected randomly from a uniform or Gaussian distribution (these images are also often called *independent, identically distributed* or *i.i.d* noise). The intensity of each pixel is therefore independent of the values of all other pixels. Such images have minimal redundancy. Within the space of all possible images, these white noise patterns represent random points in this space. For a given variance, they have the least redundancy (the highest entropy). This also means that white noise images are statistically quite different from each other. In other words, they would be maximally distributed throughout the image space. White noise images as a category can be characterized as being at the extreme of statistical difference from one another, but at the opposite, negative extreme of human discriminability and memorability.

Images of '1/f noise,' on the other hand, exhibit the same power spectra as scenes that are more recognizable to human observers (Field, 1987). They contain easily discriminable structure, somewhat resembling cloudy skies, but because their phase spectra are not aligned, they lack the defined edges that normally signal objects to the human visual system. As a category, they are widely distributed in the image space, but not so widely as white noise images.



Figures 4-1a and 4-1b: Examples of white noise and 1/f noise images.

Images of natural scenes excluding human-made structures and objects represent a category that the mammalian visual system evolved to discriminate. They might be characterized as relatively closely clustered in the image space, but relatively widely distributed in some hypothetical representational space where the dimensions are those that matter to human long-term memory. This category of images represents the ideal single visual context to test for information availability. It is probably the most well-established context in human experience; although it might not represent the observer's own physical environment, it is likely to resemble some aspects of it quite closely. The category avoids some of the pictorial biases discussed earlier, especially the anthropocentric bias, that might skew memory performance. Such images are also rich in visual detail that cannot be readily given a semantic label.

Finally, images of artifacts, buildings, animals, and situations that are relevant to human behavior represent a category that can be characterized as relatively low in entropy compared to noise images, but perhaps more widely distributed in the hypothesized memory representation space. The greater the degree of visual and content variety in this category, we propose, the greater the distribution in memory space and thus the higher degree of recognition memory. Although the dimensions of the memory representation space are unknown, we may be able to infer from the composition of a given image set what its parameters might be. This category of images also replicates some of the characteristics of image sets used in picture memory research that has produced high memory performance.

2. Experiment 5

2.1 Participants

Thirty-eight Cornell University undergraduates participated in the experiment. The experiment lasted for about 35 minutes. All participants had normal or corrected-to-normal vision. They received course credit for their participation. All participants gave their informed consent.

2.2 Stimuli

We used five categories of images. The first category, which we term here “High Variety,” was collected from two sources: One was taken from the publicly available dataset at the LabelMe website, an annotation project where volunteers trace the outlines of and label objects in photographs (Russell, Torralba, Murphy, & Freeman, 2008). The specific folder from which we took images was available in 2008 at: http://labelme.csail.mit.edu/browseLabelMe/static_web_tinyimages_testset.html. The second source was the Berkeley Image Segmentation Dataset, available at http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/. We selected from these sources all photographs that were at least 256 x 256 pixels in size. We then culled the photographs to remove any that had a) human faces at a size that would make them recognizable; b) readable text in signage or other labels (although we did permit images with readable digits); c) images of web pages, advertisements, graphs, and other text-based content; and d) any pictures that in the authors' subjective opinion were too similar to each other, such as multiple images of the same group of fish. This selection process yielded approximately 700 images. We processed the photographs by reducing them in size to 256 pixels on the shorter dimension while preserving the aspect ratio, then randomly cropping them to produce a 256 x 256-pixel image. The photographs were converted to grayscale using the MATLAB rgb2gray command, which converts RGB images to grayscale by eliminating the hue and saturation information while retaining the luminance. Finally, images were saved in Portable Network Graphic (.png) format.

The second category of images was taken from the van Hateren collection of images available at: http://www.kyb.mpg.de/bethge/vanhateren/index.php. This collection of photographs contains approximately 4000 images of natural scenery taken in the countryside in the Netherlands (van Hateren, J.H., van der Schaaf, A., 1998). We randomly selected 100 photographs from the collection, processing and converting them to 256 x 256-pixel images in the same way as for the "High Variety" category described above.

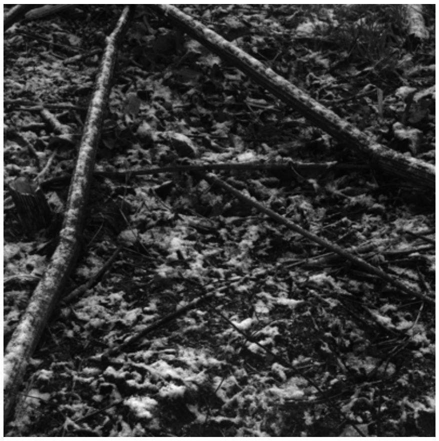


Figure 4-2: Example of natural scenery image from the van Hateren image database.

The third category of images was taken from a collection photographed by the author. This set of images was taken with a Canon PowerShot A400 digital camera with a resolution of 2048 x 1536 pixels, using fully automatic settings for focus, aperture, and speed. All images were of views from walking paths in the Peebles Island State Historic Site in Cohoes, New York taken in August, 2008. Our method was to take a single image approximately every 100 ft. while walking along the paths. Most images were taken of the view directly in front of the photographer, at normal eyeheight, although some viewpoints of terrain on either side and sometimes at the photographer’s feet were also included. The 100 images used in this experiment were thus contiguous views of approximately the first two miles of walking paths from the entrance of the site, and were presented in the order taken while the photographer walked the path. We term these images the “Walk in the Woods” category. They were prepared for presentation in the same manner as the “High Variety” and “van Hateren” images.

The fourth category consisted of 100 images of 1/f noise created by a MATLAB algorithm. The images were created to be 256 x 256 pixels in size. The final category was

100 images of white noise, which were also created with MATLAB at a 256 x 256 pixel size. The1/f noise and white noise images were multiplied by a factor of -1.8 to preserve their random distribution of intensities when presented on a computer monitor with a gamma of 1.8.

The images were displayed on a 23-in. (58.4 cm) Macintosh Cinema HD LCD computer monitor with a resolution of 1280 x 800 pixels. Each image measured 3 7/8 x 3 7/8 in. (9.8 x 9.8 cm) and subtended approximately 10.17° of visual angle horizontally and vertically. The monitor was viewed from a distance of approximately 25 in. (63.5 cm) in a normally lighted office. The gamma of the monitor was set to 1.8.

2.3 Procedure

The experiment consisted of three parts. The first part was designed to compare memory for four images from each image set. Our purpose was to see if participants were capable of remembering even a small number of the more difficult categories of images. The second part of the experiment tested memory for 100 images from each set. The purpose was to compare performance over a set that was likely large enough to reveal differences among the sets, and perhaps suggest performance limits for the more difficult categories. The third part was a side-by-side comparison of 20 pairs of images from each set. This part was designed to reveal if participants could actually discriminate between images from the more difficult categories when both images were present for comparison.

Parts 1 and 2 followed exactly the same procedure, except for the number of images presented. In each, the five sets of images were presented for a two-alternative forced-choice (2AFC) task. Each participant saw all five sets. The order of set presentation was randomly chosen for each participant. Our method was essentially the same as that used by

Standing (1973). For each set, either 4 or 100 images were first presented one after another for 1 s each, centered on the screen, surrounded by a medium gray background. Each presentation was followed by a 0.5 s ISI consisting of a gray screen. Immediately following the study presentation, the 2AFC task presented two images side by side for 2 s. One of the two images was new, while the other was randomly chosen from the previously presented images. All old images in each set were tested for the short version of this task; 40 randomly chosen old images were tested in the long version.

Participant responded by pressing either the “S” key if the image they had seen before was on the left side or the “L” key if it was on the right. Participants could take as long as they wanted to make a response; the two images were visible until a key was pressed.

Before beginning the experiment, participants were shown a poster containing sample images of each image category (none of which was used in the actual experiment). They received instructions on-screen, then began the experiment. They did not receive any performance feedback during the experiment.

The experiment was controlled by MATLAB 7.0 (R14) with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997).

2.4 Results

The results from the second part of the experiment will be discussed first, for simplicity of explanation.

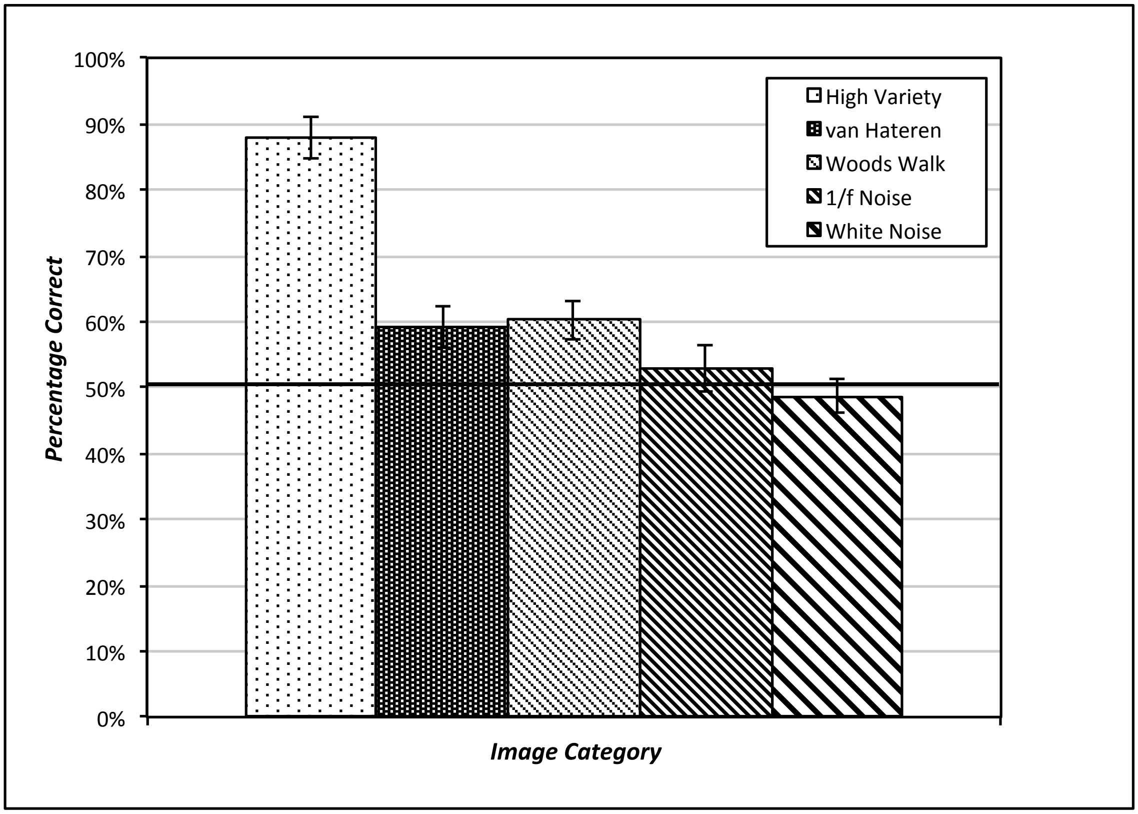


Figure 4-3: Memory performance as percentage of 100 images correctly recognized as previously seen. The error bars represent the 95% confidence interval around the mean.

As might be expected, memory performance for 100 images of both white noise and 1/f noise was not statistically significantly different from chance ($t(37) = -0.79$, $p > 1.00$ for white noise; $t(37) = 1.85$, $p = 0.35$ for 1/f noise). (Note that p values for all tests in this section have been Bonferroni-corrected for the number of comparisons made.) The simplest interpretation of these results is that the entropy of an image is not well correlated with the likelihood of its being remembered. Assuming that the entropy of both white noise and 1/f noise images is higher than the entropy of the van Hateren, Walk in the Woods, and High Variety images, the poorer performance on the first two categories indicates that the information an image conveys is probably not available when needed for recognition.

A second obvious interpretation of the results is that having a high degree of visual variety in the image set improves recognition performance by well over 20% compared to performance for any of the other categories. These results are in line with, though slightly lower than, other picture memory research. However, as with other experiments, these results have only been obtained when the pictures were specifically chosen to be highly memorable and very discriminable.

The most interesting case is that of the Walk in the Woods and van Hateren images, which are both relatively widely distributed in image space and also fairly widely distributed in memory space. Memory for these images is statistically significantly better than chance ($t(37) = 7.29$, $p < 0.0005$, $d = 2.4$ for Walk in the Woods, $t(37) = 5.91$, $p < 0.0005$, $d = 1.94$) for van Hateren), but considerably less than performance on the High Variety image category $t(37) = 24.15$, $p < 0.0005$, $d = 7.97$). Many of these images have similar content and structure, so the better-than-chance performance on these categories seemingly indicates that some quantity of visual detail is encoded and available for recognition. Yet the much lower rate of recognition compared to the High Variety category indicates that many of these images are being confused. Thus, it appears that either the amount of or the type of visual detail is insufficient to discriminate among images. From these results, however, we cannot infer which situation obtains.

The poor memory performance for white noise and 1/f noise obtained in Part 2 of the experiment is not solely due to the number of images the participants were asked to remember. Part 1, where they were presented with only four images from each category, produced results that were still not significantly better than chance for these categories ($t(38) = 1.13$, $p > 1.00$ for white noise; $t(38) = 2.55$, $p = 0.75$ for 1/f noise). Yet participants were able to remember a small number of both the natural scenery categories at rates much better than chance ($t(38) = 13.22$, $p < 0.0005$, $d = 4.29$ for Walk in the Woods; $t(37) = 17.15$, $p < 0.0005$, $d = 5.64$ for van Hateren), and only a little less than for the High Variety category. Participants remembered 100% of the four High Variety images.

