Scene perception and memory

Marvin M. Chun

Dept. of Psychology, Center for Integrative and Cognitive Neuroscience,

Vanderbilt Vision Research Center, and Kennedy Center

Vanderbilt University

IN PRESS

To appear in D. Irwin and B. Ross (Eds.) Cognitive Vision.

Address correspondence to:

Marvin M. Chun

Department of Psychology

Vanderbilt University

531 Wilson Hall

Nashville, TN 37203

e-mail: marvin.chun@vanderbilt.edu

phone: (615) 322-1780

fax: (615) 343-8449

TABLE OF CONTENTS

I. Introduction

Everywhere we look a visual scene is in view. Scenes embody most of the objects and events that we must locate and identify to guide our thoughts and actions. Thus, it may not be an exaggeration to state that to understand scene processing would be to understand vision.

The ability to perceive one's local visual environment is so important for navigation and other daily activities that it is perhaps not surprising that a region of the brain appears to be specialized for processing scene information. The parahippocampal cortex responds robustly to visual scenes, namely, depictions of visual space (Aguirre, Detre, Alsop, & D'Esposito, 1996; Aguirre, Zarahn, & D'Esposito, 1998; Epstein, Harris, Stanley, & Kanwisher, 1999; Epstein & Kanwisher, 1998). This region has been dubbed the parahippocampal place area (PPA) (Epstein & Kanwisher, 1998), and it can be readily identified within subjects using functional magnetic resonance imaging (fMRI) by localizing the cortical regions that respond significantly stronger to scene stimuli compared to face, object, or scrambled scene stimuli. Figure 1 shows a sampled region of the PPA within medial temporal cortex in a human subject. These data were collected in our lab, and the bar graph indicates mean signal strength of the fMRI signal that correlates with neural activity. The results indicate that the PPA region is more active to scenes than to faces or scrambled stimuli.

Insert Figure 1 about here

Despite great strides in understanding <u>where</u> scenes are perceived in the brain, not enough is known about <u>how</u> people perceive scenes and use scene information to

guide their actions. Theoretical insights into scene recognition have been hampered by the fundamental question of how to classify and characterize scenes. Unlike faces, which share a similar configuration of commonly shared diagnostic features such as two eyes, a nose, and a mouth, the tremendous variety of scenes we experience do not appear to share much in common, except for the fact that scenes depict a 3-dimensional layout containing objects and surfaces. Researchers lack a grammar to describe scenes or even criteria to distinguish different scenes. These limitations pose a fundamental challenge for the study of scene recognition because any scientific investigation requires at least some common language and rules for characterizing what is being studied.

As a step towards understanding scene recognition and memory, this chapter will review studies from the literature and from my lab that describe how visual scenes and scene properties are learned and represented in the brain. Another aim of this chapter is to identify outstanding issues in scene perception and memory that deserve further research. In Section III-A, I will sketch a dual-path model of scene representation as one possible framework to guide future work.

The chapter will begin with a brief review of some basic properties of scenes. Despite a lack of consensus on how to operationalize different scenes, visual scenes do share a number of properties that are uncontroversial. I will describe three of these characteristics below.

A. Scenes are complex

Most everyday scenes are complex in detail, presenting a rich multitude of objects and surfaces to the observer. In fact, the amount of information in any given

scene greatly exceeds what can be handled by the brain at any given time: the well-known problem of information overload (Broadbent, 1958; Chun & Wolfe, 2001; Pashler, 1998). Such complexity leads to rather dramatic gaps in people's perceptual grasp of the visual world, and it also has led to rather sophisticated attentional selection mechanisms that efficiently locate and detect important information within complex scenes (Chun & Marois, 2002).

Some of the most compelling lab demonstrations of limited capacity in scene processing are based on what has come to be known as the "change blindness" paradigm (Rensink, 2002; Simons & Levin, 1997). One of the most dramatic examples was in a study that demonstrated real-world failures to detect a switch in a person's identity when that switch happened behind a brief occluding event, such as a door passing in between the observer and the switched person (Simons & Levin, 1998). Simpler, though no less compelling, demonstrations of change blindness from the lab involved failures to detect a change between two otherwise identical pictures of scenes flickering back and forth with an intervening mask to disrupt visual transients (Rensink, O'Regan, & Clark, 1997). In these "flicker tasks," subjects have trouble detecting salient changes such as a bridge disappearing and reappearing across flicker. Even in situations where a scene does not flicker, subjects have difficulty detecting changes that are introduced into the scene during eye movements (Irwin, 1991; McConkie & Currie, 1996) or with other visual transients (O'Regan, Rensink, & Clark, 1999).

Such powerful demonstrations of blindness to details in scenes appear to support proposals that very little visual information is retained from one moment to the next (Horowitz & Wolfe, 1998; O'Regan, 1992). Although this view is probably too extreme,

in light of recent demonstrations of good memory for objects in scenes (Gibson, Li, Skow, Brown, & Cooke, 2000; Hollingworth & Henderson, 2002; Hollingworth, Williams, & Henderson, 2001; Kristjansson, 2000; Peterson, Kramer, Wang, Irwin, & McCarley, 2001; Shore & Klein, 2000), there is no doubt that human observers must constantly contend with a burdensome amount of visual information.

What's remarkable is that the visual environment typically does not "feel" so burdensome, because we can usually find and attend to the information we need without much time and effort (Chun, 2000; Rensink, 2000). This highlights the efficiency of powerful attentional mechanisms that direct limited capacity cognitive processing to the most important object or event that is relevant to our current behavioral goals. For example, while driving, we rapidly detect and usually obey traffic signals and stop signs without much second thought. Yet, such important, but seemingly easy tasks daunt the abilities of the many computer chips that control so many other functions within our automobiles these days. Biological perception is more powerful and more intelligent, based on the brain's ability to utilize both bottom-up and top-down cues (Treisman & Sato, 1990; Wolfe, 1994). Bottom-up cues within a scene include abrupt onsets or salient visual features that are unique in the color, size, orientation, motion direction, or other visual primitive (Bravo & Nakayama, 1992; Theeuwes, 1992; Treisman & Gelade, 1980; Yantis & Jonides, 1984). Top-down cues include perceptual set (Egeth, Virzi, & Garbart, 1984; Folk, Remington, & Johnston, 1992), novelty (Johnston, Hawley, Plew, Elliott, & DeWitt, 1990), and scene context (Biederman, Mezzanotte, & Rabinowitz, 1982; Chun & Jiang, 1998, 1999). These factors can be combined to drive selection in an efficient manner.

The efficiency of bottom-up and top-down cues is typically studied using visual search tasks, where observers are asked to search for a target appearing amongst a variable number of distractors. The visual search displays form artificial scenes, which can be controlled to study the factors that influence attentional selection. For inefficient search tasks, target detection time increases with set size, while for efficient search tasks, target detection time is independent of set size. For example, uniquely colored targets are detected rapidly, while targets that are more similar to distractors take more time to find (Duncan & Humphreys, 1989).

B. Scenes have invariant structure

The visual world is not random, and the statistics of the environment do not change radically over time. Rather, scenes contain "structure," an obvious, but underappreciated feature of everyday scenes that we consider to be extremely important (Chun, 2000; Fiser & Aslin, 2001, 2002; E. J. Gibson, 1969; J. J. Gibson, 1966; Olshausen & Field, 2000; Reber, 1989; Saffran, Aslin, & Newport, 1996). By structure, we are referring to the fact that the visual environment contains regularities, properties that recur over time: cars travel on roads, people walk on sidewalks, windows can be found on buildings, and so on. Even novel scenes tend to resemble those we've experienced in the past, allowing us to drive through new neighborhoods and stroll in new shopping malls. In sum, natural environments tend to be stable over time, and when dynamic features exist, they tend to move about and change in fairly regular, predictable ways.

We consider the invariant structure of scenes to be a key to understanding scene perception, and so this property provides the motivation for much of my lab's work on

scene perception and memory. Our basic proposal is that observers are exquisitely sensitive to visual information that is invariant. For example, the configuration of furniture in one's office or the layout of buildings on one's campus tends to be stable. Even local "scenes", such as the instrumentation panel of one's car, do not change from moment to moment or day to day. Encoding such regularities should facilitate one's interactions with these "scenes" on future encounters. Thus, understanding how scene information is processed and used by the brain can be studied as a problem of learning and memory. How does the brain encode invariant visual information, and how does invariant information benefit visual behaviors and action?

