Annu. Rev. Psychol. 1999. 50:243–71

HIGH-LEVEL SCENE PERCEPTION

John M. Henderson and Andrew Hollingworth

Department of Psychology, Michigan State University, East Lansing, Michigan 48824; e-mail: john@eyelab.msu.edu, andrew@eyelab.msu.edu

KEY WORDS: eye movements, vision, scene identification, object identification, change blindness

ABSTRACT

Three areas of high-level scene perception research are reviewed. The first concerns the role of eye movements in scene perception, focusing on the influence of ongoing cognitive processing on the position and duration of fixations in a scene. The second concerns the nature of the scene representation that is retained across a saccade and other brief time intervals during ongoing scene perception. Finally, we review research on the relationship between scene and object identification, focusing particularly on whether the meaning of a scene influences the identification of constituent objects.

CONTENTS

0084-6570/99/0201-0243$08.00

INTRODUCTION

To a first approximation, research in human vision can be divided into three areas of investigation. Low-level or early vision is concerned with extraction of physical properties such as depth, color, and texture from an image as well as the generation of representations of surfaces and edges (Marr 1982). Intermediate-level vision concerns extraction of shape and spatial relations that can be determined without regard to meaning but that typically require a selective or serial process (Ullman 1996). Finally, high-level vision concerns the mapping from visual representations to meaning and includes the study of processes and representations related to the interaction of cognition and perception, including the active acquisition of information, short-term memory for visual information, and the identification of objects and scenes. In this chapter we review three important areas of investigation in the study of high-level scene perception. First, we examine eye movements in scene perception, focusing on the cognitive control of eye movements and the degree to which meaning and ongoing cognitive processes influence eye movement behavior. Second, we review recent work on the nature of the scene representation that is retained across a saccade and other similarly brief intervals during ongoing scene perception. Finally, we review work on the interaction of cognition and perception, focusing on object and scene identification. Although these topics have a long tradition of empirical investigation, they each have received a flurry of new work in the past few years.

In research on high-level scene perception, the concept of *scene* is typically defined (though often implicitly) as a semantically coherent (and often nameable) view of a real-world environment comprising background elements and multiple discrete objects arranged in a spatially licensed manner. Background elements are taken to be larger-scale, immovable surfaces and structures, such as ground, walls, floors, and mountains, whereas objects are smaller-scale discrete entities that are manipulable (e.g. can be moved) within the scene. Clearly, these definitions are neither exact nor mutually exclusive. For example, the distinction between a scene and an object depends on spatial scale. An office scene may contain a desk as one of its component objects. But in a more focused view, the desktop might become a scene, with its surface forming the background and a stapler, phone, and pen serving as individuated objects. It is difficult to determine precisely when the spatial scale becomes too small or too large to call the resulting view a scene. Is the inside of a desk drawer a scene? Is a box of paperclips a scene? Most research on scene perception has avoided this problem of definition by using views of environments scaled to a human size. So an encompassing view of a kitchen or a playground would be considered a good scene, whereas a view of a box of paperclips or an aerial view of a city would not. For the current purposes we adopt this imprecise, intuitive, and

not wholly satisfying definition, holding to the belief that definitions are often best refined as a product of empirical investigation.

EYE MOVEMENT CONTROL IN SCENE PERCEPTION

Because of the optical structure of the eyes, the gradient in cone density in the retina, and the preferential mapping of foveal photoreceptors onto visual cortical tissue, acuity is highest at the point of fixation and drops off precipitously and continuously with increasing visual eccentricity (Anstis 1974, Riggs 1965). The highest-quality visual information is acquired from the region of the scene that projects to the fovea, a region of the retina corresponding to about the central 2º of the viewed scene (about the size of a thumbnail at arm's length). The human visual-cognitive system takes advantage of the high resolving power of the fovea by reorienting the fixation point around the viewed scene an average of three times each second via saccadic eye movements. During a *saccade*, the point of regard sweeps rapidly across the scene at velocities of up to 900º/s as the eyes rotate in their sockets (Carpenter 1988). During a *fixation*, the point of regard is relatively (though not perfectly) still. Pattern information is acquired during the fixations; information useful for ongoing perceptual and cognitive analysis of the scene normally cannot be acquired during a saccade (Matin 1974, Volkmann 1986).

A complete understanding of scene perception requires understanding the processes that control *where* the fixation point tends to be centered during scene viewing and *how long* the fixation position tends to remain centered at a particular location. In this section we review the literature on eye movements during scene perception. The scope of this review is restricted in two important ways. First, we focus on eye movements during the viewing of pictorial representations of static scenes. Eye movements during viewing of dynamic scenes have recently been reviewed by Land & Furneaux (1997). Second, we focus on molar-level eye movement behavior associated with ongoing perceptual and cognitive processing. We ignore, for the purposes of this review, other types of eye movements (e.g. smooth pursuit, vergence, slow drifts, microsaccades, and stabilization reflexes; see Carpenter 1988) as well as stimulus-based oculomotor effects like the global effect (Findlay 1982) and the optimal viewing position effect (O'Regan 1992a). Although these phenomena are important, they represent aspects of eye movement behavior that do not directly reflect ongoing visual-cognitive processing related to high-level scene perception.

Fixation Position During Scene Perception

In a classic study, Buswell (1935) reported the first systematic exploration of the spatial distribution of fixations during scene perception. Two hundred

viewers examined 55 pictures of different types of artwork, such as architecture, sculpture, and paintings, under a variety of viewing instructions. Buswell found that fixation positions were highly regular and related to the information in the pictures. For example, viewers tended to concentrate their fixations on the people rather than on background regions when examining the painting *Sunday Afternoon on the Island of La Grande Jatte* by Georges Seurat. These data provided some of the earliest evidence that eye movement patterns during complex scene perception are related to the information in the scene and, by extension, to ongoing perceptual and cognitive processing.