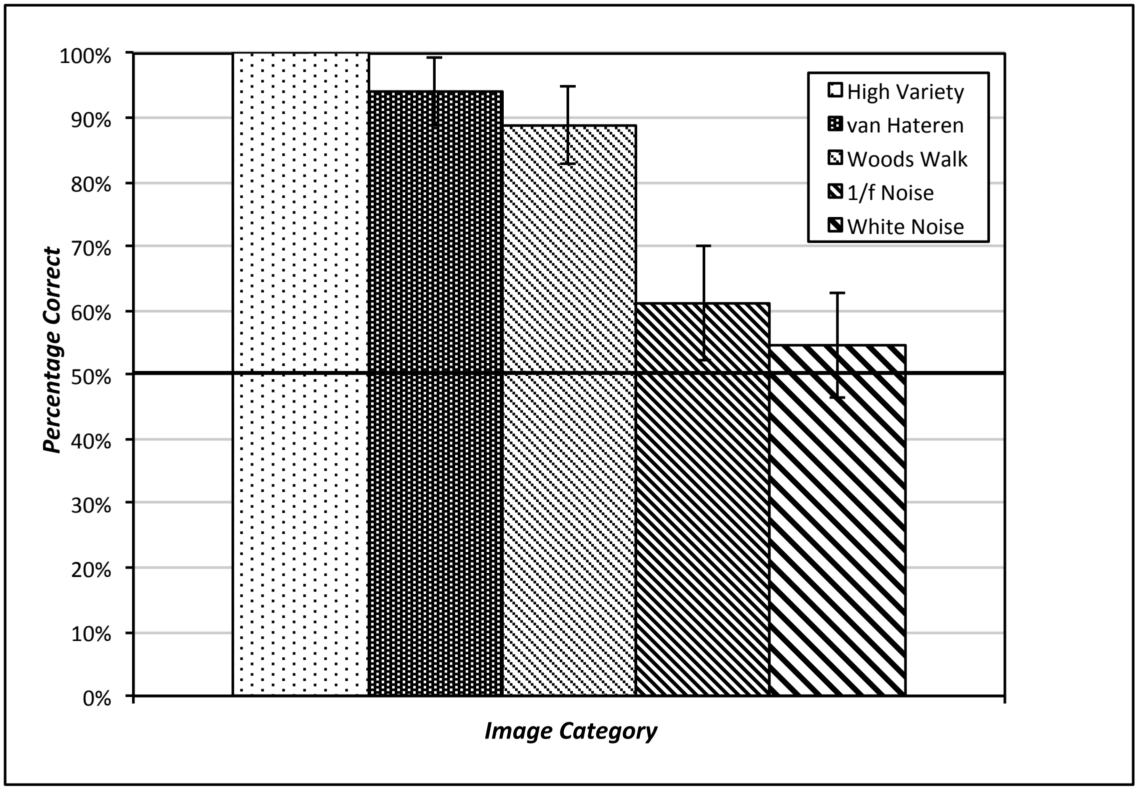


Figure 4-4: Memory performance as percentage of four images correctly recognized as previously seen. The error bars represent the 95% confidence interval around the mean.

Nor is the poor memory performance for white and 1/f noise caused by an inability to perceive the differences between images when they are both present for comparison. The results of Part 3, shown in Figure 4-5, clearly indicate that participants can discriminate among images. Performance for 1/f noise patterns was almost as good as performance for the Walk in the Woods and the van Hateren natural scenery images. Performance for white noise patterns, while considerably worse than the other categories, was still statistically much better than chance ($t(37) = 9.66$, $p < 0.0005$, $d = 3.18$).

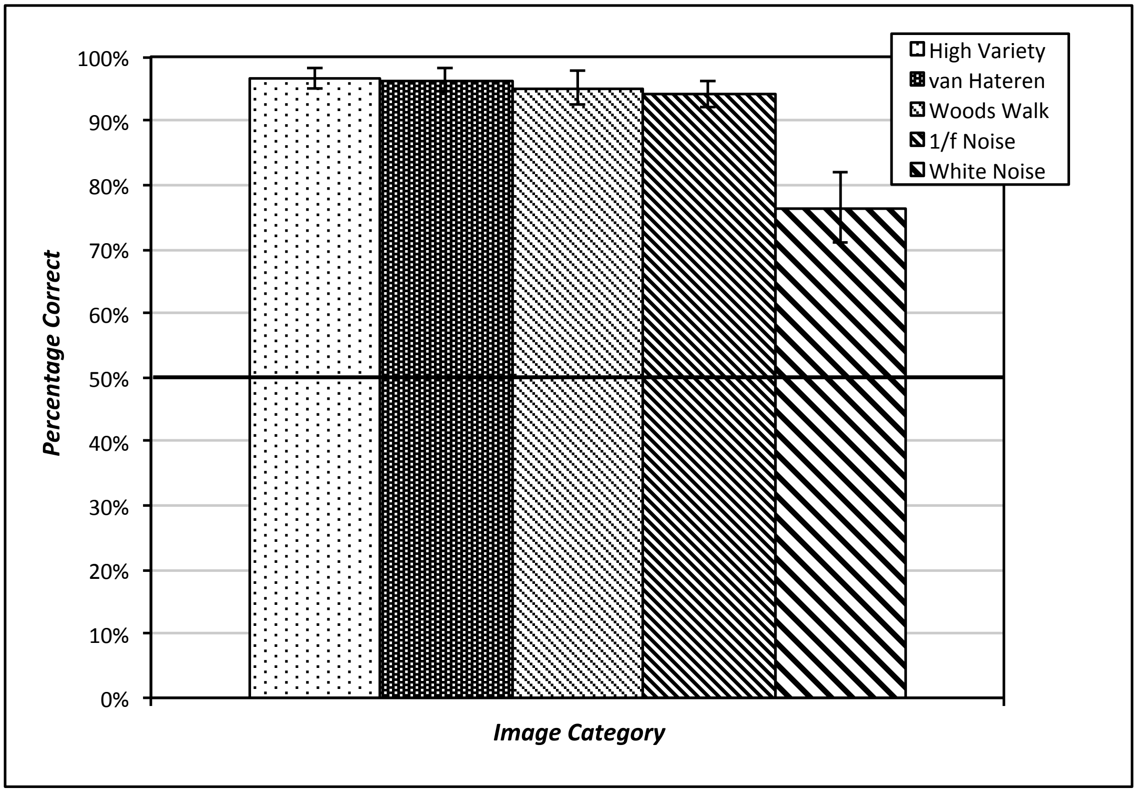


Figure 4-5: Memory performance as percentage of 20 pairs of images correctly categorized as same or different in side-by-side comparisons.

The error bars represent the 95% confidence interval around the mean.

2. Discussion

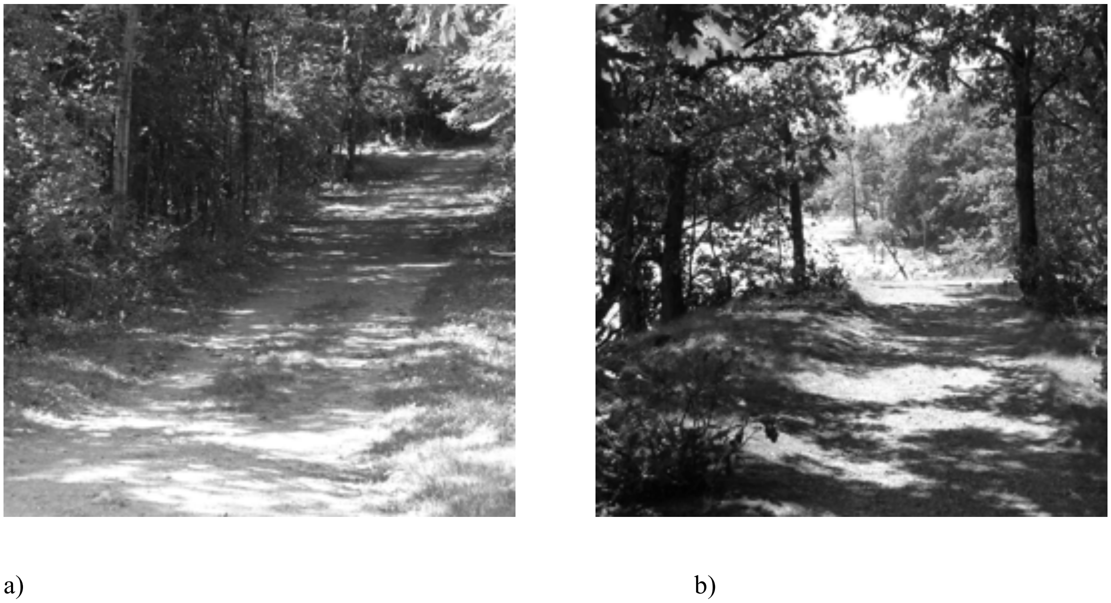
The results of this demonstration imply a number of points. Within an overall well-established context—the Walk in the Woods and van Hateren images—some scenes are producing an invariant response. This invariance, however, is occurring not at the level of online perception, but at the level of long-term memory formation. Scenes from these image categories are almost perfectly discriminable when they are both present to perception as shown in Part 3, and very discriminable in memory when an observer has only a few images to remember as shown in Part 1. As the images in memory accumulate, however, some become confused with other images already held. One implication we might draw from these results, therefore, is that the visual mechanisms that produce a context also reduce an observer's ability to store individual episodic memories of scenes that conform to that context.

It also appears that even if human observers do perceive all or most of the information available in an image, they cannot retain it long enough to discriminate among the images in long-term memory. Although a vector of pixel intensities provides a unique descriptor of a scene, sufficient to form its retinal image, and thus to discriminate any scene from any other, this information is not well-suited for describing contexts or forming long-term memories of scenes. The participants' poor performance for the white noise and 1/f noise picture categories, which are both rich in information, further demonstrate that this kind of image information is not very available when needed for recognition.

Anecdotally, participants reported that they had trouble performing on the Walk in the Woods and van Hateren scenes because the images were "too similar." There could potentially be measurable redundancy across the images in these categories that might lead to confusion. For example, we might expect to find higher pixel-to-pixel intensity correlations among the images if a number of them contained a dark ground plane and a lighter sky plane, or if they depicted a large focal object against a contrasting background. However, the average Pearson's product-moment correlation for the set of van Hateren images used in this experiment was only 0.06 and the average correlation for the Walk in the Woods images was 0.05, while the average correlation for the High Variety set was 0.02. Correlations among the natural scenery images could be low for any of a number of reasons, including translations of the objects and surfaces from one image to another, differences in illumination, and myriad differences in the shapes of fine-grained elements like twigs, leaves, and stones. During online perception, the visual system may be capable of perceiving small differences among these elements, but during memory formation, their shapes and textures may become more generalized. As Chapters 2 and 3 of this thesis demonstrated, the visual system has the ability to generalize a “what + where” response over a fairly large range of shape and location differences. Thus, the near-chance memory performance on the longer task for the Walk in the Woods and van Hateren categories indicates that information in these images is probably not available at recall any more than it is for the noise images, while the High Variety category, with a similar amount of information, elicits a much higher memory performance. The intuitive similarity that observers feel exists among the Walk in the Woods and van Hateren images of natural scenery based on their comments during debriefing, therefore, does not seem to translate into this kind of objective measure. These results reinforce the notion that whatever common elements comprise a context, they are probably at the level of recognizable scene structures rather than Shannon-type information.

The question arises whether we would find such poor memory for scenes from some other well-established context besides the natural scenery categories we used here. We think it is unlikely that participants in this experiment or others would have had as much exposure to any other single context as they would have had to the natural landscape context, except for their own work/life environments. But recall that our rationale for using the natural scenery categories was that they did not exhibit most of the biases that we contend are responsible for high memory performance in other picture memory experiments. The good results we obtained for our High Variety category in this demonstration reinforce the rationale. This category indisputably exhibits two characteristics that the other categories do not: it has more content and/or semantic diversity than the others, and that content is much more anthropocentric in nature. These characteristics, we believe, are likely much more influential in forming long-term picture memories than either information or mere visual diversity.

For example, the High Variety set used here contains only one image of an elephant. All other images in the set contain some other content (although there are a few content repetitions that differ in structure and visual details). This is the kind of diversity that has been used in most research on picture memory. The Walk in the Woods and van Hateren categories, on the other hand, contain a great deal of extremely similar content (bushes, trees, paths, etc.) but the content is structured differently in each image; that is, these categories have visual but not content diversity. Visual diversity has recently become more interesting to vision researchers, as Konkle *et al*. (2010a), Brady *et al*. (2008) and Vogt & Magnusson (2007) showed. They demonstrated that picture memory contains considerable visual detail. However, our results show that within a conceptual category, (i.e., "bushes and trees") where there is considerable room to remember visual details, memory performance is poor, at least among young observers of the modern age presumably raised in urban and suburban settings. Figures 4.6a and 4.6b show two examples of images from the Walk in the Woods category. Each can be described conceptually as "path through trees," and both are structured similarly, but the images vary greatly in the actual detail of tree and branch locations, shadow patterns, horizon line, angle of path, etc. These two images were also taken at widely spaced intervals along the path—the first is from near the trailhead, the last is from a location approximately 0.5 miles further along. Thus, the two images are not contiguous in either time or space, but are still similar enough to be confused in long-term memory.

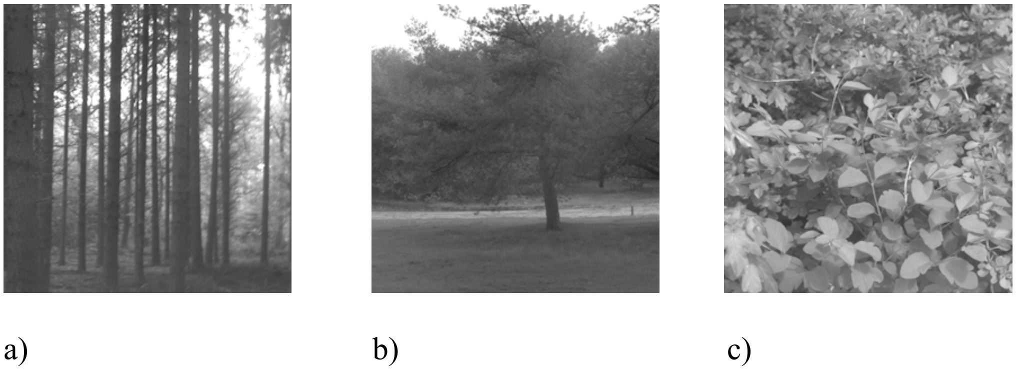


a) b)

Figures 4-6a and 4-6b: Two “Walk in the Woods” images.

These images show similar content and structure but considerable difference in visual details The images were taken from widely spaced points along the trail in Peebles Island State Park, Cohoes, New York.

Although we are calling the natural scenery categories a single context, we are doing so only in relationship to the many different contexts present in the High Variety category. According to our own definition, the natural scenery category must exhibit a number of individual contexts according to the objects and spatial structures they contain. Figure 4.7a, for example, shows a “multiple trees” context, Figure 4.7b a “single tree” context, and Figure 4.7c a “foliage close up” context, all from the van Hateren image category.



Figures 4-7a, 4-7b, and 4-7c: Three different contexts within the natural scenery category.

Anthropocentrism's effect on picture memory

The categorically distinct images used in the Konkle *et al*. (2010a) study demonstrate the power of an anthropocentric bias on image selection for picture memory experiments. Images used in many studies are composed of content that is of inherent interest to humans—animals, buildings, food, all kinds of human-made tools and objects, as well as scenes containing people. Any of this content is likely to draw the attention of a human observer. It is likely to be memorable because, above and beyond its visual interest, this kind of content is more relevant to human behavior and tasks than pictures of natural scenery. Because the visual system and memory evolved through experiencing this natural world, the tasks of foraging, mating, and finding shelter involved perceiving and remembering objects of behavioral interest in order for individuals to survive. The visual system may have become finely honed at finding such objects in what seems to modern urban- or suburban-raised observers an undifferentiated mass of foliage or other terrain. Thus, when confronted with a plethora of behaviorally interesting content, visual memory easily retains many thousands of scenes. Perhaps as a result of evolutionary pressures, a number of areas of visual cortex respond preferentially to objects, faces, and places rather than to surfaces and textures (Aguirre *et al*., 1998; Chao *et al*., 1999; Epstein & Kanwisher, 1998; Grill-Spector, 2003; Martin *et al*., 1996). It might be that more of the visual brain's resources are recruited to remember such content than other types of scenes and objects.