One may first approach this problem by first cataloging the different types of structure that scenes contain. In a recent review, Henderson and Hollingworth (1999) defined a visual scene as "a semantically coherent view of a real-world environment comprising background elements and multiple discrete objects arranged in a spatially licensed manner." Accordingly, we can identify the following key features of everyday scenes. First, scenes contain spatial configuration information about where objects are located relative to each other. Such spatial regularities can be very stable, such as buildings in a neighborhood, or approximate, such as paper on a desk or forks on a table. Second, scenes contain object shapes that covary with each other. A kitchen typically contains a sink, a stove, dishes, cups, and so on. In a living room one is more likely to see a sofa than an elephant. Thus, regularities exist in the range of objects that tend to co-occur within a scene. Finally, in addition to spatial and object shape information, scenes viewed over time also contain rich temporal structure. In dynamic environments such as driving or basketball, there are regularities in how objects move

about and change over time, allowing us to anticipate what would happen next. Thus, it is important to understand how scene information is integrated over time. We will review studies that illustrate these points in SectIon II on contextual cuing.

C. Scenes provide contextual information to object recognition

Objects in natural scenes rarely occur in isolation, but are almost always presented within a rich, detailed mosaic of other features, surfaces, objects, and events. These properties form the global visual context that exists for most of our perceptual interactions with the world. As noted earlier, global context is the source of information overload that complicates the task of individual object recognition. However, there are redundancies and regularities in this flux of information (Biederman, 1972). In most natural scenes, objects and events tend to correlate with each other providing a rich, invariant *covariational texture* of information that serves to decrease complexity and increase predictability (E. J. Gibson, 1969). Although presented in a different theoretical framework and level of analysis, both E. J. Gibson (1963; 1966) and J. J. Gibson (1966) spoke about the attunement of perceptual systems to invariant information in the physical world. In short, sensitivity to regularities in the environment is informative and helpful, and perceptual experience educates and optimizes attention. Reber (1989) makes a similar point in stating that when the stimulus environment is structured, people learn to exploit the structure to coordinate their behavior in a coherent manner.

Such theoretical considerations lead to the simple prediction that global visual context should provide important constraints to visual processing. We propose that one important role of visual context is to guide the deployment of visual attention (Chun, 2000). Attention handles how information is extracted from scenes and how this

information can be used to guide behavior. For example, context and scene meaning may guide eye movements towards important regions within scenes that are consistent with the ongoing goals of the observer. Numerous eye movement studies have shown that fixations indeed tend to cluster around regions deemed to be central to the meaning of the scene or relevant to an ongoing task (Loftus & Mackworth, 1978; Mackworth & Morandi, 1967; Shinoda, Hayhoe, & Shrivastava, 2001; Yarbus, 1967).

II. Contextual Cuing

My colleagues and I have developed a number of tasks to study how the invariant nature of complex scenes comprises contextual information that guides visual behavior. We use the term contextual cuing to refer to the process by which scene context information guides visual attention to important locations, objects and events within scenes. Unlike most prior work in scene recognition that uses real-world scenes or depictions of real scenes, we employ rather impoverished, artificial "scenes." What we lose in realism, we gain in our ability to operationalize and control different components of scenes such as their layout and content. More importantly, by using novel scenes, we can explore how scene information is learned. In relation to this, we aim to elucidate the neural mechanisms involved in representing complex scene information. Note that the principles that benefit performance in our artificial displays have correlates in studies that employ more naturalistic, real-world images (Ryan, Althoff, Whitlow, & Cohen, 2000; Sheinberg & Logothetis, 1998).

A. How does spatial layout cue location?

As reviewed above, a primary feature of scenes is that objects are arranged in a "spatially licensed manner" (Chun & Jiang, 1998; Henderson & Hollingworth, 1999). Buildings maintain their configurations over time, as does the furniture in one's office. Certainly variation occurs, but by and large, the positions of most objects in the visual world are fairly stable, especially from one moment to the next. Such regularities are presented to observers in the form of invariant visual context, such that encoding such contextual information is not only critical for navigating around the environment, but also for orienting to objects within scenes.

Our first study on contextual cuing examined how spatial context cues attention (Chun & Jiang, 1998). We required subjects to quickly detect a target, a rotated T, appearing amongst 11 other rotated L shapes (See Figure 2). This is a difficult search task that requires careful scanning of the display using focused visual attention, and we measured the time it took to locate the target. Such displays can present clearly defined multiple objects in a flexible, but fully controlled manner. But what is "context" for such sparse displays? Our insight was to define context as the spatial layout of the distractor items surrounding the target. To make this scene property "invariant," we repeatedly presented a set of 12 different scenes (search arrays) across blocks throughout the entire session. To make the scene property useful and predictive, for each repeated scene, we had the target appear in a consistent location relative to its visual context (global configuration). If observers are sensitive to the invariant spatial configuration surrounding the target, then subjects should be able to detect the target within repeating displays more quickly as they experience more repetitions. Search for

targets appearing in the repeated old scenes was compared to that for targets appearing in new contexts, randomly generated in each block to serve as a baseline. Subjects were significantly faster at detecting targets appearing in old displays compared to targets appearing in new displays. We call this the contextual cuing effect because visual context served to cue attention to the target, facilitating search. In addition, subjects were not aware of which displays were old or new, making this task an implicit one, a point that we will return to in Section III-B. Similar results were observed using pseudo-naturalistic displays with 3-dimensional perspective (Chua & Chun, in press).

Insert Figure 2 about here

What exactly is contextual information guiding? We had proposed that context guides "attention" based on the assumption that the allocation of attention to a target precedes any action directed towards it. However, we had to infer this based on manual response times. An example of a more direct visual behavior would be eye movements that direct foveal resolution to a target item. Indeed, a recent study that measured eye movements showed that fewer saccades were needed to acquire a target appearing in an old display compared to a new display (Peterson & Kramer, 2001). Similar results have been observed in monkeys making eye movements to targets embedded in natural scene backgrounds (Sheinberg & Logothetis, 1998). Interestingly, such contextual cuing of eye movements may even override the powerful pull of salient visual events such as abrupt onsets (Peterson & Kramer, 2001ab).

Although the contextual cuing paradigm was developed to get a better handle on the notion of "context" in visual processing, a number of questions arise from the

demonstration of robust, implicit contextual learning. Namely, what is the limit? Any given scene contains a prohibitively large amount of information, all of which need not be encoded. So what counts as context? To begin to address this issue, we raised two questions to examine what counts as context in the artificial displays used in Chun and Jiang's (1998) study.

First, is the entire display of 12 items encoded as global context, or does local context around the target suffice? Olson and Chun (2002) tested this by making only half of each display invariant, while the other half of the display changed randomly from repetition to repetition. The invariant half of the display could either be on the side containing the target or on the opposite side. Thus, for each old scene, half of the display was always invariant and predictive of target location. What varied was whether the target was embedded within the invariant, predictive side or within the random side. Contextual cuing was only observed when the side containing the target was invariant, suggesting that local context is sufficient, and that random local context is not.

Second, Jiang and Chun (2001) explored the role of selective attention in implicit learning of background context information. Jiang and Chun presented displays of rotated L distractors. Half were colored green and the other half were colored red. Each subject had a target color that was red or green, and they were instructed to always attend to that color because the rotated T target never appeared in the unattended color. Thus, for any given display of intermixed red and green items, half of the items were attended and the other half was unattended. Jiang and Chun varied whether the attended context (spatial layout of distractors) was repeated or whether the unattended context was repeated. Only the attended displays produced contextual

cuing; unattended items did not, even though they were repeated the same number of times as the attended items, and even though all of the items were interleaved with each other. This finding demonstrates the importance of selective attention in controlling learning, even implicit learning, to items of behavioral relevance. Thus, in the real world, we propose that when contextual information is encoded, such learning is restricted to the subset of items within a complex scene that is most relevant to the ongoing task.

Broadly speaking, contextual cuing illustrates the importance of learning and memory mechanisms in visual perception. The predictive context information was learned as subjects performed the search task. In other words, observers encoded the invariant visual information that benefited target detection. We propose that such learning occurs most of the time that observers are interacting with their visual environment. However, learning is not indiscriminate and it does not have infinite capacity. Thus learning is strongest for local context and especially for attended information. Not all that repeats gets encoded.

B. How does shape context cue an object?

Another key feature of scenes is that they contain objects that tend to co-occur with each other. Modern day classrooms contain desks, chairs, whiteboards, and computer projection systems, and they are unlikely to contain bottles of scotch or ashtrays. Such statistical information provide another form of “structure” that should be useful for the observer. Importantly, covariation information acquired through perceptual experience allows each object within a scene to cue the presence of other related objects.