Examination of the scene categories used in the Konkle *et al*. (2010a) study showed that of the 16 image categories containing 64 exemplars each, 15 categories contained scenes that were primarily composed of human-made objects or human-built settings. For example, they used images of concert halls, shop fronts (with signage), gas stations, offices, train yards, and living rooms. In contrast, in our results, the two image categories without an anthropocentric bias (the Walk in the Woods and van Hateren categories) achieved only 60% and 59% performance respectively with 100 exemplars each. Our High Variety category, the only one with a degree of anthropocentric bias, achieved a similar level of performance to Konkle *et al*.'s at 88%. The difference between 100 exemplars of an anthropocentrically oriented scene category and 100 exemplars of unremarkable natural scenery may simply reflect an inherent difference in the interest and attention paid to the scenes when encoding into memory.

3. Conclusions

For most observers, their lifetime of visual experience consists of an enormous number of repetitions of the same content (spouse, children, pets, house, office, route to work, friends' houses, local businesses, etc.). An individual observer would find it difficult to experience as much content diversity in a year as he would encounter in a few minutes or hours of a typical picture memory experiment, yet from this type of research, vision researchers have drawn broad conclusions about the capacity of visual memory. We contend that some as-yet unidentified factor accounts for the huge capacity of visual memory in such experiments, but that the capacity of visual memory for the natural world in general is probably not nearly so large.

We have shown here that this factor is unlikely to be information in Shannon's sense. It is easy to say that the unidentified factor is semantic diversity or anthropocentrism. It is much more difficult to conceive how these attributes might be measured, and more significantly, compared across different kinds of pictures. We anticipate many questions regarding the usefulness of metrical properties of natural scenes as applied to long-term memory for scenes.

CHAPTER 5
CONCLUSIONS AND FUTURE DIRECTIONS

1. Summary of Findings

The work in this thesis was done in order to understand more about how visual context is formed and what effects it has on long-term memory for scenes. The results reported here give us information about the ease of learning spatial context, the potential difficulties of learning identity context, and the pronounced effects of well-established contexts on long-term memory for scenes.

In Chapter 2, we showed that visual context formation is quite sensitive to spatial learning. We were able to produce a "spatial violation" effect when objects appeared in unexpected regions of the scene after short periods of training where they only appeared within expected regions. The regions where objects could appear included both relatively large areas of the scene (a full quadrant) and very constrained regions (a single-object-width stripe, either vertical or horizontal). These findings illustrate that spatial constraints on context do not necessarily depend on semantic knowledge about the laws of physics, as Biederman *et al*. (1982) claimed, or other real-world knowledge, but can be learned directly from the probability of an object's appearance in implicitly constrained areas of the visual field. The ease and precision with which these constraints are learned also leads us to believe that visual contexts are relatively insensitive to changes in an object's apparent position due to the observer's common types of movements, leading to some degree of viewpoint invariance. At the same time, the marked effects of objects' appearing outside expected regions suggest to us that the visual system is very sensitive to changes in position caused by an object's self-movement (using the term "object" for any entity lulls us into thinking mostly of inanimate objects, but clearly animals and other humans are quite prone to self-movement).

In Chapter 3, the first experiment reproduced the well-known “inconsistent object” effect from a baseline using abstract objects. Like the experiments in Chapter 2, it shows that object inconsistency can be learned solely from the statistics of object co-occurrence rather than requiring semantic knowledge of the role an object plays in the real world. More importantly, however, the first experiment suggests that as a context is formed, it may become relatively tolerant of shape variation. With objects of the same size, the inconsistency effect did not emerge unless the shape plus a variety of other textural and contour attributes that contribute to a different visual style or appearance were varied. The second experiment confirmed that shape variation alone is probably not enough to drive the inconsistent object effect; spatial configuration predominated in the results, at least for circumstances where shape variation is relatively slight and the objects are the same size.

These results also tell us something about the tabulating properties of the visual system. They suggest, potentially, that not all instances of presentation to the visual system are treated alike, as is true of artificial algorithms. If scenes contain configurations of objects that seem to have a relationship based on visual similarity, it appears probable that the visual system discounts the perceivable shape differences (and perhaps other visual characteristics as well), preserving only the spatial configuration. Thus, it might not be true that observers learn co-occurrence statistics from every presentation of a given type of object. Unless those shape and other characteristics become important differentiators among scenes, then similar objects appearing in appropriate locations may have equal status for context formation. For example, it is easy to imagine that a scene containing a grey

Toyota Camry might be confused with a scene containing a grey Honda Accord in the same spatial location; important implications for eye-witness testimony and other memory confabulations can be readily inferred.

In Chapter 4, we saw how scenes with well-established contexts are quite difficult to remember, especially when the established context is devoid of objects that are of interest to human beings. The results of such a context are a set of memory traces that are quite similar to one another, perhaps to the extent that only pixel-level information could help the observer discriminate among them. We showed that this information is probably not available in the formation of long-term memories, although it may be available at the time of online perception.

The context we used is perhaps singular in human experience; that is, natural scenery without any human artifacts, animals, or other humans in it. This context might conceivably be the most difficult case for forming distinct long-term memories. However, this kind of content is what remains after any anthropocentric content has been eliminated, and moreover, is the material that the mammalian visual system evolved to perceive. Possibly the much-researched immense capacity of long-term visual memory for pictures is the evolutionarily honed result of struggling to perceive and remember the behaviorally relevant details within the context of natural scenery.

2. Future Directions

2.1 The role of visual distinctiveness in forming identity contexts

As we noted in Chapter 3, the visual system must tabulate the co-occurrence of objects in scenes, as evidenced by "inconsistent object" effects. It simply makes sense that observers inventory objects as part of scene comprehension (Hollingworth & Henderson, 2002; Hollingworth, Williams, & Henderson, 2001). We found difficulty in reproducing object identity contexts, however, for the reasons we discussed in previous sections. Clearly, the next step in this research should focus on formation of object identity context when the objects vary in size, but are still related in visual appearance. The second step should be to introduce a mix of objects of varying visual appearance in training a visual context.

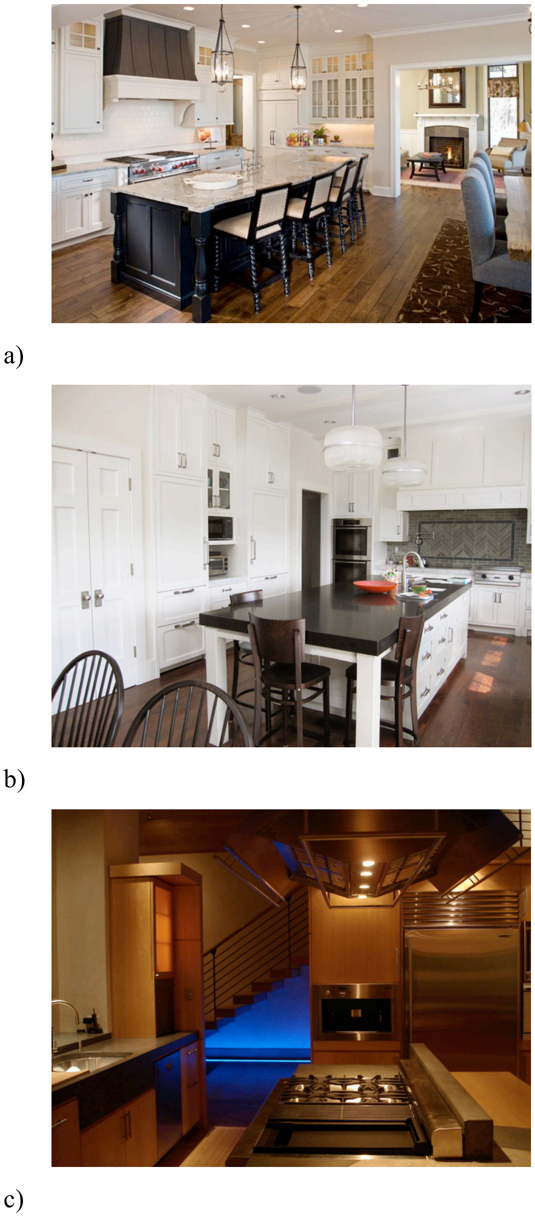
We should consider ways of quantifying the aspects of visual appearance that might cause the visual system to allocate more attention or memory resources to them. Similarly, we should construct experiments that test whether identity context relies more on semantic relations among objects and scenes than spatial context does.

2.2 Long-term memory across visual contexts

The major inference we have drawn from the results in this thesis is that well-established contexts negatively affect the ability to form detailed long-term memory traces of individual scenes. This inference was supported in one context, that of natural scenery without anthropocentric content. Although we argue that this context represents the baseline content of human vision in its evolutionary development, the inference cannot be accepted without showing its utility across a range of contexts. If statistical learning is the sole mechanism for establishing a context, the content of scenes should not matter. It might be the case, however, that memory for scenes with more human-oriented contents is innately better, perhaps because more visual brain areas are allotted to recognition of human-made objects and environments (Aguirre *et al*., 1998; Chao, Haxby, & Martin, 1999; Epstein & Kanwisher, 1998; Grill-Spector, 2003; Martin *et al*., 1996). In that case, we should find that even if the initial performance rate is higher relative to that for natural scenery, performance should decline as the set of scenes to be remembered increases in size. In other words, we might find that observers can remember 100 scenes of an alligator in a bathtub with a higher rate of accuracy than 100 scenes of trees and bushes, but by the time they have seen 1000 similar alligator/bathtub scenes, their performance will have degraded to near chance. The best evidence for the statistical learning hypothesis as it applies to visual contexts would be to find similar declining curves for memory performance across a wide range of contexts.

The gold standard test for well-established contexts would, of course, be taken from an individual observer's own physical environment. One's personal rooms, workplace, and recreational sites could provide scenes that would include both human-oriented and natural scenery, avoiding the criticism that natural scenery environments may not be easily remembered by modern observers. It should be possible to gather sufficient material from participants, although it might be difficult to equate scenes for similarity of content and complexity.

Clearly, in designing future experiments to test this hypothesis, we will encounter problems of experimenter bias in selecting the scenes. It will be necessary to define what counts as an instance of the context in question. A fair test of visual context is one in which similar-looking objects occupy similar positions in the scene. A test in which the scenes are semantically categorized as belonging to the context but varying widely in viewpoint, object appearance, and spatial layout would not test the hypothesis. The photographs of kitchens below illustrate the problem.


a)

b)

c)

Figures 5-1a, 5-1b, and 5-1c: Differences between visual and functional contexts.

The top two photographs show kitchens with similar visual contexts, while Figure 5.1c shows a kitchen with semantic, functional similarity but dissimilar visual context.

Figure 5.1a and 5.1b show two kitchens photographed from approximately the same viewpoint, with appliances, cabinets and other surfaces of similar luminance values. The appointments of these two kitchens differ substantially in appearance, but they resemble each other more than the appointments of the kitchen in Figure 5.1c. This kitchen is photographed from a different viewpoint, has much darker surfaces, and appointments that are visually different from those in the first two photographs. All three photographs could be included in the semantic category "kitchens" but only the first two could plausibly test the statistical learning of context hypothesis. We contend that observers learn to include scenes that are functionally but not visually similar in their semantic categories; their interactions with such environments allow them to infer functional equivalence beyond the purely visual categorization of scenes. If future experiments are not controlled for this functional bias, we might find misleadingly high memory performance in some context categories.

While increasing familiarity with a context may decrease observers' ability to remember individual scenes, other mechanisms may operate to offset the decrease. For example, *visual expertise* develops alongside contextual familiarity. Through prolonged exposure to multiple examples of a given kind of content, observers can learn to hone their perceptions of subtle differences among the examples. What appears to the novice birder as an undifferentiated LBB ("little brown bird") is easily identified by the expert birder or ornithologist as a song sparrow by its conformation and plumage. Visual expertise probably accrues from directing attention to details of shape, color, and texture that novices ignore as they make broader, categorical judgments. This kind of expertise may develop automatically as observers spend time with examples, but almost certainly it develops when discrimination among examples is useful to the particular observer. Whether the expertise is in breeding animals, authenticating Old Master paintings, or identifying terrorist activity, the need to pay attention to details works at cross purposes to the effects of context formation. One future direction for this work might be to track both effects in a naïve population, to determine which predominates over the other. Another interesting question is whether scenes that are differentiated by fine details are as durable in memory as scenes of more broadly different content. Even if details can be perceived well enough to make the discrimination, we would like to know if they have a lasting effect on memory.

This suggestion leads to a consideration of what content there is to perceive in the average observer's world. We propose that there is less visual variety in that world than we like to think, apart from the creations of art and entertainment technologies. An observer cannot instantly change from one environment to another, except in relatively unusual situations (the obvious exception is moving from an indoor to an outdoor environment, or vice versa). As a result, she spends most of her waking hours in just a few environments where she has probably lived most of her life: a particular dwelling, neighborhood, and office; familiar shopping areas, parks or recreation facilities; friends' and acquaintances' homes and offices; etc. These environments are where most of her visual contexts form, and they probably change little from year to year. If the observer travels, she exposes herself to unfamiliar environments and new visual contexts, but the number of hours she spends in them, and thus the strength of these new contexts, can never equal the strength of the contexts from her familiar environments. If these are indeed the conditions under which most long-term visual memories are formed, what does that imply about the distinct, highly memorable scenes that have been well-tested by picture memory research? Possibly two forms of memories may exist side by side: generalized, archetypical scenes of the observer's familiar environments, which may equate to the schemata or frames suggested by many researchers over the years, and individually memorable scenes that depart in some as-yet undefined way from the firmly established contexts. This suggestion deserves further research.

3. Conclusions

The points of empirical evidence gathered in the experiments of the preceding chapters can be placed in a narrative about how visual contexts form. As observers move around the environment, the anatomy and physiology of the eyes, head, neck, and trunk dictate that most glimpses of the environment are viewed from a limited number of viewpoints. Further, research has shown that observers habitually scan the environment directly ahead of them most of the time, while they report much less often looking to either side (Wagner, Baird, & Barbaresi, 1981). Thus, the overlapping patches of foveated vision from fixations that make up the visual field are taken from a relatively constrained set of viewpoints; though this set may be large, it is much lower in dimension than the possible set of viewpoints. Each scene patch containing part of the structure of a scene is repeatedly encountered multiple times during the period the observer spends in the environment. Obviously, some viewpoints are repeated more often. Equally obviously, no scene patch can be an exact replica of another: alignment of the observer's eyes, head, neck, and torso is unlikely to be identical. Meanwhile, illumination on the scene changes moment by moment, with time of day or year, as well as cloud, wind, and animal movements. Seasonal changes due to growth, flowering, fruiting, ripening, etc. change the look of the landscape. Shapes of objects in the environment change through the same causes. Animals in the environment are obviously usually not still. Even if they are recurrently present in a scene, they might change appearance due to growth and maturation.

Many of these changes may be incidental to laying down memory traces of the scene. Some are so fine-grained or gradual as to go unnoticed—the illumination on a particular leaf diminishing by a few lux, or the shade of green in a field of ripening cereal grass from one day to the next. Likewise, small changes in the positions of a previously seen object (the angle of the pen left on the desktop) and the shape changes induced by changes in the observer's viewpoint (the greater reveal of the edge of the laptop as one shifts in one's chair) may be perceptually noted but unimportant to memory. If each instance of a scene patch is sufficiently like another to activate the same set of neural responses, it reinforces the invariant response and thus the context of that scene.

From the beginnings of modern psychology, memory theorists have agreed that human memory does not consist of a complete, veridical record of an observer's every waking moment (Bartlett, 1932; Neisser, 1982). Instead, they posit that VLTM is *constructive*, rather than *reproductive* (Schacter & Addis, 2007; Schacter, Norman, & Koutsaal, 1998). The thinking goes that, because the future will probably resemble the past, there is little need to record every detail of each scene. According to this view, visual memory retains the elements of scenes that are most useful in constructing perceptions and interpretations of future scenes. One benefit of a constructive visual memory is to keep VLTM from filling up with uninformative near-copies of the same visual content, consolidating individual memories and keeping room for more distinctive autobiographical memories. Another obvious function of a contextually based visual memory is as a sort of alerting mechanism, to keep the observer continuously aware of what is normal in his environment and when opportunities or threats arise. The findings we have presented here contribute to our understanding of precisely what the useful elements needed for a constructive theory of visual memory are, and the constraints on their structure in scenes.

Finally, as the work in this thesis has demonstrated, these potential benefits are simply an outgrowth of the way that visual contexts are formed. The distinctive, unusual, “out-of-context” scene is always a function of what the individual observer has already seen and incorporated in his long-term memory. This fact implies that even the most distinctive, beautiful, meaningful, *memorable* scene an individual has ever viewed has that status only as long as he has seen it just once. The highly plastic nature of visual contexts guarantees that a thousand viewings of the Mona Lisa will reduce the chance of remembering a single representation of her enigmatic smile to a throw of the dice.

REFERENCES

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). An area within human ventral cortex sensitive to 'building' stimuli: Evidence and implications. *Neuron*, 21, 373-383.

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, 78(4), 266-278.

Auckland, M. E., Cave, K. R., & Donnelly, N. (2007). Nontarget objects can influence perceptual processes during object recognition. *Psychonomic Bulletin & Review,* 14 (2), 332-337.

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

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J, Schmiddt, A. M., Dale, A. M., Hamalainen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R., & Halgren, E. (2006), Top-down facilitation of visual recognition. *PNAS*, 103(2), 449–454.

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology.* Cambridge: Cambridge University Press.

Becker, M. W., Pashler, H., & Lubin, J. (2007). Object-intrinsic oddities draw early saccades. *Journal of Experimental Psychology: Human Perception and Performance*, 33(1), 20-30.

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy and J. R. Pomerantz, (Eds.) *Perceptual Organization*. Hillsdale, NJ: Erlbaum.

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

Brady, T. F. & Chun, M. M. (2007). Spatial constraints on learning in visual search: Modeling contextual cuing. *Journal of Experimental Psychology: Human Perception and Performance*, 33(4), 798–815.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Science,* 105(38), 14325–14329.

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

Franks, J. J., & Bransford, J. D. (1971). Abstraction of visual patterns. *Journal of Experimental Psychology*, 90(1), 65.

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.

Brockmole, J. R., & Henderson, J. M. (2008). Prioritizing new objects for eye fixation in real world scenes: Evidence from eye movements. *Visual Cognition*, 16(2/3), 375-390.

Bruce, N. D. B., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):5, 1–24, http://journalofvision.org/9/3/5/, doi:10.1167/9.3.5.

Castelhano, M. S. & Henderson, J. M. (2005). Incidental visual memory for objects in scenes. *Visual Cognition*, 12 (6), 1017-1040.

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience*, 2, 913-919.

Chua, K.-P. & Chun, M. M. (2003). Implicit scene learning is viewpoint dependent. *Perception & Psychophysics*, 65 (1), 72-80.

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

Chun, M. M. & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, 10(4), 360-365.

Chun, M. M. & Jiang, Y. (2003). Implicit, long-term spatial contextual memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(2), 224-234.

Conway, C. M., Goldstone, R. L., & Christiansen, M. H. (2007). Spatial constraints on visual statistical learning of multi-element scenes. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 185-190). Austin, TX: Cognitive Science Society.

Coppola, D. M., Purves, H. R., McCoy, A. N., & Purves, D. (1998). The distribution of oriented contours in the real world. *Proceedings of the National Academy of Sciences*, 95, 4002-4006.

Davenport, J. (2007). Consistency effects between objects in scenes. *Memory & Cognition*, 35(3), 393-401.

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

Dill, M. & Edelman, S. (2001). Imperfect invariance to object translation in the discrimination of complex shapes. *Perception*, 30, 707-724.

Dill, M. & Fahle, M. (1998). Limited translation invariance of human visual pattern recognition. *Perception & Psychophysics*, 60 (1), 65–81.

Diwadkar, V. A. & McNamara, T. P. (1997). Viewpoint dependence in scene recognition. *Psychological Science*, 8, 302-307.

Edelman, S., Intrator, N., & Jacobson, J. S. (2002). Unsupervised learning of visual structure. *Lecture Notes in Computer Science*, 2525, 165-178.

Edelman, S. (1999). *Representation and recognition in vision.* Cambridge: MIT Press.

Endo, N. & Takeda, Y. (2004). Selective learning of spatial configuration and object identity in visual search. *Perception & Psychophysics*, 66 (2), 293-302.

Endo, N. & Takeda, Y. (2005). Use of spatial context is restricted by relative positive in implicit learning. *Psychonomic Bulletin & Review*, 12(5), 880-885.

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

Fei-Fei, L, Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4), 594-611.

Fei-Fei, L., VanRullen, R., Koch, C., & Perona, P. (2005). Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli. *Visual Cognition*, 12(6), 893-924.

Fergus, R., Fei-Fei, L., Perona, P., & Zisserman, A. (2010). Learning object categories from Internet image searches. *Proceedings of the IEEE, Special Issue on Internet Vision*, 98(8), 1453-1466.

Field, D. J. (1987). Relations between the statistics of natural images and the response profiles of cortical cells. *Journal of the Optical Society of America A,* 4, 2379-2394.

Field, D. J., Hayes, A., & Hess, R. (1993). Contour integration by the human visual system: evidence for a local 'association field.' *Vision Research*, 33(2), 173-193.

Fiser, J. & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 28(3), 458-467.

Fiser, J. & Aslin, R. N. (2005). Encoding multielement scenes: statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General*, 134(4), 521–537.

Franconeri, S. L, Hollingworth, A., & Simons, D. J. (2005). Do new objects capture attention? *Psychological Science*, 16(4), 275-281.

Franks, J. J. & Bransford, J. D. (1971). Abstraction of visual patterns. *Journal of Experimental Psychology*, 90(1), 65-74.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, 108(3), 316-355.

Gao, Z., Li, J., Liang, J., Chen, H., Yin, J., & Shen, M. (2009). Storing fine detailed information in visual working memory—Evidence from event-related potentials. *Journal of Vision*, 9(7):17, 1–12, http://journalofvision.org/9/7/17/, doi:10.1167/9.7.17.

Gajewski, D. A. & Henderson, J. M. (2005). Minimal use of working memory in a scene comparison task *Visual Cognition*, 12 (6), 979-1002.

Garsoffky, B., Schwan, S., & Hesse, F. W. (2002). Viewpoint dependency in the recognition of dynamic scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(6), 1035–1050.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology,* 59, 167–192.

Gentner, D. & Loftus, E. F. (1979). Integration of verbal and visual information as evidenced by distortions in picture memory. *American Journal of Psychology*, 92(2), 363-375.

Goujon, A., Didierjean, A., & Marmeche, E. (2007). Contextual cueing based on specific and categorical properties of the environment. *Visual Cognition*, 15(3), 257-275.

Graham, D. J., Chandler, D. M., & Field, D. J. (2008). Can the theory of ‘‘whitening’’ explain the center-surround properties of retinal ganglion cell receptive fields? *Vision Research*, 46, 18, 2901-2913.

Greene, M.R., & Oliva, A. (2009a). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, 20(4), 464-472.

Greene, M.R., & Oliva, A. (2009b). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, 58(2), 137-179.

Grill-Spector, K. (2003). The neural basis of object recognition. *Current Opinion in Neurobiology*, 13, 1-8.

Grill-Spector, K. & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science*, 16(2), 152-160.

Hansen, B. C. & Essock, E. A. (2004). A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes. *Journal of Vision*, 4, 1044-1060.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *TRENDS in Cognitive Sciences*, 7(11), 498-504.

Hollingworth, A. (2005). The relationship between online visual representation of a scene and long-term scene memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3), 396-411.

Hollingworth, A. (2005). Memory for object position in natural scenes. *Visual Cognition*, 12(6), 1003-1016.

Hollingworth, A. (2006). Scene and position specificity in visual memory for objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(1), 58-69.

Hollingworth, A. & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, 127(4), 398-415.

Hollingworth, A. & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 28(1), 113–136.

Hollingworth, A, Williams, C. C., & Henderson, J. M. (2001). To see and remember: Visually specific information is retained in memory from previously attended objects in natural scenes. *Psychonomic Bulletin & Review*, 8(4), 761-768.

Intraub, H. & Richardson, M. (1998). Wide-angle memories of close-up scenes. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15(2), 179-187.

Irwin, D. E. & Zelinsky, G. J. (2002). Eye movements and scene perception: Memory for things observed. *Perception and Psychophysics*, 64(6), 882-895.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems*, 18, 1-8.

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

Jiang, Y., Olson, I. R., & Chun, M. M. (2000). Organization of visual short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(3), 683-702.

Jiang, Y. & Song, J.-H. (2005). Hyperspecificity in visual implicit learning: Learning of spatial layout is contingent on item identity. *Journal of Experimental Psychology: Human Perception and Performance*, 31(6), 1439–1448.

Jiang, Y. & Wagner, L. C. (2004). What is learned in spatial contextual cuing—configuration or individual locations? *Perception and Psychophysics*, 66(3), 454-463.

Joseph, J. S., Chun, M. M., & Nakayama, K. (1997). Attentional requirements in a 'preattentive' feature search task. *Nature*, 387, 805-807.

Joubert, O. R., Fize, D., Rousselet, G. A., & Fabre-Thorpe, M. (2008). Early interference of context congruence on object processing in rapid visual categorization of natural scenes. *Journal of Vision*, 8(13):11, 1–18.

Joubert, O. R., Rousselet, G. A., Fabre-Thorpe, M. & Fize, D. (2009). Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise. *Journal of Vision,* 9(1):2, 1–16

Judd, T., Durand, F., & Torralba, A. (2011). Fixations on low-resolution images. *Journal of Vision*, 11(4):14, 1-20.

Juttner, M. & Rentschler, I. (2008). Category learning induces position invariance of pattern recognition across the visual field. *Proceedings of the Royal Society B*, 275, 403-410.

Karacan, H. & Hayhoe, M. M. (2007). Is attention drawn to changes in familiar scenes? *Visual Cognition*, 16(2), 356-374.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, 46(11), 1762-1776.

Kobatake, E. & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, 71(3), 856-867.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010a). Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological Science*, 21(11), 1551-1556.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010b). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, 139(3), 558-578.

Kravitz, D. J., Kriegeskorte, N., & Baker, C. I. (2010). High-level visual object representations are constrained by position. *Cerebral Cortex*, 20(12): 2916-2925.

Kravitz, D. J., Vinson, L. D., &Baker, C. I. (2008). How position dependent is visual object recognition? *Trends in Cognitive Sciences*, 12(3), 114-122.

Levin, D. T., Simons, D. J., Angelone, B. J., & Chabris, C. F. (2002). Memory for centrally attended changing objects in an incidental real-world change detection paradigm. *British Journal of Psychology*, 93, 289–302.

Li, B., Peterson, M R., & Freeman, R. D. (2003). Oblique effect: A neural basis in the visual cortex. *Journal of Neurophysiology,* 90, 204-217.

Li, L.-J., Su, H., Lim, Y. & Fei-Fei, L. (2010). Objects as attributes for scene classification. *Trends and Topics in Computer Vision: Lecture Notes in Computer Science*, 6553, 57-69.

Li, N. & DiCarlo, J. J. (2008). Invariant object representation in visual cortex. *Science,* 321(5895), 1502-1507.

Lleras, A., Rensink, R. A., & Enns, J. T. (2005). Rapid resumption of interrupted visual search: New insights on the interaction between vision and memory. *Psychological Science*, 16(9), 684-688.

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), 565-572.

Logothetis, N. K., Pauls, J. & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5), 552-563.

Luck, S. J. & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(20), 279-281.

Mack, M. L. & Palmieri, T. J. (2010). Modeling categorization of scenes containing consistent versus inconsistent objects. *Journal of Vision*, 10(3):11.

Maljkovic, V. & Nakayama, K. (1994). Priming of popout: I. Role of features. *Memory and Cognition,* 22, 657-672.

Mandler, J. M. & Johnson, N. S. (1976). Some of the thousand words a picture is worth. *Journal of Experimental Psychology: Human Learning*, 2(5), 529-540.

Mandler, J. M. & Parker, R. E. (1976). Memory for descriptive and spatial information in complex pictures. *Journal of Experimental Psychology: Human Learning and Memory,* 2(1), 38-48.

Mandler, J. M. & Ritchey, G. H. (1977). Long-term memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory,* 3(4), 386-396.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY: Henry Holt and Co. Inc.

Martin, A., Wiggs, C. L., Ungerleider, L. G., & Haxby, J. V. (1996). Neural correlates of category-specific knowledge. *Nature*, 379, 649-652.

McCotter, M., Gosselin, F., Sowden, P., & Schyns, P. (2005). The use of visual information in natural scenes. *Visual Cognition*, 12 (6), 938-953.

McNamara, T. P., Rump, B., & Werner, S. (2003). Egocentric and geocentric frames of reference in memory of large-scale space. *Psychonomic Bulletin & Review*, 10(3), 589-595.

Melcher, D. (2006). Accumulation and persistence of memory for natural scenes. *Journal of Vision*, 6, 8–17.

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

Neider, M. B. & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, 4, 614-621.

Neisser, U. (1982). Memory: What are the important questions? In U. Neisser & I. Hyman (Eds.), *Memory Observed: Remembering in Natural Contexts,* pp. 3-18. San Francisco: Freeman.

Newell, F. N., Sheppard, D. M., Edelman, S., & Shapiro, K. L. (2005). The interaction of shape- and location-based priming in object categorization: Evidence for a hybrid “what + where” representation stage. *Vision Research*, 45, 2065-2080.

Nickerson, R. S. (1965). Short-term memory for complex meaningful visual configurations: A demonstration of capacity. *Canadian Journal of Psychology/Review of Canadian Psycholology*, 19(2), 155-160.

Nickerson, R. S. (1968). A note on long-term recognition memory for pictorial material. *Psychonomic Science*, II, 58.

Oliva, A. (2004). Gist of the scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Encyclopedia of Neurobiology of Attention,* (pp. 251-256). San Diego, CA: Elsevier.