We studied this in the lab using novel shapes (Chun & Jiang, 1999). Subjects searched for a target that was the only shape in the display that was symmetric around the vertical axis. The other distractors were novel shapes symmetric around a non-vertical axis (See Figure 3). Thus, we were able to define a target task without specifying or labeling the precise shape of the target, which could be any one of a large number of vertically symmetric shapes. Upon target detection, subjects pressed a key as quickly as possible, and their response time was measured. The display was then replaced with an array of probe letters, each appearing in a location previously occupied by an object. Subjects reported the probe letter that appeared in the same location as the target on the prior search display. The probe task simply allowed us to ensure that the target was properly identified.

Insert Figure 3 about here

We controlled the statistics of this novel visual world by varying whether the target shape was correlated with its distractor shapes (old condition) or whether the target and distractor shapes were not correlated (new condition). In other words, target shapes were consistently mapped to distractors in the old condition, and variably mapped in the new condition (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). If subjects are sensitive to covariation information, they should be faster in the old condition, and indeed, they were. Importantly, the locations of targets and distractors were completely random in this experiment, so that any cuing effects could be attributed to shape association learning alone. Presumably, this type of learning subserves the intra-object priming effects observed with real-objects (Henderson, 1992; Henderson, Pollatsek, & Rayner, 1987).

C. How does ongoing temporal context cue an upcoming event?

Spatial layout and shape association information are prominent features of static scenes, but they fail to encompass the fact that the visual environment is dynamic. Not only do objects move about within scenes, our perception of scenes changes from moment to moment as we navigate around them. Thus, there is rich temporal structure in the environment that may guide our expectations for what will happen in future time steps.

First, let's consider situations with moving objects. A classic example would be a basketball or soccer game where players move about along with the ball. The movements are obviously not random, and moreover, there are regularities not only in how a single player may move, but how the field of players moves relative to the ball. Effective athletes have what is called "field sense," which basically refers to their above-average ability to read the dynamic field of players to predict how key players will move and where the ball will go in the next time step. This ability is not just an index of natural talent but also of perceptual experience, which tunes the player to important regularities in how plays unfold during the game.

We studied this in a dynamic search task, where subjects were asked to quickly detect a T target that was moving about amongst other moving L distractors (Chun & Jiang, 1999). The movements of all of the items were independent and seemingly random with the constraint that they could not run into each other. However, for half of the displays, the target trajectory was perfectly correlated with its distractor trajectories, such that the dynamic context of moving distractor items cued the target trajectory. For the other half of the displays, the target trajectory was not correlated with the distractor

trajectory. Although the displays were seemingly quite arbitrary, subjects were faster to detect targets appearing along trajectories that were correlated with their distractor trajectories. They demonstrated contextual cuing from dynamic displays without awareness of which dynamic display was old and which was new.

Another form of temporal context exists in how visual events change and unfold over time, even in the absence of explicit motion in the display. Namely, an invariant sequence of events forms a temporal context that benefits visual processing for upcoming events. Olson and Chun (2001) presented sequences of letters and varied whether the letter identities appeared in a fixed sequence or randomly. When the onset of the target letter was preceded by a fixed sequence of letter identities, subjects detected the target more quickly. Thus, when visual events unfold in a previously experienced manner, then the sequential information helps observers predict what's forthcoming. Such temporal context learning undoubtedly benefits everyday perception.

How do subjects acquire such temporal associative information? Fiser and Aslin (2002) demonstrated that subjects are tuned to transitional probabilities between successive shapes. In fact, even passive viewing allowed observers to extract temporal correlations from an ongoing stream of different visual shape sequences. Thus, the acquisition of temporal structure may be understood as a problem of statistical learning, important for both the visual and auditory domains (Saffran, Johnson, Aslin, & Newport, 1999).

D. Scene structure and contextual cuing

To sum, our perceptual environment is highly structured, such that knowledge of such structure, presented in the form of visual context, may guide perceptual processes

to rapidly orient to a location, identify an object, or prepare for an upcoming event. The meaningful regularities in the environment may be extracted and internalized using powerful statistical learning mechanisms within the brain. Contextual cuing is a paradigm for studying how regularities are learned through perceptual experience, and how such visual knowledge facilitates behaviors such as search. Understanding the neural mechanisms that encode such regularities should provide insights into how the brain stores visual knowledge for everyday perception.

III. Issues for the study of scene recognition and learning

In the following sections, we will discuss three issues that deserve further research. For each topic, we will summarize outstanding problems, review existing work, and outline directions for future investigation.

A. How are scenes represented?

What is the nature of scene representations in the mind? This seemingly basic question does not have a straightforward answer. We will divide our discussion into two sections. The first concerns whether scenes are more critically defined by the collection of objects they contain or whether the background configuration is important. The second section develops a dual-path model of scene processing that is based on evidence that spatial layout information and object association information may make separable contributions to scene recognition and may have dissociable substrates in the brain.

1. Objects or Background?

Are scenes merely collections of co-occurring objects or is the background structure of a scene important as well? This question has been traditionally asked by studies that probe the effects of scene context on object recognition. In addition, novel insights have recently been obtained from functional neuroimaging.

Consider an office scene. An office contains objects that co-occur in the real world: chairs, computers, telephones, pens, papers, books, etc. In addition to these objects, offices typically contain a certain background structure: four walls, floor, ceiling, windows, and perhaps some built-in bookshelves and desk countertops attached to the wall. This background structure depicts a sense of 3-dimensional space within which objects can be arrayed in coherent spatial relations to each other. Of course, in principle, the distinction between object and background is much less clear than described above. However, to start, we wish to follow the convention that objects tend to be things that either move around or can be moved around, while backgrounds depict more stable, fixed entities, thus providing reference points to define the space they appear in (Boyce, Pollatsek, & Rayner, 1989).

With such a distinction in hand, researchers differ in the relative importance they place on objects versus backgrounds in defining scenes and in understanding scene context effects. Several authors propose that global scene information, formally called “schemas” or “frames,” is extracted based on the overall spatial organization of objects appearing within a background context. Such information may be extracted even before individual objects are identified. The schemas serve to facilitate recognition of the embedded objects (Antes & Penland, 1981; Biederman et al., 1982).

Alternatively, scene recognition and scene context effects may be dependent on recognition of the objects that typically comprise a scene (Friedman, 1979; Henderson, 1992; Henderson et al., 1987). Scene context facilitation of object identification would occur by priming from other objects within the scene. Scene recognition itself is largely driven by rapid identification of diagnostic objects within scenes (e.g., an oven to define a kitchen scene, or a car for a garage scene).

Boyce et al. (1989) supported the schema hypothesis to explain scene context effects on target facilitation. Namely, global background information appeared to be more critical than surrounding objects. They observed that objects were more difficult to identify within a semantically inconsistent background even when related objects were present. Moreover, for their displays, whether simultaneously presented objects were related or unrelated did not matter.

Other studies support an intra-level object-to-object priming account (de Graef, 1992; Henderson et al., 1987). This account is based on facilitation effects observed from related objects that were fixated prior to the target object (Henderson et al., 1987). Even when spatial layout was unstructured, extended viewing of a scene containing statistically correlated objects yielded robust item-to-item priming effects (Chun & Jiang, 1999).

The answer to this debate perhaps lies in between the two accounts (de Graef, 1992). Within the first few hundred milliseconds of analysis of a scene, it is likely that global scene properties, which may include diagnostic color information (Oliva & Schyns, 2000), are rapidly registered and used to guide exploration of the scene (Chun & Jiang, 1998; de Graef, 1992; Henderson, Weeks, & Hollingworth, 1999; Oliva &

Schyns, 2000; Schyns & Oliva, 1994). Thus, experimental studies that rely on briefly flashed scenes are more likely to observe global schema effects rather than local object priming effects. As scene viewing progresses across multiple fixations, object-to-object priming is likely to augment how the scene is processed and how component objects are identified. We will develop this idea in further detail below.

2. A dual-path model of scene processing

It seems likely that global spatial structure and object shape covariation information make joint contributions to the recognition of scenes as well as objects within scenes. This is reasonable given that scenes contain both spatial layout and object shape information. However, are spatial layout information and shape information stored in an integrated manner or are the internal representations for these somewhat independent? This question immediately brings to mind the popular "what" versus "where" distinction, where spatial information is processed primarily through a dorsal pathway, and object information through a ventral pathway (Ungerleider & Mishkin, 1982). Although the distinction is not absolute, it has proven useful for understanding how spatial or object shape information may make separable contributions to a variety of behavioral tasks. For example, damage to the dorsal pathway impairs the ability to utilize spatial cues in a choice task while damage to the ventral pathway impairs the ability to use shape cues (Pohl, 1973). In working memory, holding spatial locations in mind typically activates the dorsal stream while holding object shape information in mind activates the ventral stream (Kohler, Kapur, Moscovitch, Winocur, & Houle, 1995).