Oliva, A. & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41, 176–210.

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.

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

Olson, I. R., Jiang, Y., & Moore, K. S. (2005). Associative learning improves visual working memory performance. *Journal of Experimental Psychology: Human Perception and Performance*, 31(5), 889-900.

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

Op de Beeck, H., Wagemans, J., & Vogels, R. (2001). Macaque inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience,* 4, 1244–1252.

O'Regan, J. K. & Noe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5), 939-1031.

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, 9(4), 441-474.

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.

Pezdek, K., Maki, R., Valencia-Laver, D., Whetstone, T., Stoeckert, J., & Dougherty, T. (1988). Picture memory: Recognizing added and deleted details. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3), 468-476.

Pezdek, K., Whetstone, T., Reynolds, K., Askari, N., & Dougherty. T. (1989). Memory for real-world scenes: The role of consistency with schema expectation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(4), 587-595.

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

Poggio, T., Fahle, M., & Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science*, 256, 1018-1021.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, 2(5), 509-522.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, 81(1), 10-15.

Potter, M. C., Staub, A., & O'Connor, D. H. (2004). Pictorial and conceptual representation of glimpsed pictures. *Journal of Experimental Psychology: Human Perception and Performance*, 30(3), 478-489.

Rao, S. C., Rainer, G., & Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. *Science*, 276, 821-824.

Reich, D.S. Mechler, F. & Victor, J.D. (2000). Formal and attribute-specific information in primary visual cortex. *Journal of Neurophysiology*, 57, 132–146.

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

Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get the “gist” of real-world natural scenes? *Visual Cognition*, 12(6), 852-877.

Russell, B., 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.

Sanocki, T. (2003). Representation and perception of scenic layout. *Cognitive Psychology*, 47, 43–86.

Sanocki, T. & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science*, 8(5), 374-378.

Schacter, D. L. & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B*, 362, 773-786.

Schacter, D. L., Norman, K. A., & Koutsaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology,* 49, 289–318.

Schmidt, B. K., Vogel, E. K., Woodman, G. F., & Luck, S. J. (2002). Voluntary and automatic attentional control of visual working memory. *Perception and Psychophysics*, 64(5), 754-763.

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.

Shannon, C. E. (1948). A mathematical theory of communication. Reprinted with corrections from *The Bell System Technical Journal*, 27, 379–423, 623–656.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6, 156-163.

Simoncelli, E.P. & Olshausen, B.A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24, 1193–1216.

Simons, D. J. & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin and Review,* 5(4), 644-649.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology*, 25, 207-222.

Standing, L., Conezio, J., & Haber, R. N. (1970). Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science*, 19(2), 73-74.

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

Strong, E. K. (1912). The effect of length of series upon recognition memory. *Psychological Review*, 1912, 19, 447-462.

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

Torralba, A. (2003b). Modeling global scene factors in attention. *Journal of the Optical Society of America A*, 20(7), 1407-1418.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. M. (2006) Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review,* 113(4), 766–786.

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

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

Tversky, T., Geisler, W. S., & Perry, J. S. (2004). Contour grouping: Closure effects are explained by good continuation and proximity. *Vision Research*, 44, 2769-2777.

van Hateren, J.H. & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in visual cortex. *Proceedings of the Royal Society B*, 265(1394), 359-366.

Vidal, J. R., Gauchou, H. L., Tallon-Baudry, C., & O'Regan, J. K. (2005). Relational information in visual short-term memory: The structural gist. *Journal of Vision*, 5(3):8.

Viera, C. L. & Homa, D. L. (1991). Integration of nonthematic details in pictures and passages. *American Journal of Psychology,* 104(4), 491-516.

Vinje, W.E. & Gallant, J.L. (2002). Natural stimulation of the nonclassical receptive field increases information efficiency in V1. *Journal of Neurosci*ence, 22, 2904–2915.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 27(1), 92-114.

Vogt, S. & Magnussen, S. (2007). Long-term memory for 400 pictures on a common theme. *Experimental Psychology*, 54(4), 298–303.

Wagner, M., Baird, J. C., & Barbaresi, W. (1981). The locus of environmental attention. *Journal of Environmental Psychology*, 1, 195-206.

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

Wheeler, M. E. & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General*, 131(1), 48-64.

Wolfe, J. M. (2003). Moving towards solutions to some enduring controversies in visual search. *TRENDS in Cognitive Sciences*, 7(2), 70-76.

Wolfe, J. M. & Bennett, S. C. (1996). Preattentive object files: Shapeless bundles of basic features. *Vision Research*, 37(1), 25-43.

Zimmerman, E., Schnier, F., & Lappe, M. (2010). The contribution of scene context on change detection performance. *Vision Research*, 50, 2062–2068.

APPENDIX A
ADDITIONAL EXPERIMENT IN SPATIAL CONTEXT FORMATION

A.1 Introduction

The experiments described in Chapter 2 gave us some evidence that spatial constraints are easily learnable; observers are sensitive to constraint violations even when the spatial displacements outside the learned constraints are quite small in actual extent. Because of this evidence, we sought to test this ability in an even more difficult task.

Some objects in the natural environment can appear in more than one constrained area. Examples might include animals or birds that spend time in both arboreal and ground habitats. Some human-made objects are frequently stored in one area or orientation, but used in another (garden tools hanging on a garage wall vs. being handled by a gardener, or books stored on a shelf but open on a desk). A violation of such multiple spatial constraints might be more difficult for an observer to spot, but could be within their capabilities, based on the results in our previous experiments. We designed the following experiment to test participants' ability to learn two spatially constrained areas for a single probe object, where displacements outside the area were of relatively small extent.

A.2 Experiment 6

The method for this experiment was similar to that of Experiment 1 in Chapter 2.

A.2.1 Participants

Twenty Cornell University undergraduates participated in this experiment. All had normal or corrected-to-normal vision. They received course credit for their participation. All participants gave their informed consent.

A.2.2 Stimuli

The stimuli for this experiment were artificial scenes constructed using the same black abstract shapes generated by the Op de Beeck, Wagemans, & Vogel (2001) algorithm as were used in the experiments in Chapters 2 and 3. The scenes consisted of five objects each, in which each object was 50 x 50 pixels, subtending approximately 1.4° of visual angle. The objects appeared in randomly chosen positions within the scene, subject to the constraints described below.

The spatial configuration of each scene was generated by a MATLAB 7.0 algorithm using the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). The objects occupied a 750 x 750 pixel white square centered on the screen, subtending approximately 21.5° of visual angle, surrounded by a medium gray background extending to the edges of the screen. Two spatial context areas were chosen within the white square. A single probe object could appear in either one of these areas. Area 1 was a 150 x 150 square in the upper right quadrant of the screen, while Area 2 was of the same dimensions in the lower left quadrant. Each spatial context area was positioned within its quadrant so that a 100-pixel margin bordered each of its sides.

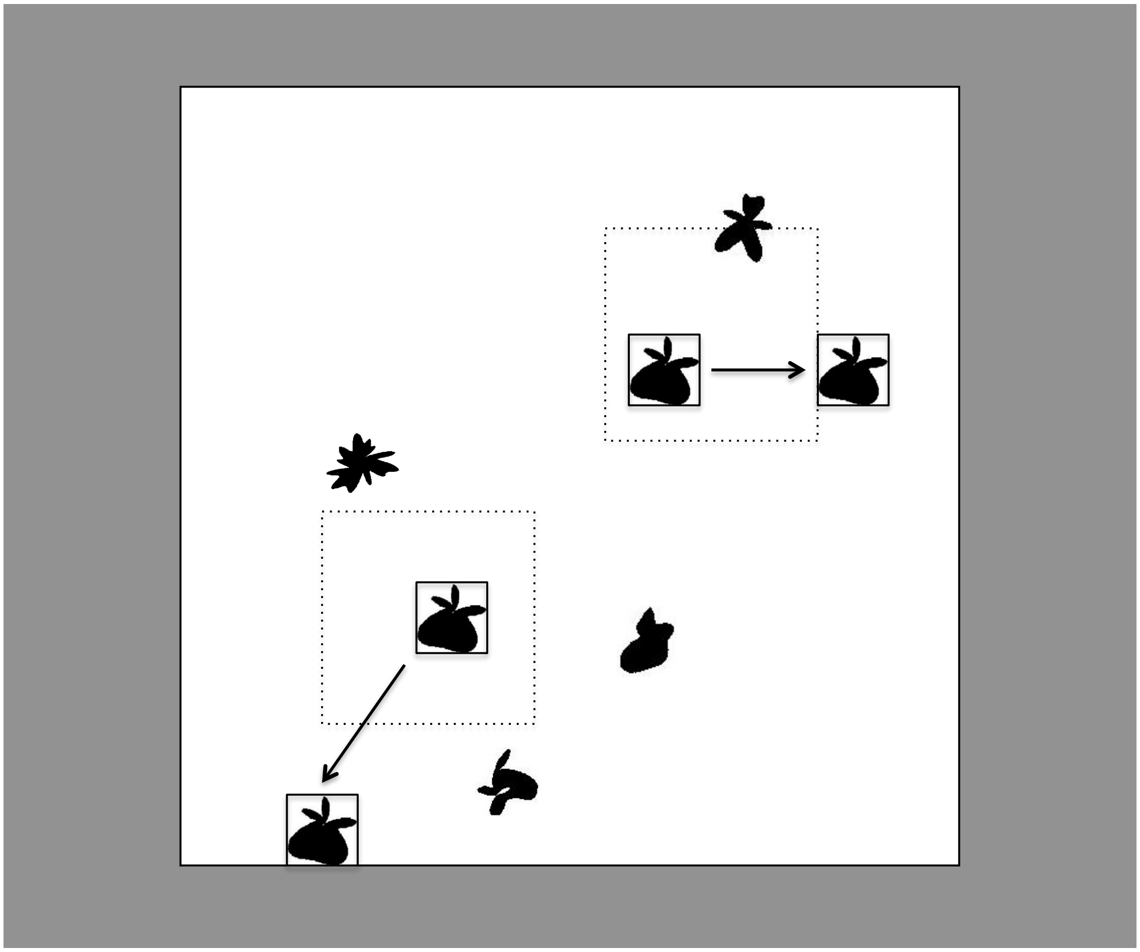


Figure A-1: Translations of the single probe object for Experiment 6.

Small (upper right) and large (lower left) translations of a single probe object that could appear in either one of two bounded areas, indicated by the dotted squares. (Boundaries are shown for illustrative purposes only.)

The scenes were displayed on a 17-in. (43.2 cm) Dell computer monitor with a resolution of 1024 x 768 pixels. The monitor was viewed from a distance of approximately 25 in. (63.5 cm) in a normally lighted office, so that the entire viewing area from screen edge to edge subtended approximately 24.3º of height and 32.3º of width. The experimental protocol was controlled using MATLAB 7.0.

Scenes were presented for 2 s, followed by a 1-s mask consisting of a 750 x 750 pixel square of blurred white noise. A medium gray screen appeared for 0.5 s before the start of a new trial.

A.2.3 Procedure

The experiment consisted of 276 trials with a delayed match to sample task. The scenes were presented with one intervening scene between each sample and test pair, so that each intervening scene became the sample for the next test scene. Thus, participants always had to remember at least one additional scene before they responded to a test scene. In test scenes, a red square surrounded the probe object. The response measure was a forced choice to the question, "have you seen *that* object in *that* location?"

Participants read instructions for the experiment on screen, then completed a short practice session before beginning the main experiment. The practice session included one trial of each of the experimental conditions. Participants indicated their yes-no response with a keypress. They were permitted to view the test scene as long as they wanted before responding.

The experiment began with a training phase consisting of 168 scenes. The single probe object could appear equally often in each of the two bounded areas. In the training phase, the constrained probe object was translated only within one of its bounded areas from sample scene to test scene. A second, unconstrained object also served as a probe object in one-third of the scenes; it could be translated in any direction from its original position by either 50 or 100 pixels. All scenes were equally divided into 'move' and 'no-move' trials.

A brief rest period followed the training, after which participants were told to continue with the experiment. In the test phase, the probe objects could be translated outside their bounded areas, although participants were not informed of this fact. The test phase consisted of an additional 180 trials, also divided equally between ‘move’ and ‘no-move’ trials. Within ‘move’ trials, there were equal numbers of moves within the constrained areas and outside the constrained areas for each probe object. (The unconstrained probe’s moves were equally often of smaller and larger extent.)

The unconstrained probe object served as a control object; its initial positions were randomly selected anywhere within the scene area. The other three objects were also randomly positioned in the scene area; they never overlapped with the other objects, although they were permitted to appear in either of the constrained object’s bounded areas.

A.2.4 Results

A mixed-model regression analysis performed on the *d’* data (where *participant* was entered as a random factor and *area* [Area 1, Area 2, and Unconstrained Area] and *move type* [in area, out of area, and no move] were entered as fixed factors) yielded no statistically significant results. Participants were no more sensitive to the probe object’s displacements within its constrained area than outside of it (F = 0.51, df = 156, p = 0.60). Sensitivity to object displacements were no better for the trained objects, in either Area 1 or Area 2, than to the untrained object (F = 0.38, df = 156, p = 0.67).

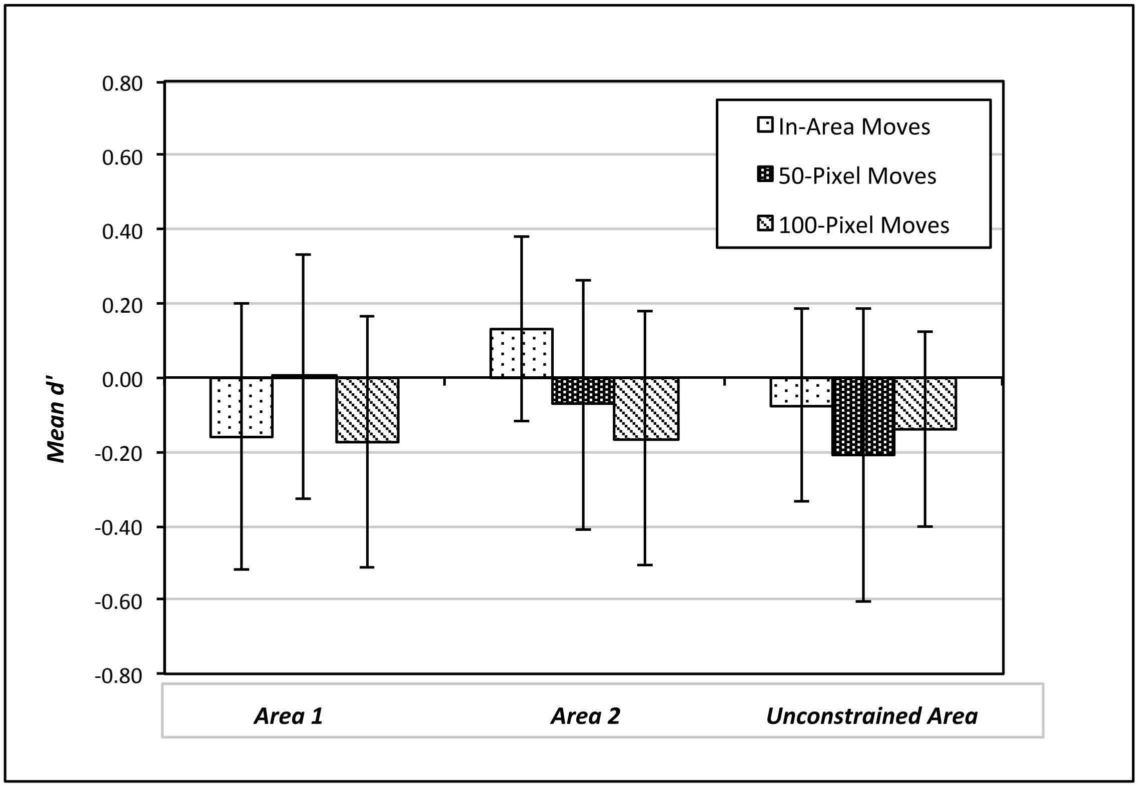


Figure A-2: Performance results from Experiment 6.

Participants were not statistically significantly sensitive to any of the transfer conditions in the experiment. Their tendency toward negative *d'* values indicates that they may have believed the probe object had been translated from sample to test when it had actually not (a high “false positive” rate in signal detection terms). The error bars represent the 95% confidence interval around the mean.

A.2.5 Discussion

The results of this experiment unfortunately did not add to our evidence that spatial constraints are easily learned. However, participants’ inability to discriminate between the appearances of the single probe object inside and outside of either of its bounded areas is perhaps easily explained. The pattern of *d'* results is almost uniformly negative; negative *d*’ values usually indicate a high rate of false positive responses. Because so much of the screen area was allotted to the transfer conditions where the trained probe object could appear outside its bounded area, either by 50 pixels or 100 pixels in any direction, the area where the object could appear within bounds was relatively very small (only about 8% of the total test scene pixels were within the bounded areas). Participants may have responded with false positives because it appeared to them that almost any translation was outside of bounds. A better design for a follow-up experiment might be to arrange somewhat larger bounded areas that abut the margins of the test scene area, so that they do not need a 100-pixel margin on all four sides. This would allow more area for translations within the bounded area, while preserving some area for testing whether participants are sensitive to small translations outside the bounded area. Alternatively, the experimental design could include moves outside the bounded area into any part of the screen area, as we did in the first experiment in Chapter 2. We could then match the extent of such moves within and outside the bounded area for statistical comparisons.

Another potential explanation for the null results of this experiment is that the training period is simply too short. Admittedly, the situations where a single object can appear in more than one spatial context are somewhat rare. The unusual character of the training period might necessitate more training trials in order for the participants to learn the rules of the situation.

APPENDIX B
ADDITIONAL EXPERIMENTS IN IDENTITY CONTEXT FORMATION

B.1 Introduction

Several experiments described in this section were performed in order to understand the interaction of identity context and spatial context. As we discussed in Chapter 2, spatial context effects tend to override identity context effects except where experimental conditions are carefully designed to let identity context emerge. As we found, at least one requirement for this effect is a marked difference in the visual appearance of some of the objects in the context. However, the experiments described below showed us some of the aspects of experimental protocol, object identity, and spatial configuration that are not conducive to identity context formation. In this manner, these experiments point out which directions for future research need to be modified in order to be productive.

B.2 Experiment 7

The purpose of this experiment was the same as for the two experiments described in Chapter 3: to investigate whether experience with a scene includes the association of a particular spatial configuration with all the identities of the objects that comprise it. This ability would correspond with observers' real-world extraction of associations from their experience; for example, with the "kitchen" configuration, between the identities of chairs, tables, sinks, and refrigerators and their specific locations in the configuration. If so, does the association of object identity with location break down when the configuration is composed of different object identities, for example, if the location of the refrigerator is occupied by an office desk? What happens if the objects remain the same, but the configuration changes; for example, if the refrigerator, sink, and kitchen table appear hanging on the walls and ceiling, as Biederman, Mezzanote, & Rabinowitz (1982) conjectured?

We would expect the ability to remember the location of the objects in the scene to be best when the association of object identity and configuration is the same as during learning. Memory should naturally be less accurate when the configuration differs but the object identities do not, which amounts to a rearrangement of the previous scene content. The most interesting comparison, however, would be between scenes where the configuration is the same, but the identities of the objects that comprise it are either the same or different at test. This comparison comprises a test of the notion that new objects comprising the same configuration should be regarded as anomalous by observers who have learned to associate a specific set of objects with that configuration. As other researchers have found, observers may attend more to the new identities (Becker, Pashler, & Lubin, 2007; Friedman, 1979; Loftus & Mackworth (1978), which may cause longer reaction time performance in the anomalous condition, but perhaps better performance for remembered location.

The experimental protocol was very similar to the protocol for the second experiment described in Chapter 3. It used the same abstract objects, which we called Type 1 in that chapter. However, any of the five objects in the test scenes could serve as the probe object; additionally, we included a transfer condition in which all the object identities remained the same, but the configuration was changed from sample to test. Thus, the memory load for this experiment was greater than in our previous experiments, because it required participants to hold a greater number of object identities and their spatial locations in memory.

B.2.1 Participants

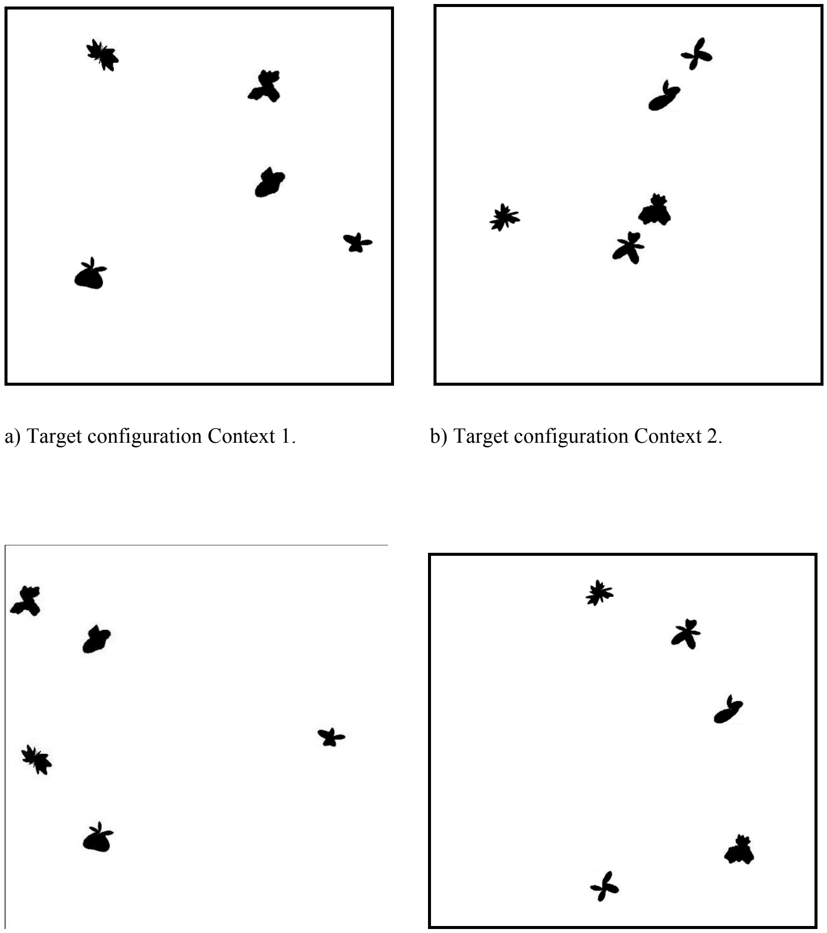
Twenty-one Cornell University undergraduates participated in the experiment. All participants had normal or corrected-to-normal vision. They received course credit for their participation. All participants gave their informed consent.

B.2.2 Stimuli

Artificial scenes were constructed using black abstract shapes randomly placed on a white background. Each scene consisted of five objects. The object shapes were generated by an algorithm created and used by Op de Beeck, Wagemans, & Vogel (2001). Each object was 50 x 50 pixels in size, subtending approximately 1.4° of visual angle.

The scenes were laid out by an algorithm created with MATLAB 7.0 with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). The objects were placed on a 600 x 600-pixel white square centered on the monitor screen and surrounded by a medium gray background extending to the edges of the screen. The generation algorithm ensured that the objects did not touch or overlap; they could appear within 4 pixels of each other, although in practice this did not occur.

Four target scene configurations were generated, referred to here as Contexts 1, 2, 3, and 4. Each context consisted of a given set of object shapes in a given overall configuration (Figures B-1a and B-1b). The configurations were grouped in pairs (Contexts 1 and 2, or Contexts 3 and 4), and each pair was learned by a randomly assigned group of one-half of the participants. Within a pair, the two contexts did not share any object shapes, although the same set of object shapes was used for both pairs. Distractor scenes were also generated, containing a different set of object shapes from either of the two contexts within a pair. Each distractor scene contained a different configuration of the distractor objects.



a) Target configuration Context 1.

b) Target configuration Context 2.

c) Target configuration Context 3.

d) Target configuration Context 4.

Figures B-1a, B-1b, B-1c, and B-1d: Target configurations for Contexts 1 through 4.

The scenes were displayed on a 17-in. (43.2 cm) Dell computer monitor with a resolution of 1024 x 768 pixels. The monitor was viewed from a distance of approximately 25 in. (63.5) cm in a normally lighted office, so that the entire viewing area from screen edge to edge subtended approximately 24.27º of height and 32.34º of width. The experimental protocol was controlled using MATLAB 7.0.

B.2.3 Procedure

After signing the informed consent and reading the experimental instructions on-screen, the participants began with a practice session consisting of 10 trials, including one representative trial each of the four experimental conditions for each context (described below), plus four distractor trials. The practice trials were not used in the actual experiment.

Each participant learned two target contexts, either Contexts 1 and 2, or Contexts 3 and 4. During the learning phase of the experiment, each of the target contexts was repeated 30 times, plus 60 distractor trials, for a total of 120 trials.

For each trial, the sample scene appeared for 2.0 s, followed by a 1.0-s mask consisting of a scrambled image of a natural scene, followed by the test scene. In test scenes, the scene configuration reappeared, but a single probe object that had appeared in the sample scene was missing. For both experimental condition trials and distractor trials, the probe object was randomly chosen from the five objects comprising the scene. Participants were instructed to click with the computer mouse as close as possible to the center of the location where they believed the probe object had appeared in the sample scene. They were permitted to view the test scene as long as they wanted before responding. Participants could take a short break between the first and second phases. The trial presentation order was randomized for each participant.

For the transfer phase of the experiment, the learned context scenes appeared as the sample scenes, but the test scenes were altered according to one of four transfer conditions:

- ***Same configuration/same objects condition***: the same configuration was occupied by the same objects as in the sample scene. In other words, sample and test scenes were identical.
- ***Same configuration/different objects condition***: the same configuration as in the sample scene was occupied by different objects, chosen randomly for each trial from among the set of distractor objects and the objects for the opposite context.
- ***Different configuration/same objects condition***: the same objects from the sample scene were arranged in a new configuration.
- ***Different configuration/different objects condition***: an entirely new configuration from the sample scene appeared whose positions were occupied by objects chosen from among the set of distractor objects and the objects for the opposite context.

The transfer phase consisted of 20 trials of each transfer condition for each context, plus an equal number of distractor trials constructed in the same manner as for the learning phase, for a total of 240 trials. Participants were not told that the configurations or the shapes that comprised them were being altered in this phase. As in the learning phase, the probe object was randomly chosen from the five objects comprising the sample scene for each transfer condition trial. The entire experiment took approximately 45 minutes.

B.2.4 Results

The dependent measures were the spatial error (in pixels) between the center of the probe object in the sample scene and the coordinates of the mouse click the participant made on the test scene, and RT. Each measure was analyzed in a four-factor mixed-model regression, where *participant* was entered as a random factor, and the three experimental condition factors were entered as fixed factors (*context–1, 2, 3, 4*; *positions–same, different;* and *object identities–same, different*). Trials with RTs greater than two standard deviations from the mean for all trials within a group of participants (those learning either Contexts 1 and 2, or 3 and 4) were discarded; approximately 4.6% of trials were discarded for the

Context 1 and 2 pair, and 4.9% for the Context 3 and 4 pair. Distractor trials were also excluded from the analyses.

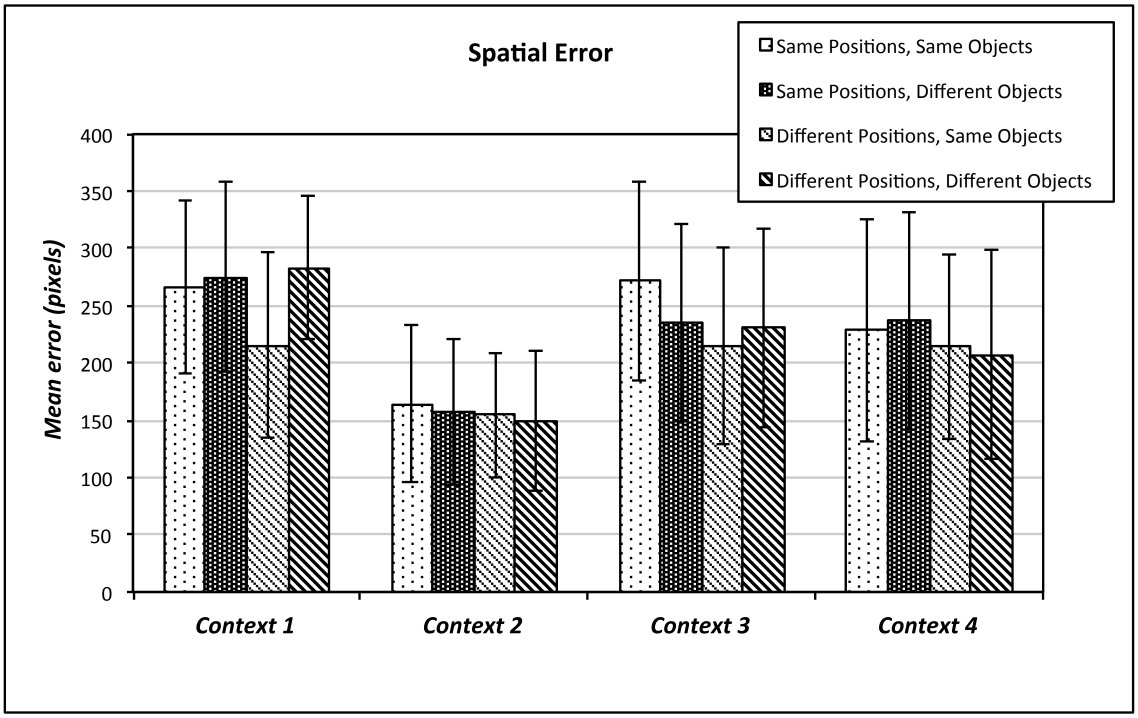


Figure B-2: Spatial error data from Experiment 7.
Error bars represent the 95% confidence interval around the mean.