The dorsal versus ventral stream distinction does not map directly on how scenes may be represented in long-term memory, but it is interesting to note that there

is some evidence that spatial and object shape information in scenes may be stored in anatomically distinct regions of medial temporal cortex.

For example, the brain area that is sensitive to scene stimuli appears to care more about spatial structure than component objects. In a seminal neuroimaging study that characterized the parahippocampal place area (PPA), Epstein and Kanwisher (1998) demonstrated that the neural activity in this region was substantially higher for an "empty" room than for a 2-dimensional array of multiple related objects (e.g., furniture from a room on a blank background that lacked 3-dimensional spatial context). Based on this and other converging evidence, they concluded that the PPA was most sensitive to information that depicted the layout of local space.

Then where are object associations stored? One promising candidate is perirhinal cortex, which is located at the ventromedial aspect of the primate temporal lobe. It plays an important role in both the perception and memory of objects, especially associations among objects (Gaffan & Parker, 1996; Murray & Bussey, 1999; Murray & Richmond, 2001). Although most work in this cortical region has been conducted in non-human primates, our lab is currently pursuing a number of hypotheses to establish a role for perirhinal cortex in object association learning.

We believe that the behavioral work and neurophysiological data reviewed here points to a dual-path model of scene recognition. Soon after a scene comes into the eyes, global features of its spatial layout that depict 3-dimensional space will activate the parahippocampal place area. This initial "gist" is available within 200 ms (Thorpe, Fize, & Marlot, 1996), even when a mask is present. The global information serves to guide further exploration of the scene (Chun & Jiang, 1998; de Graef, 1992; Henderson

et al., 1999; Oliva & Schyns, 2000; Schyns & Oliva, 1994). As interrogation of a scene progresses, multiple eye movements will foveate different objects within a scene. The sequential pattern of these highly detailed object fixations will activate object representations in temporal areas such as perirhinal cortex, where activation will spread on to neuronal representations of other associated objects. These two streams of information should interact with each other, such that global spatial information processed in the PPA may guide the deployment of eye movements and access to associated object shape information in perirhinal cortex. In turn, object shape information may help the PPA to discriminate one local layout from another, as well as cue the presence of other objects within the scene based on associative knowledge stored in perirhinal cortex.

B. How do people learn environmental regularities in scenes?

A very important question that is related to the issue of scene representation is to understand how people encode scenes from perceptual experience. More broadly speaking, how do observers encode important environmental regularities? One thing that we do know about scene memory is that it is exceptionally good. Behavioral studies have revealed that observers can recognize thousands and thousands of scene images that were novel to them prior to a brief study phase (Shepard, 1967; Standing, 1973; Standing, Conezio, & Haber, 1970). Although such memory performance probably relies more on scene gist rather than a detailed engram, it is still remarkable how many scene images can be encoded, sometimes even based on a single trial of exposure. Furthermore, we suspect that remarkable scene memory performance measured in such recognition tasks may actually be a gross underestimate of the

brain's capacity to encode and discriminate scene information. We base this conjecture on the hypothesis that conscious recognition memory, measured in these prior studies, has smaller capacity than that of unconscious, implicit recognition memory.

A considerable bulk of memory research is organized around the distinction between explicit and implicit memory (Roediger, 1990; Schacter, 1987; Squire, Knowlton, & Musen, 1993). Explicit (declarative) memory supports the ability to consciously retrieve and declare past facts and events. Implicit (nondeclarative) memory supports improved performance in a variety of perceptual and motor tasks, although observers cannot recall or articulate the learned information. The basic feature of implicit memory is that much information that cannot be consciously retrieved can produce effects on behavior due to prior exposure. In fact, amnesic patients with very little explicit memory show intact implicit memory for a variety of perceptual and motor tasks (Cohen & Squire, 1980; Corkin, 1968). Thus, implicit memory may be more sensitive than explicit memory in revealing traces of past experience. Another related feature of implicit memory is its robustness over time. Information that fades away from explicit retrieval over time may be accessed with implicit memory tasks (Cave, 1997; Cave & Squire, 1992; Jacoby & Dallas, 1981; Tulving, Schacter, & Stark, 1982).

Returning to scene context learning, our own lab's work on contextual cuing also shows that "scene" memory can be remarkably powerful, even for the rather sparse, similar-looking displays. Another interesting key feature of contextual cuing is that it is implicit (Chun & Jiang, 1998, 1999, 2003; Olson & Chun, 2001). Most observers do not consciously notice the predictive relationship between repeating contexts and embedded target locations or identities. In fact, most subjects do not even notice that

scene layouts or object shapes were repeating. When probed to explicitly discriminate old displays from new displays, subjects performed at chance. Even when alerted to the fact that displays were repeated and should be noted, subjects did not show more contextual cuing or better performance on the explicit recognition task. (Chun & Jiang, 2003). Fiser and Aslin (2001, 2002) have also observed that subjects may implicitly learn important statistical regularities from structured spatial arrays or temporal sequences of visual objects.

Such implicit learning is perhaps essential for visual perception, because as a number of authors have argued (Lewicki, 1986; Reber, 1989), implicit learning allows the learner to extract statistical regularities in a more efficient manner than may be possible through explicit learning. As noted above, a practical feature of implicit learning is that it tends to be more robust and sensitive than explicitly learned information. For example in the spatial contextual cuing task, it is quite remarkable to observe such a specific contextual cuing effect based on 12 arbitrary artificial scenes that were not discriminable from the other novel scenes they appeared with. Even more notable is the finding that such implicitly learned artificial scene information may persist for up to an entire week (Chun & Jiang, 2003).

Characterizing contextual scene learning as implicit need not imply that different mechanisms or brain systems should be involved for implicit perceptual learning versus conscious, explicit perceptual learning. Indeed, an amnesic patient study suggested that explicit and implicit learning may share the same neural substrates. Chun and Phelps examined contextual learning in amnesic patients with damage to the hippocampus, which is a brain structure important for encoding relational, configural

information, critical for a variety of memory tasks such as spatial learning, contextual learning, and episodic encoding (Cohen & Eichenbaum, 1993; Hirsh, 1974; McClelland, McNaughton, & O'Reilly, 1995; O'Keefe & Nadel, 1978; Rudy & Sutherland, 1994). However, in humans, the hippocampus and neighboring medial temporal lobe structures are also essential for explicit, declarative memory (Squire, 1992), such that damage to these structures produce profound amnesia. In contrast, implicit memory, as expressed in perceptual priming studies or motor skill learning tasks, relies on other non-hippocampal brain structures. Does this mean that spatial contextual cuing, which requires spatial learning but is also implicit, does not rely on the hippocampus? Interestingly, Chun and Phelps (1999) demonstrated that amnesic patients with hippocampal and neighboring medial temporal lobe damage were impaired in their ability to benefit from repeating spatial layouts. The patients showed no contextual cuing, suggesting that the hippocampus and neighboring structures are important for spatial scene learning, regardless of whether the learning is conscious or unconscious.

The Chun and Phelps (1999) finding supports views that the hippocampus is important for configural and relational processing. However, further work is needed. One complication is the finding that partial hippocampal damage is not sufficient to observe contextual cuing impairments (Manns & Squire, 2001), suggesting that complete hippocampal damage is necessary to observe a deficit. Given that the hippocampal patients in the Chun and Phelps study had damage that also extended into other medial temporal lobe structures, it is possible that these other areas play a critical role in contextual cuing. However, a recent neuroimaging study has provided further evidence for hippocampal involvement (Preston, Salidis, & Gabrieli, 2001). Thus, the

hippocampus is likely to be essential for spatial contextual learning, independent of whether other medial lobe structures also contribute or not.

Another limitation is that the amnesic patients were only tested with the spatial context task. Thus, it is possible that other non-spatial forms of implicit configural learning may not be impaired by hippocampal damage. It would be very useful to test the object shape contextual cuing task in a group of amnesic subjects with hippocampal damage. If the hippocampus is important for any type of contextual, configural learning, then the patients should not show object contextual cuing. However, if the hippocampus is only relevant for configural learning that involves spatial relations, then hippocampal patients should show normal object contextual cuing. Following similar logic, it would be useful to test hippocampal patients in the temporal contextual cuing tasks as well. An advantage of the contextual cuing paradigm is its flexibility to test spatial, object, and temporal factors separately. Thus, further studies with the contextual cueing task promise to yield further insights into how different components of scene memory are represented in long-term memory.