For the spatial error data, overall there were statistically significant main effects of the context (configuration) itself (*F* (3, 152) = 151.103, *p* < 0.0001) and position (*F* (1, 3153) = 6.543, *p* = 0.011, *d* = 0.09), but no main effect of the object identities (*F* (1, 3150) = 1.680, *p* = 0.195) nor any interaction between position and object identities (*F* (1, 3150) = 0.234, *p* = 0.629). Clearly, some contexts are easier to remember than others; the errors were much smaller for Context 2 over all the transfer conditions. Spatial accuracy was in general greater for the two transfer conditions where positions changed, either with the same objects or different objects (except in Context 1).

Separate mixed-model regressions with the same factors were performed on each context. Within each context, the pattern of results was inconsistent. For Context 1, neither the positions of the objects nor their identities was statistically significant (F (1, 751) = 1.901, p = 0.168 for positions; F (1, 751) = 0.895, p = 0.344 for identities). For Context 2, the results showed the same pattern (F (1, 732) = 2.612, p = 0.106 for positions; F (1, 732) = 0.431, p = 0.512 for identities). For Context 3, there was a main effect of object identity (F (1, 822) = 12.026, p = 0.001, d = 0.24) but no main effect for positions (F (1, 823) = 1.521, p = 0.218). In contrast, for Context 4, there was a main effect of position (F (1, 820) = 10.061, p = 0.002, d = 0.22), but no main effect for object identity (F (1, 820) = 0.765, p = 0.382). There was no interaction between position and identity for any of the contexts (all $p > 0.05$).

The inconsistent main effects within the individual conditions might be attributable to the overall difficulty of the task. Yet it is important to note that within most of the individual contexts, there was no statistically significant difference between the two conditions that are of most interest, *the same positions/same objects* condition and the *same positions/different objects* condition (for Context 1, Fisher's LSD, df = 750, p = 1.000; for Context 2, df = 732, p = 1.000; for Context 4, df = 819, p = 1.000). Only in Context 3 was a significant difference between these two conditions found (Fisher's LSD, df = 821, p = 0.016). (All probability values were Bonferroni-corrected for the number of comparisons made.)

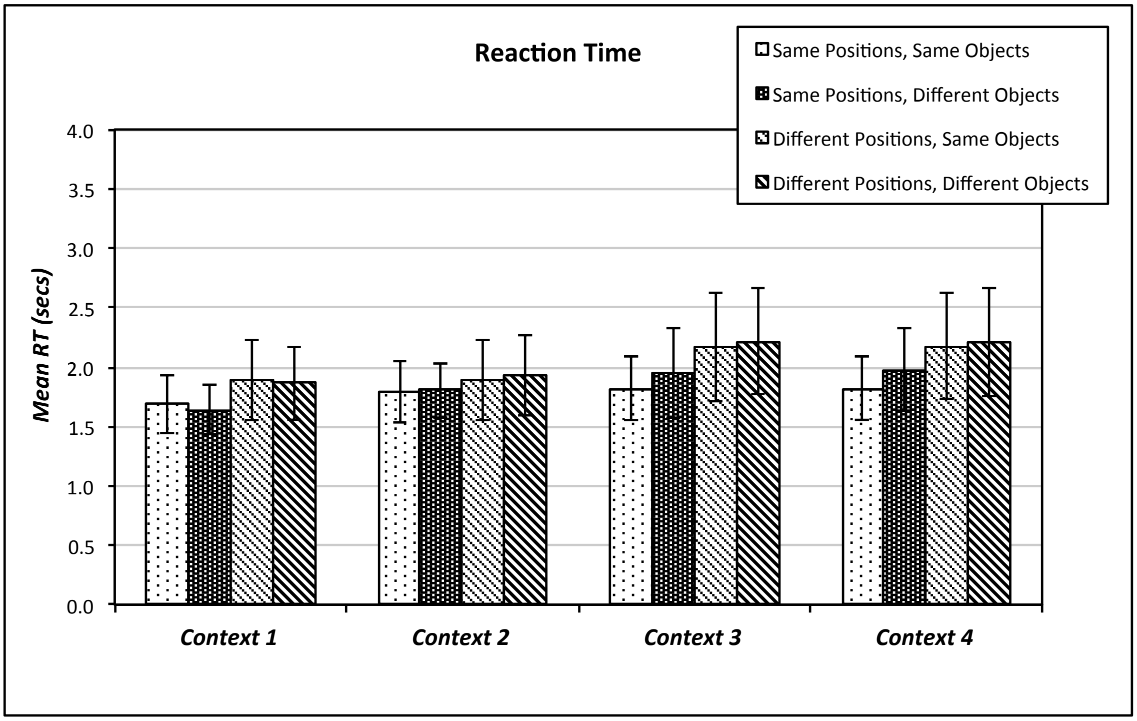


Figure B-3: Reaction time data from Experiment 7.

Participants were slightly slower in remembering the location of probe objects in the *same positions, same objects* and the *same positions/different objects* conditions only in Contexts 3 and 4. Error bars represent the 95% confidence interval around the mean.

Overall, the reaction time data showed a somewhat more interpretable pattern. Statistically significant main effects of context (*F* (3, 53) = 8.469, *p* < 0.0001), positions (*F* (1, 3149) = 258.350, *p* < 0.0001, *d* = 0.57), and object identities (*F* (1, 3149) = 13.119, *p* < 0.0001, *d* = 0.13) were all found. There was no interaction between positions and object identities (*F* (1, 3149) = 1.244, *p* = 0.265). Participants' RTs generally increased with each kind of change to the sample scene, whether it was a change in position or in object identities; clearly they found it more difficult to remember the positions of objects when the overall configuration changed from sample to test, although as noted, their remembered positions were slightly more accurate.

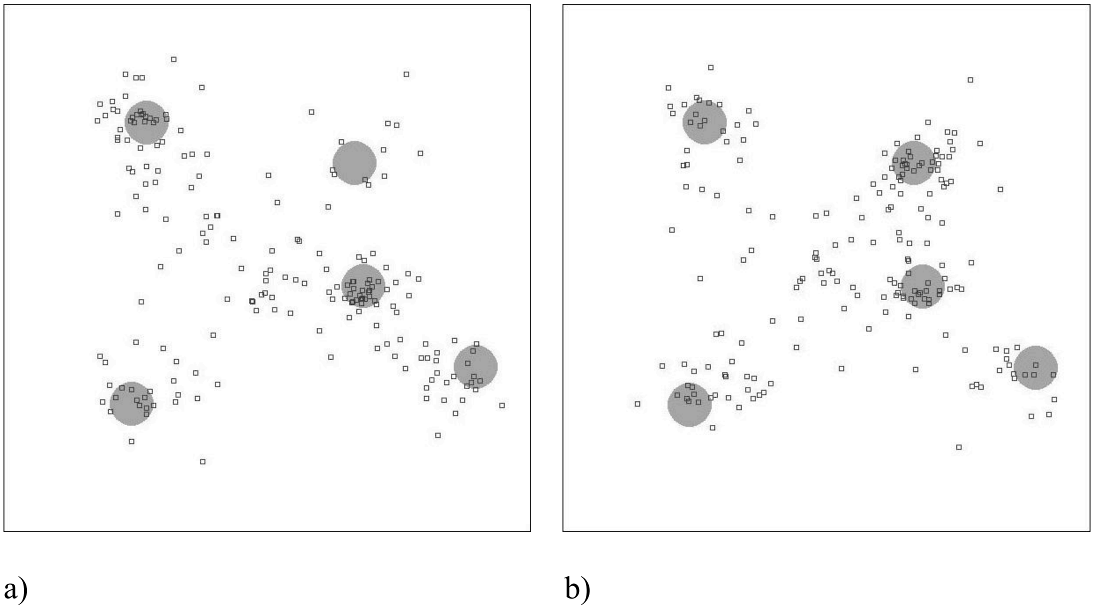
Additional mixed-model regressions performed on each context showed that, for Context 1, positions were significant (*F* (1, 750) = 69.242, *p* < 0.0001, *d* = 0.61), but object identities were not (*F* (1, 750) = 0.180, *p* = 0.677). Context 2 also showed a significant main effect of positions (*F* (1, 731) = 23.207, *p* < 0.0001, *d* = 0.36), but not of object identities (*F* (1, 731) = 0.614, *p* = 0.433). In Contexts 3 and 4, both positions and object identities were statistically significant (Context 3, positions [*F* (1, 821) = 94.832, *p* < 0.0001, *d* = 0.68]; Context 3, object identities [*F* (1, 821) = 7.928, *p* = 0.005, *d* = 0.20]; Context 4, positions [*F* (1, 819) = 82.998, *p* < 0.0001, *d* = 0.64]; Context 4, object identities [*F* (1, 819) = 10.697, *p* = 0.001, *d* = 0.23]).

Once again, individual comparisons between the same *positions/same objects* transfer condition and the *same positions/different objects* condition showed no statistically significant effects for two of the contexts (for Context 1, Fisher's LSD, *df* = 749.997, *p* = 0.955; for Context 2, *df* = 731.033, *p* = 1.000). However, for Context 3, (*df* = 821.021, *p* = 0.019) and for Context 4, (*df* = 819.023, *p* = 0.004), significantly longer RTs emerged.

B.2.5 Discussion

Clearly the individual context (configuration) affects the ease with which participants can remember the position of the missing probe object. Beyond that, the results seem to show that there is little difference between test scenes where the same positions are occupied by either the same objects or different objects as the scene observed in the sample. It takes slightly longer, in some contexts, to call to mind the position of the probe, but in general, memory for the position itself is poor.

Because any of the objects in the sample scene could serve as the probe object, participants likely attended to the identities of most or all of the objects in the sample scene in order to attempt to remember their locations in the test scene. Figure B-4a shows the actual locations of the participants' mouse clicks for the remembered positions overlaid on a schematic representation of the objects' actual positions. It shows a moderate degree of clustering of remembered locations for all the objects in the *same positions/same objects* condition, while Figure B-4b shows the clustering for the *same positions/different objects* condition. However, because the spatial error is relatively high for both these transfer conditions, one explanation might be that the participants remembered the positions, but not the identity of the objects that occupied them. In this task, all but one of the positions of the configuration are re-presented at test. Participants therefore have comparatively less to hold in memory about the configuration than they do about the object identities, but their memory resources may be occupied with the attempt to remember all of the object identities. Because the spatial error is equally great whether the positions are occupied by the same or by different objects upon test, we might conclude that the participants have relatively poor memory for any of the object identities. The difficulty of remembering the identity and locations of five objects per context may overwhelm the participants' visual memory resources, diminishing the possibility of associating the identities with the positions of the context itself. A number of psychophysical studies support this explanation (Irwin & Zelinsky, 2002; Luck & Vogel, 1997; Vogel, Woodman, & Luck, 2001); the consensus of these studies is that four to five objects can be maintained in visual short-term memory. A recent neurophysiology study by Gao, Li, Liang, Chen, Yin, & Shen (2009) corroborates this view, having found that the brain areas implicated in representing complex shape information show a diminished response when the number of objects in a scene exceeds four.



Figures B-4a and B-4b: Plots of remembered locations of probe objects.

(a) Remembered locations of probe object from Context 1, *same configuration/same objects* transfer condition, for all trials The gray-filled circles represent the locations of the objects in the scene, any of which could serve as the probe object. (b) Remembered locations of probe objects for Context 1, *same configuration/different objects.*

It could equally be argued, however, that the participants remembered most or all of the objects in the sample scenes through the intervening mask, and were somewhat distracted by the substitution of other objects in the remaining slots of the test scene, but simply performed so poorly on the spatial memory task that this result was obscured. The RT data for two of the contexts at least shows that the substitution of different objects in the same configuration slows the participants' response to the task, implying that the unfamiliar identities are interfering with this process. We believe that a clearer understanding of the causes involved was provided by limiting the demands on visual memory, as we did in the second experiment in Chapter 3. There, we found that spatial configuration was probably the more salient cause of spatial error and longer RTs.

B.3 Experiment 8

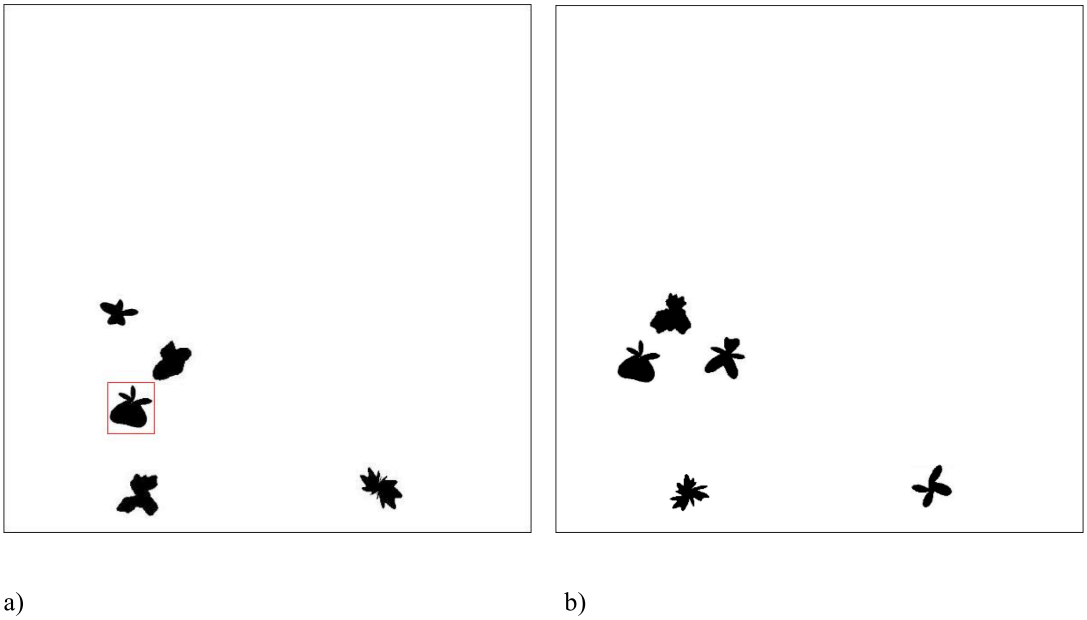
One difficulty in determining whether object identity makes a difference to the formation of context is that observers may use either the global configuration or a local portion of it as the cue to the probe object's location instead of its identity (Jiang & Wagner, 2004). Another way to pit identity against configuration is to design a situation where the location of a probe object depends entirely on the identities of the other objects in the context, but not their locations, but without decoupling the two aspects (as Endo and Takeda [2004] did). To that end, our next experiment presented participants with two contexts that differed in only one small respect; the probe object could appear in two slightly different positions, depending on which of two sets of other objects appeared with it. The slight displacement of the probe object means that while learning the two contexts, the global configuration of the objects is extremely similar. In this way, participants should be less able to code the probe object's location in memory by reference to the configuration itself.

B.3.1.Participants

Eleven Cornell University undergraduates participated in this experiment. All participants had normal or corrected-to-normal vision, received course credit for their participation, and gave their informed consent.

B.3.2 Stimuli

In this experiment, the scenes were composed of the same abstract objects as in the previous experiments. Two target scene configurations, Context 1 and Context 2, were generated, which were identical except for the location of the probe object. The probe's position was changed to approximately one object width upward and leftward from Context 1 to Context 2. The second difference between the contexts was that two separate sets of objects comprised the other objects in the scene. Thus, during the learning phase, the two contexts appeared very similar, as shown in Figures B-5a and B-5b:



Figures B-5a and B5b: Target configurations for Contexts 1 and 2.

The probe object is outlined in Figure B-5a (for purposes of illustration only; the outline was not visible in the actual experiment). Note in Figure B-5b its slight displacement from Context 1 to Context 2.

Distractor scenes were also constructed of a different set of objects from either Context 1 or Context 2.

B.3.3 Procedure

The procedure was the same as for previous experiments, except that a single object served as the probe object in all the test scenes. The mask between sample and test scenes was a medium gray screen displayed for 1.0 s.

Participants began the experiment with a practice session of nine trials, one each of the four transfer conditions (including the *same configuration/same objects* condition for both Context 1 and Context 2), plus five distractor trials.

In the learning phase, participants saw 30 trials of each of the two contexts, plus 60 distractor trials, for a total of 120 trials.

For the transfer phase of the experiment, the learned context scenes appeared as the sample scenes, but the test scenes were altered according to one of four transfer conditions:

- ***Context1/Context 1 condition***: the same configuration was occupied by the same objects as in the sample scene for Context 1 (20 trials).
- ***Context 2/Context 2 condition***: same as above, for Context 2 (20 trials).
- ***Context 1/Context 2 condition***: the sample scene was Context 1, while the test scene switched to Context 2 (20 trials).
- ***Context 2/Context 1 condition***: the sample scene was Context 2, while the test scene switched to Context 1 (20 trials).

The transfer phase also included 120 distractor trials; the disproportionate number was included to divert participants' attention from the similarities between Context 1 and Context 2 to ensure that they did not simply learn the two locations and nothing else. In distractor scenes, the probe object was randomly chosen from among all the objects in the scene.

B.3.4 Results

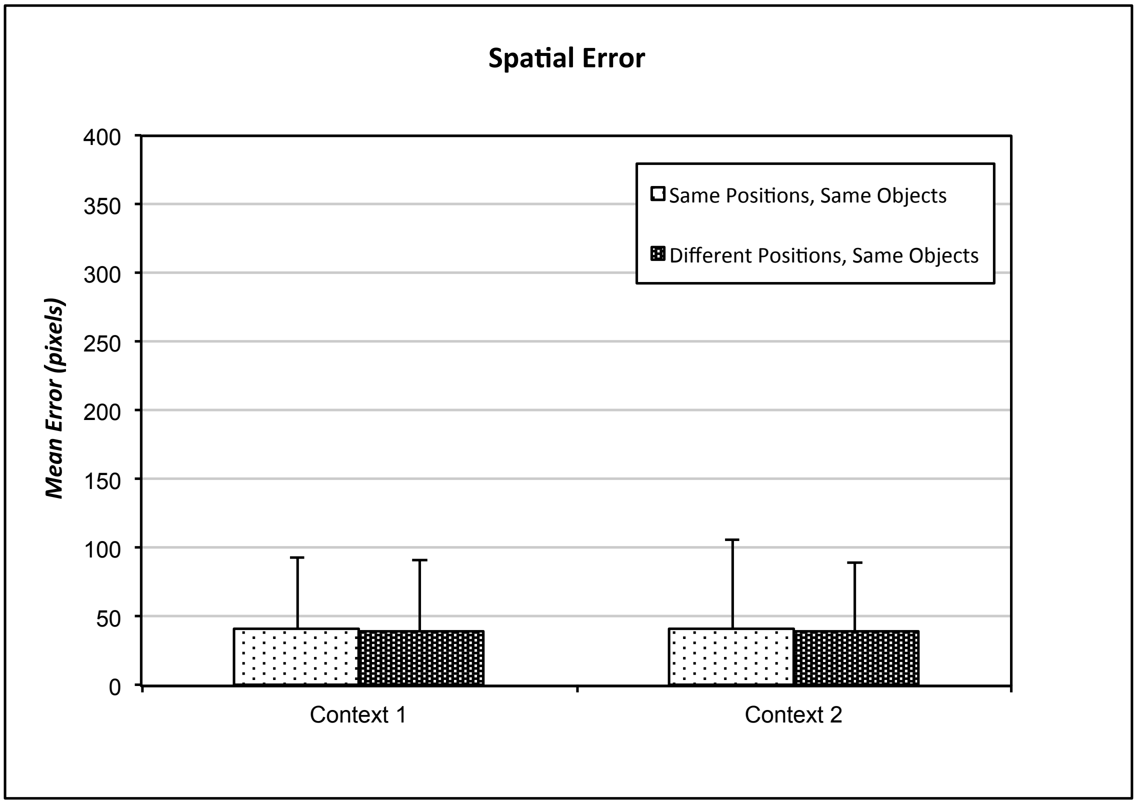


Figure B-6: Spatial error data for Experiment 8.

There was almost no difference in the remembered position of the probe object whether it appeared with the same objects in the configuration or with different objects. Error bars represent the 95% confidence interval around the mean.

A mixed-model regression analysis was performed on the spatial error data, with *participant* as a random factor. Initial context (*Context 1, Context 2*) and test context (same levels) were entered as fixed factors. No main effect of initial context (*F* (1, 866) = 0.435, *p* = 0.510), or of test context (*F* (1, 866) = 0.159, *p* = 0.690) was found. Nor was there any interaction between initial context and test context (*F* (1, 866) = 0.421, *p* = 0.517). An additional mixed-model regression with a combined experimental condition factor (*1–Context 1/Context 1; 2–Context 2/Context 2; 3–Context 1/Context 2; 4–Context*

2/Context1), showed no main effect of the condition (*F* (3, 866) = 0.338, *p* = 0.798). In addition, no comparison between any pair of conditions was statistically significant (all comparisons Fisher's LSD with Bonferroni correction, *df* = 866, *p* = 1.000).

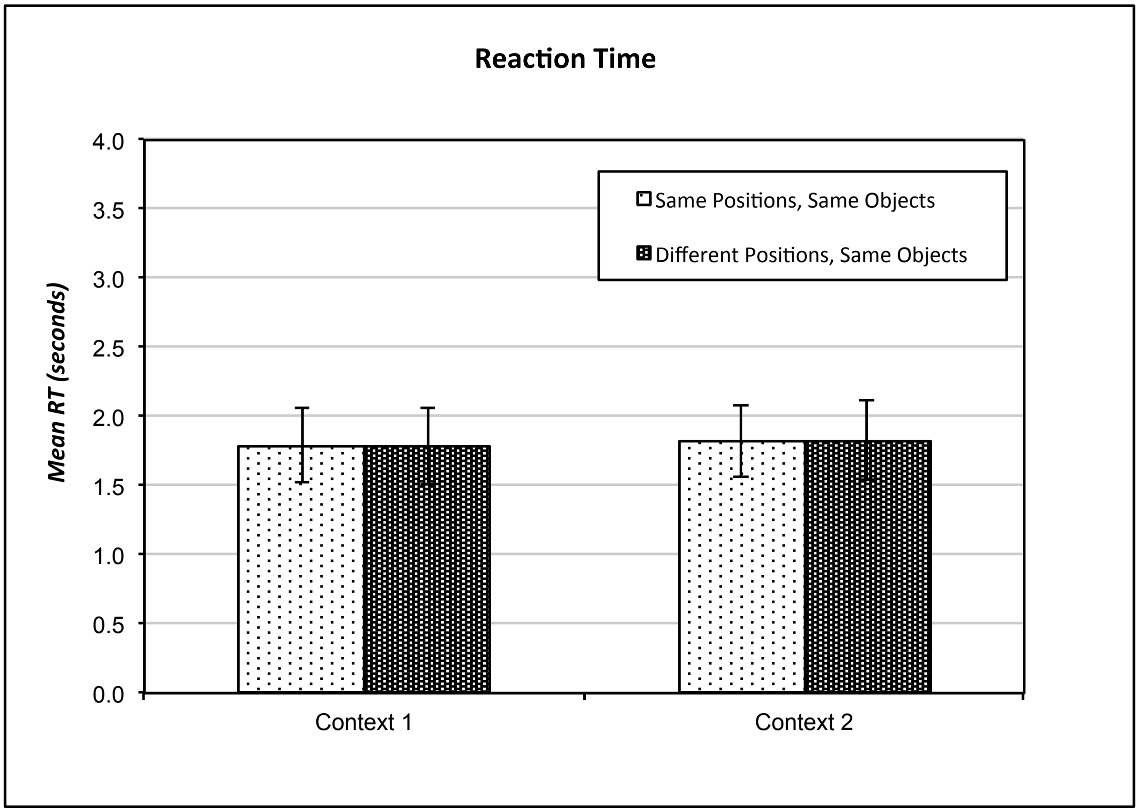


Figure B-7: Reaction time data for Experiment 8.

Participants took almost exactly the same time to remember the position of the probe object whether they were presented with the same objects in the configuration or with different objects. Error bars represent the 95% confidence interval around the mean.

Before analysis, distractor trials and trials with RTs greater than two standard deviations from the mean were excluded; approximately 2.2% of trials were removed. For the RT data, no main effect of initial context (*F* (1, 866) = 0.574, *p* = 0.449), or of test context (*F* (1, 866) = 0.097, *p* < 0.756) was found. There was no interaction between initial context and test context (*F* 1, 866 = 0.478, $p < 0.490$). An additional mixed-model regression showed no main effect of condition (*F* 3, 866 = 0.383, $p = 0.765$). No comparison between any pair of conditions was statistically significant (all Fisher's LSD with Bonferroni correction, *df* = 866, $p = 1.000$).

B.3.5 Discussion

The task of remembering two spatial locations for a single probe object, depending on which other objects appear in the scene, appears to have somewhat dissociated the implicit learning of object identity from global or local configuration. Participants were relatively accurate in remembering both locations in their appropriate conditions; their mean spatial error was lower than for any of the experiments in this series (all those in Chapter 3 and Appendix B). They were equally accurate in relocating the probe object when Context 1 switched to Context 2 at test as when Context 2 switched to Context 1. Their RTs were also nearly equal in both switch conditions.

If object identity in combination with configuration affects implicit learning of context, we might have expected that the switch conditions would show less accuracy or at least longer RTs as participants attempted to remember the location of the probe object in the previous identity/configuration. Instead, participants' attention may have been directed to the appropriate probe object location by identification of the other objects with which it appeared, and they may then have simply remembered the location without being distracted by the other identities whether the context switched at test or not.

To this point, our results seem to confirm that spatial configuration information is likely more robust in visual memory than object identity, even when the load on memory is lighter than in previous research, as in the experiments in Chapter 3, or when the information available in the configuration is reduced, as in the present experiment. Some other effect must account for humans' knack of registering co-occurrences of object identities from free viewing of the real world. Our research to date has indicated that this effect may have to do with the ways in which the visual properties of controlled stimulus arrays differ from those of the natural world, especially marked differences in visual appearance.

APPENDIX C
INSTRUCTIONS FOR EXPERIMENTS

B.1 On-Screen Instructions for Experiments 1, 2 and 6

We are interested in how people's vision works when they look at scenes composed of objects in the natural world. Although this experiment uses unfamiliar objects in unfamiliar scenes, we are interested in whether you think you have seen an object in a particular location before.

You'll see sets of objects presented on the computer screen for a few seconds. In each presentation, 4 or 5 different objects will be on the screen.

Every few presentations, one of the objects will appear circled in red. When you see the red-circled object, your task is to decide whether you have seen that object in that location before.

Press the "L" key if you think that object *has* previously been in that location.

Press the "S" key if you think that object *has not* previously been in that location.

(You should position your fingers over the "L" and "S" keys before the experiment starts.)

You have as much time as you need to decide.

You will get a chance to practice before the main experiment begins. The entire experiment will take about an hour. You can receive a full explanation of the experiment's purpose when you have completed it.

Press any key when you are ready to start.

B.2 On-Screen Instructions for Experiments 3, 4, 7, and 8

We are interested in how people's vision works when they look at a succession of simple scenes, each composed of several objects.

You'll see sets of objects presented briefly on the computer screen. In each presentation, 4 or 5 objects will be on the screen.

In Experiment 3, the following wording was used:

The screen will then change to a scrambled image. As soon as the scrambled image disappears, a blank area will reappear. One of the objects will also appear at the left of the blank area.

In Experiment 4, 7, and 8 the following wording was substituted for the paragraph above.

The screen will go blank, then reappear with one of the objects missing. As soon as the scene reappears, your task is to click where the missing object used to be. That object will appear on the left, outside the scene, to remind you what it was.

Click the point of the cursor as close as you can to the center of where the object used to be. Your response will be timed; try to be as fast and as accurate as you can.

You will get a chance to practice before the main experiment begins. The entire experiment will take about 45 minutes. You can receive a full explanation of the experiment's purpose when you have completed it.

Press any key when you are ready to start the practice session.

B.3 On-Screen Instructions for Experiment 5

Initial Instructions:

We are interested in how many pictures people can remember. In this experiment, you’ll see sets of pictures that contain images of different kinds. Although some of these sets might not look like what you think of as pictures, try to remember each image you see. Some types of pictures will be harder to remember than others; that’s okay, just try your best.

This experiment is divided into three parts, in which you’ll see several sets of images for each part. The entire experiment will last about 45 minutes. After you have completed the experiment, you can receive an explanation of its purpose.

The first part is short. There will be five sets of images in this part, and you will do the same thing for each set. First you will see several images, one after another. After you have seen all the images in a set, you’ll be presented with a pair of images side by side. Each pair will contain one of the images you have already seen from that set, plus a new image. Your task is to decide which one of the pair you have already seen.

If the image you have already seen is on the left, press the ‘F’ key. If it is the image on the right, press the ‘J’ key. A message will pop up after each pair of images to remind you of this.

Press any key when you are ready to start.

Part 2:

This part of the experiment is exactly like the first part, except that you will see many more images in each set.

Remember, first you will see images presented one after another. After you have seen all the images in a set, you will be presented with a pair of images side by side. Each pair will contain one of the images you have already seen from that set, plus a new image. Your task is to decide which one of the pair you have already seen.

If the image you have already seen is on the left, press the ‘F’ key. If it is the image on the right, press the ‘J’ key. A message will pop up after each pair of images to remind you of this.

Press any key when you are ready to start.

Part 3:

The final part of the experiment is also short. You will see five sets of images, but this time, you will only see pairs of images presented briefly side by side. Your task is to determine if the images are the same or different. Some of the pairs will be clearly different, and some will be harder to tell apart. That’s okay, just try your best. We’re interested in how hard it is to discriminate one image from another.

Exactly one-half of the pairs will contain the same image. The images in them will be exactly the same; we are not trying to trick you.

If the images are the same, press the ‘F’ key. If the images are different, press the ‘J’ key. A message will pop up after each pair of images to remind you of this.

Press any key when you are ready to start.