C. Does scene context facilitate object recognition?

As reviewed throughout this chapter, one of the most basic functions of scene context and gist is to drive eye movements and attention towards objects relevant to a scene. Eye fixations tend to cluster around regions of interest within scenes and to objects relevant to an ongoing task (Loftus & Mackworth, 1978; Mackworth & Morandi, 1967; Yarbus, 1967). Detection of changes, which requires attention, within scenes tends to be faster for features that are central to the context of a scene than for features that are less central to the context of a scene (Kelley, Chun, & Chua, in press; Rensink

et al., 1997; Shore & Klein, 2000). These findings can be extended to hypothesize that context directly facilitates the identification of consistent objects within a scene. Thus, Palmer (1975) demonstrated that the scene context of a kitchen enhanced recognition of an embedded breadbox as opposed to a drum. Biederman (1982) showed that subjects were better at detecting objects appearing in valid locations compared to invalid locations. Even using novel shapes, targets that were consistently paired with their context were detected more rapidly than those that were not. In sum, it would seem a foregone conclusion that scene context facilitates object recognition in an interactive manner.

Unfortunately, despite considerable work on this topic, a fundamental question about this basic hypothesis remains unresolved: Where is the locus of contextual effects on object perception? Does scene context bias an early stage of visual processing by biasing feature extraction? Or does it operate on higher-level representations, at the stage where perceptual representations are matched with stored descriptions of known objects? Or is scene knowledge completely isolated from object identification processes? Although prior work may appear to support the former two possibilities that place scene context effects on object recognition stages or earlier, recent studies have questioned this assumption with evidence showing that scene context effects may reflect response bias or selective encoding, rather than facilitated perception.

A wide variety of paradigms have been used to address this question, but each has specific problems, as reviewed by Henderson and Hollingworth (1999). First, in eye movement paradigms, the dwell time of fixation on an object may be interpreted as one index of object recognition efficiency. Thus, shorter fixations may be predicted for

objects consistent with their global scene context. The problem with such measures is that evidence for shorter fixations on scene-consistent objects is not clear, at least not for the first fixation within a scene. A more fundamental problem is that fixation may reflect the contribution of other mental processes beyond perception, such as an increased difficulty of remembering the item for later report or the increased time involved to cognitively assimilate an item that is incongruous with its surrounding context. Thus, eye movement measures, at least as they have been used in the past, may not afford decisive insights into the locus of scene context effects. This problem generalizes to other methods such as naming tasks, which provides response times that reflect other additional cognitive processes beyond perceptual recognition.

Given these problems with eye movement and naming measures, object detection paradigms appear more promising, at least for understanding object facilitation effects. In detection tasks, experimenters measure the accuracy of detecting a target object appearing within a briefly presented scene. A classic study demonstrated that objects appearing within intact scenes were more accurately detected than objects appearing within jumbled scenes (Biederman, 1972). One may also measure response time to objects within scenes. Accordingly, subjects take less time to find a target object within a normal scene than in a jumbled scene (Biederman, Glass, & Stacy, 1973). Although Biederman's early studies demonstrated the importance of coherent scene context, one limitation is that the findings may instead reflect an "incoherent scene disadvantage," given that the jumbled scenes introduced new contours, confounding visual complexity between intact and jumbled scenes.

Such concerns may be addressed by exploring object recognition within coherent scenes only. To manipulate scene context effects, one may vary whether the target object is consistent or inconsistent with the scene (Loftus & Mackworth, 1978; Palmer, 1975). Broadly speaking, inconsistent objects may be incongruous with scene context in their identity (a camel in a restaurant) or in their spatial position (a chair glued to the ceiling in an office scene) or both (a sofa floating in the sky of an outdoor city scene). Using signal detection measures, early studies showed that the advantage for consistent objects (Biederman et al., 1982; Biederman, Teitelbaum, & Mezzanotte, 1983) reflected higher sensitivity, a measure of perceptual discriminability, rather than bias, a measure of postidentification decision processes. However, this finding has been sharply criticized by Hollingworth and Henderson (1998) who demonstrated a problem in the experimental design that affected how perceptual sensitivity was calculated. Using a corrected design, Hollingworth and Henderson not only replicated Biederman et al's results using their original uncorrected design, they demonstrated that the advantage of context-consistent objects disappeared when the design was corrected. If anything, Hollingworth and Henderson (2000; 2001) have repeatedly observed an inconsistent object advantage, which they attribute to postperceptual selective encoding in memory. Bolstering a postperceptual explanation, Henderson and colleagues (Henderson et al., 1999) demonstrated that inconsistent objects were fixated longer, but not earlier than consistent objects during scene viewing. In sum, they favor a functional isolation model that posits that scene knowledge and object perception processes are segregated. Evidence for interactions between global scenes and embedded objects may reflect cognitive processes occurring beyond recognition, such

as guessing strategies or selective encoding strategies. In sum, current behavioral evidence is very mixed in regards to whether scene context facilitates object recognition or not.

My opinion is that scene context effects occur at both perceptual and postperceptual stages. Different tasks and dependent measures may reveal scene context effects at different levels of perceptual and cognitive processing. Thus, this question should be approached with a variety of methodologies. In particular, cognitive neuroscience methods that look into brain activity may provide novel insights, as I will review below.

To resolve the issue of how scene context influences object recognition, one must consider both anatomical and temporal factors. Anatomically speaking, scene context may influence object recognition at an early or late stage of visual processing. Early stages may include areas in temporal cortex, where object shape information is processed, and they may even include the earliest stages of visual analysis, such as areas V1, V2, and V4, where features are initially extracted from the incoming image. Conversely, scene context may not influence visual processing in the occipital or temporal cortex at all. Instead, one may only observe effects of context in frontal areas that are not specialized for visual analysis, but are more involved in working memory and response selection.

In conjunction with such anatomical factors, one may consider the time course of contextual influences as well. For example, does contextual information modulate stimulus processing as sensory information passes through visual areas, say, within

200 ms of stimulus onset? Or are contextual influences observed at a later latency that may be more consistent with postperceptual processes?

There are a variety of methods to probe the anatomical and temporal characteristics of contextual processing in the brain. We will consider three here. First, single-cell neurophysiology affords insights into contextual influences with very high spatial and temporal resolution. However, such methods are not typically available to study activity in human cerebral cortex. For human studies, there are two non-invasive methodologies that are popularly used. Event-related potentials measure stimulus and task relevant neuronal activity that can be recorded at the scalp. Although anatomical resolution is poor, temporal resolution is high. Complementary insights may be obtained from functional neuroimaging methods such as positron emission tomography (PET) or functional magnetic resonance imaging (fMRI). These methods measure changes in blood flow that correlate with neural activity. They afford more anatomical precision than ERP methods, while lacking temporal precision. The anatomical precision can be quite revealing in the case of fMRI.

When one considers the neurophysiological evidence in the literature, it becomes abundantly clear that some form of scene context benefits perceptual processing, at a fairly short latency within the earliest of visual cortical areas: V1. However, the meaning of “scene” becomes critical here, as most work has focused on processing low-level features using stimuli that do not resemble the natural scenes we typically encounter in the world. Nevertheless, if one may (momentarily) allow a collection of discrete items in an array to be called a scene, then one will find that such scene context influences processing of items within it. Consider the neural response of a cell in V1 that is

optimally tuned to an oriented line (target) within its receptive field. If the target is the only item within the display, then its orientation will determine the strength of the neural response because V1 neurons are orientation sensitive. Of course, the neuron only responds to stimuli within its receptive field. If the target is presented outside the neuron's receptive field, no response is observed, and no modulation is observed as the target moves around outside the receptive field. However, if the target is in the neuron's receptive field, and there are other items in the context of the target, outside of the receptive field, then an interesting result emerges. As the orientation of the items in the context deviates from the target orientation, the neuron's response increases. For example, the neuronal response to a vertical target is maximal when the target is surrounded by a field of horizontal lines, and it is weakened when the surrounding field is also vertical. It is as if the neuron fires to permit "pop-out," rapid segregation of the target feature relative to the background (Knierim & van Essen, 1992). What's remarkable is that such influences are being driven by stimuli outside the target's receptive field. In addition, the latency of such influences is rapid, occurring within 20 ms of stimulus array onset. Such long-range interactions in visual cortex may provide the foundation for psychophysical observations that revealed how thresholds for discriminating faint, oriented visual targets are dependent on interactions with other stimuli that spatially flank the target (Polat, Mizobe, Pettet, Kasamatsu, & Norcia, 1998; Polat & Sagi, 1993, 1994).

Similar observations of contextual influences in V1 have been observed for visual surfaces as well. When the orientation of lines within a target surface patch is different from the texture of lines in the background of the target surface patch, the neural

response to the lines within the target surface patch become enhanced, supporting the sense of perceptual segregation experienced from such displays (Lamme, 1995; Zipser, Lamme, & Schiller, 1996; but see Rossi, Desimone, & Ungerleider, 2001).

Of course, most people will resist calling these artificial displays scenes. In fact, the mechanisms described above most likely play a role in low-level visual processing, promoting texture segregation and feature pop-out. The point that I wish to draw is that one of the most fundamental stages of visual processing harbors neural mechanisms to support highly interactive processing. No feature is processed in isolation of another, and this fact encourages the search for similar processing principles within higher levels of visual processing.

One attempt to do so employed the contextual cuing paradigm. Olson, Allison, and Chun (2001) had the opportunity to collect electrophysiological recordings directly from the cortical surface of patients who were being monitored for epileptic seizure foci. We trained a group of patients on a set of spatial contexts that predicted the embedded target location. The patients showed a significant contextual cuing effect, faster detection of targets appearing in old contexts compared to targets appearing in new contexts. Because no other visual cues existed to distinguish old from new contexts, the search benefit must have been driven by learned context information. Thus, any difference in neural activity to old scenes versus new scenes must reflect some process that distinguishes the two types of trials, leading to faster detection. Olson et al. observed significant differences in the N210 component of the ERP waveform to old vs. new scenes. Thus, this finding demonstrates that learned context information can influence neural processing within 210 ms of stimulus onset. Moreover, the relatively

higher resolution of intracranial recordings permitted Olson et al. to demonstrate that much of this differential activity occurred in early visual areas such as V4, V2, and perhaps even V1. The latency of the N210 is such that it probably does not reflect modulation of activity within the initial volley of visual information through visual cortex, but rather backward feedback from higher-level stages, presumably scene representations in medial temporal cortex. Unfortunately it is not clear what the N210 is revealing: whether it simply reflects the discrimination of old vs. new displays or whether it signals the top-down control of spatial attention to the target associated with an old context. Much further work is needed. Nevertheless, this study provides some of the clearest evidence that learned context information can induce changes in neural activity within 210 ms in early visual areas.

At higher stages of visual processing, there is less direct neural evidence for contextual interactions. However, the potential for contextual influence seems high. Consistent with the dual-path model of scene processing, the first step of scene context effects is likely to be rapid recognition of global scene context and configuration information. Behavioral work has shown that scene recognition is very efficient, based on Potter's (1975) finding that the gist of a target scene can be reliably extracted from an rapid ongoing stream of different scenes. Still, behavioral work cannot pinpoint the time course of scene processing because categorization processes progress even after the stimulus is no longer present. ERP measures can provide more direct measures, and it is very interesting that ERP signals begin to distinguish scene categories by 150 ms after stimulus onset (Thorpe et al., 1996). A follow-up of this study used fMRI to

reveal that differential activation for target and distractor scenes occurs in high-level visual areas such as the fusiform and parahippocampal gyri (Fize et al., 2000).

Such solid evidence for rapid scene categorization makes it tempting to postulate that scene information develops in parallel with object information in a way that the two streams of information interact throughout the visual pathway. The next step is to establish that such scene information impacts the representations of embedded objects. Such interactions must be based on associative links between objects that tend to appear together such that the presence of one object cues the presence of the other. Towards this goal, one must demonstrate associative learning in temporal cortex, where object knowledge is thought to reside. One of the most classic studies to do so was a neurophysiological study by Miyashita and colleagues (Miyashita, 1988; Sakai & Miyashita, 1991). By training monkeys on novel visual shapes, they first showed that neurons in inferotemporal (IT) cortex become shape-selective with learning. In addition, they demonstrated that these neurons became selective to other temporally associated but geometrically unrelated stimuli. Presumably, this type of associative learning would assist the neuron's ability to link different views of the same object (Logothetis & Pauls, 1995), in addition to linking different objects that typically co-occur with each other. Of further interest is the recent suggestion that visual experience may induce the development of clusters of neurons with similar stimulus preferences (Erickson, Jagadeesh, & Desimone, 2000).

One limitation of these past studies of associative learning in visual cortex is that they were limited to temporal associations. Namely, a cue stimulus was temporally correlated with a stimulus that trailed in time. However, with respect to the dual-path

model of scene recognition, temporal cuing may play a central role, as most objects in complex scenes are fixated in a serial manner. Nevertheless, it would be important to extend these insights to understand how simultaneously presented object shapes may influence the neural activity, and corresponding behavioral response, to a target shape. Our lab is currently testing fMRI tasks that examine stimuli sets that are temporally associated and/or spatially associated, and we believe that the results will further clarify how scene context facilitates object recognition within visual processing areas in temporal cortex.

IV. Summary Remarks

Scenes are complex, but this complexity provides a rich source of contextual information that constrains visual processing in a useful manner. In particular, scenes contain many regularities in their spatial layout, object shape correlations, and dynamic features. Encoding such statistical regularities allow the observer to use ongoing contextual information to constrain their search and identification of visual objects relevant to behavior. Much scene learning appears to occur implicitly such that past experience with scenes and scene properties may influence behavior even when the observer is not consciously aware of having seen them before. We believe that implicit measures of scene memory reveal a prodigious visual memory capacity that is at least as large, if not larger than, the rich capacity for distinguishing previously viewed scenes, as measured through explicit recognition measures.

To understand how such environmental regularities are represented in the brain, it is useful to consider both behavioral and neuroscientific data. Past findings appear to converge to support a dual path model of scene processing, where global spatial

configuration information is rapidly registered and used to guide how a scene is interrogated with multiple eye movements. As fixations move from one object to the next, each object serves to define the scene as well as prime expectancies for other objects within a scene. In addition, neuroscience studies suggest that global spatial configuration information may be represented separately from object association information in the brain throughout medial temporal cortex. A rich theory of visual processing will emerge through understanding how scene knowledge is acquired, how scene knowledge is represented, and how scene knowledge interacts with early perceptual and late response selection mechanisms.

References

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

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). Neural components of topographical representation. Proceedings of the National Academy of Sciences of the United States of America, 95(3), 839-846.

Antes, J. R., & Penland, J. G. (1981). Picture context effects on eye movement patterns. In D. F. Fisher & R. A. Monty & J. W. Senders (Eds.), Eye movements: cognition and visual perception. Hillsdale, NJ: Erlbaum.

Biederman, I. (1972). Perceiving real-world scenes. Science, 177(4043), 77-80.

Biederman, I., Glass, A. L., & Stacy, E. W., Jr. (1973). Searching for objects in real-world scences. Journal of Experimental Psychology, 97(1), 22-27.

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

Biederman, I., Teitelbaum, R. C., & Mezzanotte, R. J. (1983). Scene perception: A failure to find a benefit from prior expectancy or familiarity. Journal of Experimental Psychology: Learning, Memory, & Cognition, 9(3), 411-429.

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

Bravo, M. J., & Nakayama, K. (1992). The role of attention in different visual-search tasks. Perception & Psychophysics, 51, 465-472.

Broadbent, D. E. (1958). Perception and Communication. London: Pergamon Press.

Cave, C.-B. (1997). Very long-lasting priming in picture naming. Psychological Science, 8, 322-325.

Cave, C. B., & Squire, L. R. (1992). Intact and long-lasting repetition priming in amnesia. Journal of Experimental Psychology: Learning, Memory, & Cognition, 18(3), 509-520.

Chua, K.-P., & Chun, M. M. (in press). Implicit spatial learning is viewpoint-dependent. Perception & Psychophysics.

Chun, M. M. (2000). Contextual cueing of visual attention. Trends in Cognitive Science, 4(5), 170-178.

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

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

Chun, M. M., & Jiang, Y. (2003). Implicit, long-term spatial contextual memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, Accepted pending revision.

Chun, M. M., & Marois, R. (2002). The dark side of attention. Current Opinion in Neurobiology, 12, 184-189.

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

Chun, M. M., & Wolfe, J. M. (2001). Visual Attention. In B. Goldstein (Ed.), Blackwell Handbook of Perception (pp. 272-310). Oxford, UK: Blackwell Publishers Ltd.

Cohen, N. J., & Eichenbaum, H. (1993). Memory, amnesia, and the hippocampal system. Cambridge, MA: MIT Press.

Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. Science, 210(4466), 207-210.

Corkin, S. (1968). Acquisition of motor skill after bilateral medial temporal-lobe excision. Neuropsychologia, 6(3), 255-265.

de Graef, P. (1992). Local and global contextual constraints on the identification of objects in scenes. Canadian Journal of Psychology, 46, 489-508.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. Psychological Review, 96(3), 433-458.

Egeth, H. E., Virzi, R. A., & Garbart, H. (1984). Searching for conjunctively defined targets. Journal of Experimental Psychology: Human Perception & Performance, 10(1), 32-39.

Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: recognition, navigation, or encoding? Neuron, 23(1), 115-125.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature, 392(April 9), 598-601.

Erickson, C. A., Jagadeesh, B., & Desimone, R. (2000). Clustering of perirhinal neurons with similar properties following visual experience in adult monkeys. Nature Neuroscience, 3(11), 1143-1148.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. Psychological Science, 12(6), 499-504.

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

Fize, D., Boulanouar, K., Chatel, Y., Ranjeva, J. P., Fabre-Thorpe, M., & Thorpe, S. (2000). Brain areas involved in rapid categorization of natural images: an event-related fMRI study. Neuroimage, 11(6 Pt 1), 634-643.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. Journal of Experimental Psychology: Human Perception & Performance, 18(4), 1030-1044.

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

Gaffan, D., & Parker, A. (1996). Interaction of perirhinal cortex with the fornix-fimbria: Memory for objects and "object-in-place" memory. Journal of Neuroscience, 16, 5864-5869.

Gibson, B. S., Li, L., Skow, E., Brown, K., & Cooke, L. (2000). Searching for one or two identical targets: When visual search has a memory. Psychological Science, 11, 324-327.

Gibson, E. J. (1963). Perceptual Learning. Annual Review of Psychology, 14, 29-56.

Gibson, E. J. (1966). Perceptual development and the reduction of uncertainty. Paper presented at the Proceedings of the 18th International Congress of Psychology, Moscow.

Gibson, E. J. (1969). Principles of perceptual learning and development. New York: Appleton-Century-Crofts.

Gibson, J. J. (1966). The senses considered as perceptual systems. Boston: Houghton Mifflin.

Henderson, J. M. (1992). Identifying objects across saccades: Effects of extrafoveal preview and flanker object context. Journal of Experimental Psychology: Learning, Memory, & Cognition, 18(3), 521-530.

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

Henderson, J. M., Pollatsek, A., & Rayner, K. (1987). Effects of foveal priming and extrafoveal preview on object identification. Journal of Experimental Psychology: Human Perception & Performance, 13(3), 449-463.

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

Hirsh, R. (1974). The hippocampus and contextual retrieval of information from memory: a theory. Behavioral Biology, 12(4), 421-444.

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

Hollingworth, A., & Henderson, J. M. (2000). Semantic informativeness mediates the detection of change in natural scenes. Visual Cognition, 7, 213-235.

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. Journal of Experimental Psychology: Human Perception & Performance, 28, 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 and Review, 8(4), 761-768.

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

Irwin, D. E. (1991). Information integration across saccadic eye movements. Cognitive Psychology, 23(3), 420-456.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. Journal of Experimental Psychology: General, 110, 306-340.

Jiang, Y., & Chun, M. M. (2001). Selective attention modulates implicit learning. Quarterly Journal of Experimental Psychology A, 54A, 1105-1124.

Johnston, W., A., Hawley, K., J., Plew, S. H., Elliott, J. M., & DeWitt, M. J. (1990). Attention capture by novel stimuli. Journal of Experimental Psychology: General, 119, 397-411.

Kelley, T. A., Chun, M. M., & Chua, K.-P. (in press). Effects of scene inversion on change detection of targets matched for visual salience. Journal of Vision.

Knierim, J. J., & van Essen, D. C. (1992). Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. Journal of Neurophysiology, 67(4), 961-980.

Kohler, S., Kapur, S., Moscovitch, M., Winocur, G., & Houle, S. (1995). Dissociation of pathways for object and spatial vision: a PET study in humans. Neuroreport, 6(14), 1865-1868.

Kristjansson, A. (2000). In search of remembrance: Evidence for memory in visual search. Psychological Science, 11, 328-332.

Lamme, V. A. (1995). The neurophysiology of figure-ground segregation in primary visual cortex. Journal of Neuroscience, 15(2), 1605-1615.

Lewicki, P. (1986). Processing information about covariations that cannot be articulated. Journal of Experimental Psychology: Learning, Memory, & Cognition, 12(1), 135-146.

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

Logothetis, N. K., & Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. Cerebral Cortex, 3, 270-288.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. Perception & Psychophysics, 2, 547-552.

Manns, J., & Squire, L. R. (2001). Perceptual learning, awareness, and the hippocampus. Hippocampus, 11, 776-782.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychological Review, 102(3), 419-457.

McConkie, G. W., & Currie, C. B. (1996). Visual stability across saccades while viewing complex pictures. Journal of Experimental Psychology: Human Perception & Performance, 22(3), 563-581.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature, 335(6193), 817-820.

Murray, E. A., & Bussey, T. J. (1999). Perceptual-mnemonic functions of the perirhinal cortex. Trends in Cognitive Sciences, 3, 142-151.

Murray, E. A., & Richmond, B. J. (2001). Role of perirhinal cortex in object perception, memory, and associations. Current Opinion in Neurobiology, 11(2), 188-193.

O'Keefe, J., & Nadel, L. (1978). The hippocampus as a cognitive map. Oxford: Clarendon Press.

O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. Special Issue: Object perception and scene analysis. Canadian Journal of Psychology, 46(3), 461-488.

O'Regan, J. K., Rensink, R. A., & Clark, J. J. (1999). Change-blindness as a result of 'mudsplashes'. Nature, 398(6722), 34.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. Cognitive Psychology, 41(2), 176-210.

Olshausen, B. A., & Field, D. J. (2000). Vision and the coding of natural images. American Scientist, 88, 238-245.

Olson, I. R., & Chun, M. M. (2001). Temporal contextual cueing of visual attention. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 1299-1313.

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

Olson, I. R., Chun, M. M., & Allison, T. (2001). Contextual guidance of attention: ERP evidence for an anatomically early, temporally late mechanism. Brain, 124, 1417-1425.

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

Pashler, H. (1998). The Psychology of Attention. Cambridge, MA: MIT Press.

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

Peterson, M. S., & Kramer, A. F. (2001b). Contextual cueing reduces interference from task-irrelevant onset distractors. Visual Cognition, 8, 843-859.

Peterson, M. S., Kramer, A. F., Wang, R. F., Irwin, D. E., & McCarley, J. S. (2001). Visual search has memory. Psychological Science, 12, 287-292.

Pohl, W. (1973). Dissociation of spatial discrimination deficits following frontal and parietal lesions in monkeys. Journal of Comparative and Physiological Psychology, 82, 227-239.

Polat, U., Mizobe, K., Pettet, M. W., Kasamatsu, T., & Norcia, A. M. (1998). Collinear stimuli regulate visual responses depending on cell's contrast threshold. Nature, 391(6667), 580-584.

Polat, U., & Sagi, D. (1993). Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments. Vision Research, 33(7), 993-999.

Polat, U., & Sagi, D. (1994). Spatial interactions in human vision: from near to far via experience- dependent cascades of connections [see comments]. Proceedings of the National Academy of Sciences, U S A, 91(4), 1206-1209.

Potter, M. C. (1975). Meaning in visual search. Science, 187(4180), 965-966.

Preston, A. R., Salidis, J., & Gabrieli, J. D. E. (2001). Medial temporal lobe activity during implicit contextual learning. Society for Neuroscience Abstracts.

Reber, A. S. (1989). Implicit learning and tacit knowledge. Journal of Experimental Psychology: General, 118, 219-235.

Rensink, R. A. (2000). The dynamic representation of scenes. Visual Cognition, 7, 17-42.

Rensink, R. A. (2002). Change detection. Annual Review of Psychology, 53, 245-277.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. Psychological Science, 8(5), 368-373.

Roediger, H. L., III. (1990). Implicit memory: Retention without remembering. American Psychologist, 45, 1043-1056.

Rossi, A. F., Desimone, R., & Ungerleider, L. G. (2001). Contextual modulation in primary visual cortex of macaques. Journal of Neuroscience, 21(5), 1698-1709.

Rudy, J. W., & Sutherland, R. J. (1994). The memory-coherence problem, configural associations, and the hippocampal system. In D. L. Schacter & E. Tulving (Eds.), Memory Systems 1994 (pp. 119-146). Cambridge, MA: MIT Press.

Ryan, J. D., Althoff, R. R., Whitlow, S., & Cohen, N. J. (2000). Amnesia is a deficit in relational memory. Psychological Science, 11(6), 454-461.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants [see comments]. Science, 274(5294), 1926-1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. Cognition, 70(1), 27-52.

Sakai, K., & Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. Nature, 354, 108-109.

Schacter, D. L. (1987). Implicit memory: History and current status. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 501-518.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing : I. Detection, search and attention. Psychological Review, 84(1), 1-66.

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

Sheinberg, D. L., & Logothetis, N. K. (1998). Implicit memory for scenes guides visual exploration in monkey. Society for Neuroscience Abstracts, Vol. 24, Part 2, 1506.

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

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing : II. Perceptual learning, automatic attending and a general theory. Psychological Review, 84(2), 127-190.

Shinoda, H., Hayhoe, M. M., & Shrivastava, A. (2001). What controls attention in natural environments? Vision Research, 41(25-26), 3535-3545.

Shore, D. I., & Klein, R. M. (2000). The effects of scene inversion on change blindness. Journal of General Psychology, 127, 27-43.

Shore, D. I., & Klein, R. M. (2000). On the manisfestations of memory in visual search, Spatial Vision, 14, 59-75.

Simons, D., & Levin, D. (1997). Change Blindness. Trends in Cognitive Science, 1, 261-267.

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

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. Psychological Review, 99, 195-231.

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

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 picture: Single-trial learning of 2500 visual stimuli. Psychonomic Science, 19, 73-74.

Theeuwes, J. (1992). Perceptual selectivity for color and form. Perception & Psychophysics, 51, 599-606.

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

Treisman, A., & Sato, S. (1990). Conjunction search revisited. Journal of Experimental Psychology: Human Perception & Performance, 16(3), 459-478.

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

Tulving, E., Schacter, D. L., & Stark, H. (1982). Priming effects in word fragment-completion are independent of recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 8, 336-342.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle & M. A. Goodale & R. J. W. Mansfield (Eds.), Analysis of Visual Behaviour (pp. 549-586). Cambridge, MA: MIT Press.

Wolfe, J. M. (1994). Guided Search 2.0 : A revised model of guided search. Psychonomic Bulletin & Review, 1(2), 202-238.

Yantis, S., & Jonides, J. (1984). Abrupt visual onsets and selective attention: Evidence from visual search. Journal of Experimental Psychology: Human Perception & Performance, 10(5), 601-621.

Yarbus, A. L. (1967). Eye movements and Vision. New York: Plenium.

Zipser, K., Lamme, V. A., & Schiller, P. H. (1996). Contextual modulation in primary visual cortex. Journal of Neuroscience, 16(22), 7376-7389.

Figure Legends

Figure 1. The brain image shows a coronal slice of the human parahippocampal place area (PPA), defined as the region (outlined with a black square) with higher activity to scenes than to faces, objects, and scrambled scenes. The bar graph shows the percent signal strength of the fMRI signal, relative to fixation baseline, in the PPA when the subject was viewing scenes, face, scrambled scenes, or scrambled faces. Activity was highest for scenes.

Figure 2. A sample search trial display from the spatial contextual cuing task (Chun & Jiang, 1998). The task was to search for a T rotated to the right or to the left. The L shapes were also rotated in random directions, and the layout of the distractors form a "visual context" around the T target. When the distractor configuration was repeated and correlated with a consistent target position, search performance improved in comparison to displays where the distractor configuration was randomly generated.

Figure 3. A sample search trial display from the object shape contextual cuing task (Chun & Jiang, 1999). The task was to search for a vertically symmetric shape. All of the other shapes were symmetric around a non-vertical axis. When the target shape was correlated with the distractor shapes, then search was faster in comparison to a control condition where the target and distractor shapes were repeated but not correlated with each other.

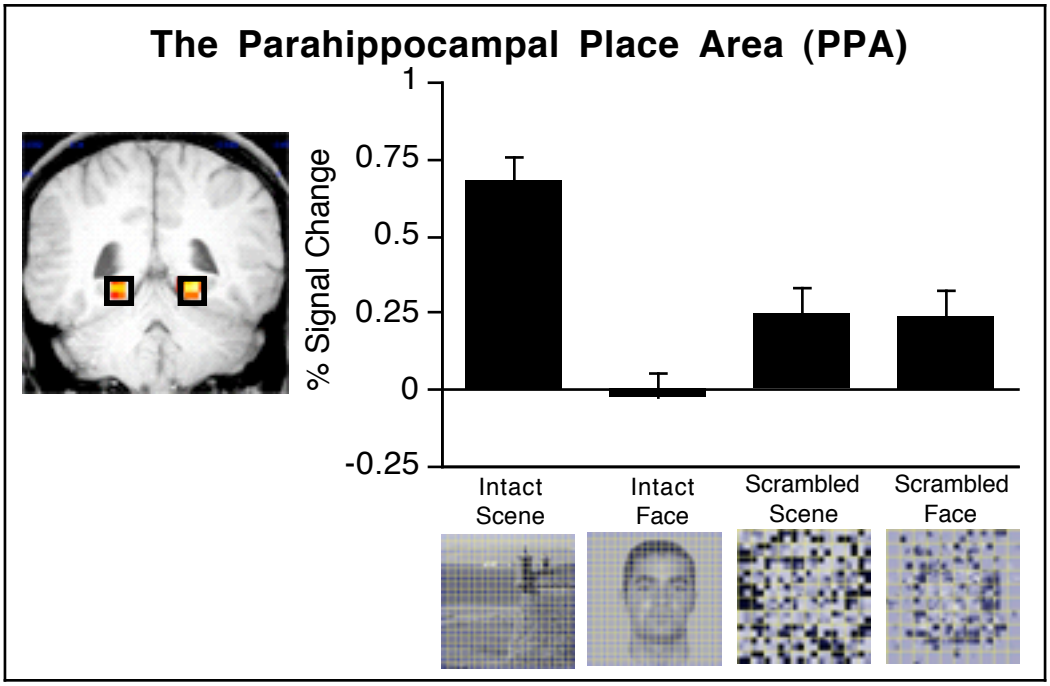

Figure 1 (Chun)

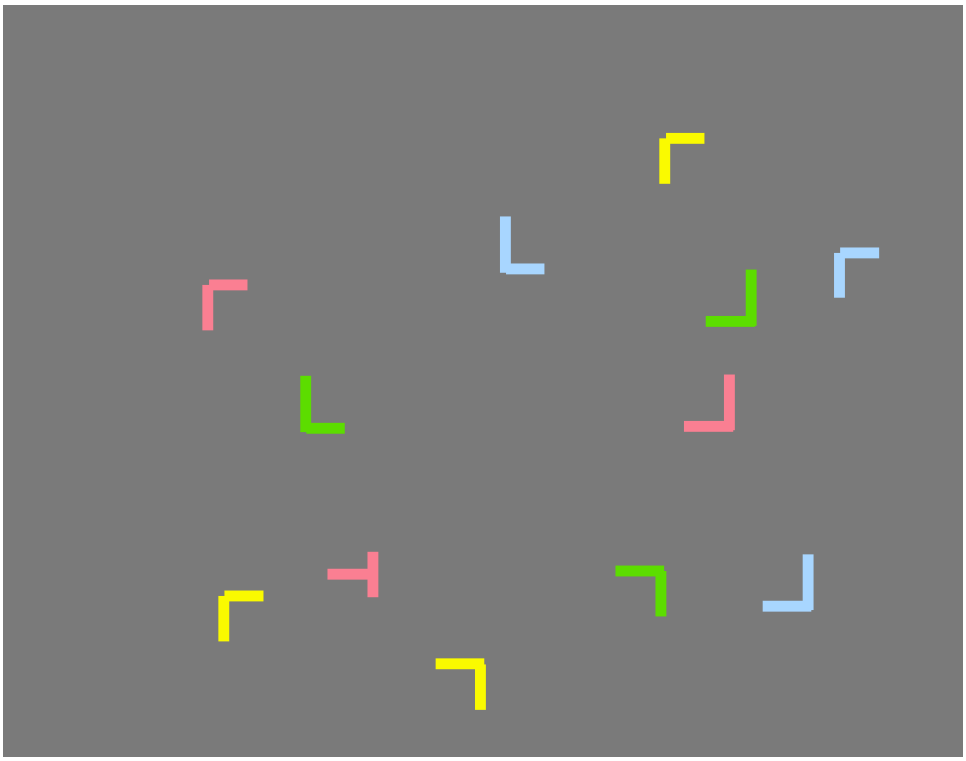

Figure 2 (Chun)

Figure 3 (Chun)