In another classic study, Yarbus (1967) asked viewers to examine color paintings of scenes and other artwork over extended viewing times. Yarbus found that when viewers examined a picture of IE Repin's *An Unexpected Visitor* to determine the ages of the people in the scene, they tended to concentrate their fixations on the people and particularly on the faces of those people. When viewers were instead attempting to estimate the material circumstances of the family in the scene, they distribute their fixations more widely over the scene. Yarbus observed similar systematicity in eye movements for other scenes and for other types of pictures such as faces and drawings of objects and suggested that the eyes tend to land on regions containing information that is either actually or in the viewer's opinion "useful or essential for perception."

The observation by Buswell (1935) and Yarbus (1967) that more informative scene regions receive more fixations has been replicated many times. In the first study to explore this relationship analytically, Mackworth & Morandi (1967) divided each of two color photographs into 64 square regions, and a first group of viewers rated the informativeness of each region based on how easy it would be to recognize on another occasion. A second group of viewers then examined the photographs with the task of deciding which of the two they preferred. *Fixation density* (the total number of discrete fixations in a given region over the course of scene viewing) in each of the 64 regions in each scene was found to be related to the rated informativeness of the region, with regions rated more informative receiving more fixations. In addition, viewers were as likely to fixate an informative region in the first two seconds of scene viewing as in other two second intervals, suggesting that region informativeness could be detected relatively early during scene viewing. Furthermore, regions that received low informativeness ratings were often not fixated at all, suggesting that uninformative regions could be rejected as potential fixation sites based on information acquired from the visual periphery.

The two pictures used by Mackworth & Morandi (1967) were visually and informationally simple: One picture depicted a pair of eyes within a hooded mask and the other a coastal map. In both, large regions were relatively uniform in their visual properties. Using scenes taken predominantly from the Thematic Apperception Test, Antes (1974) provided evidence that region in-

formativeness also affects fixation position in relatively complex scenes. Like Mackworth & Morandi (1967), Antes first asked a group of participants to rate the informativeness of scene regions. A separate group of viewers then examined the scenes while their eye movements were recorded. Antes found that the very first fixation position selected by a viewer (following the experimenter-induced initial fixation position at the center of the scene) was much more likely to be within an informative than an uninformative region of a scene, suggesting rapid control of fixation position by scene characteristics.

The studies reviewed thus far suggest that the positions of individual fixations in scenes, including initial fixations, are determined by the informativeness of specific scene regions. However, because informativeness was defined in these studies on the basis of experimenter intuition (Buswell 1935, Yarbus 1967) or ratings provided by other viewers (Antes 1974, Mackworth & Morandi 1967), and because a subjective assessment of informativeness may be based on either visual or semantic factors (or both), it is not possible to determine from these studies whether the eyes were controlled by perceptual factors, semantic factors, or both. If fixation position reflects ongoing cognitive operations as well as perceptual processes during scene viewing, then semantically informative regions should also be more likely to receive fixations than semantically uninformative regions, holding visual informativeness constant.

Loftus & Mackworth (1978) reported the first study designed to investigate directly the influence of semantic informativeness on fixation position while holding visual informativeness constant. Participants viewed line drawings of scenes in which a manipulated target object was either high or low in semantic informativeness. Semantic informativeness was defined as the degree to which a given object was predictable within the scene, with unpredictable objects taken to be more informative. An attempt was made to control visual informativeness by exchanging objects across scenes. For example, a farm scene and an underwater scene were paired so that either scene could contain an octopus or a tractor. Participants viewed the scenes for 4 s each in preparation for a later memory recognition test. Loftus & Mackworth reported three important results. First, fixation density was greater for semantically informative than uninformative regions, suggesting that fixation position was controlled by the semantic informativeness of a region with respect to the scene. This result accords with the qualitative data available in the figures of Buswell (1935) and Yarbus (1967). Second, viewers also tended to fixate the semantically inconsistent objects earlier than the consistent objects during the course of scene viewing, suggesting that the semantics of the extrafoveal region could control fixation placement. Third, viewers were more likely to fixate the semantically informative objects immediately following the first saccade within the scene. Because the average distance of the saccade to the target object was greater than 7º of visual angle, these data suggest that fixation sites could be selected

based on a semantic analysis of scene regions relatively distant in the visual periphery.

Two recent studies have called into question the conclusion that fixation placement is initially affected by a semantic analysis of scene regions that have only been viewed peripherally. First, De Graef et al (1990) manipulated semantic informativeness in a visual search task. Participants searched line drawings of scenes for nonobjects, objectlike figures that were meaningless. Using the same manipulation as Loftus & Mackworth (1978), prespecified meaningful target objects were placed in the scenes, and these objects were either semantically inconsistent (informative) or consistent (uninformative) with the rest of the scene. In contrast to Loftus & Mackworth, De Graef et al found no evidence that informative objects were initially fixated first or were fixated earlier than uninformative objects. In fact, viewers were equally likely to fixate the two types of objects for the first eight fixations in each scene. After the first eight fixations, viewers tended to fixate the uninformative objects sooner than the informative objects. These data thus contradict the finding that the eyes are immediately drawn to semantically informative objects in scenes and so call into question the conclusion that a semantic analysis of peripheral scene regions can control fixation placement.

Henderson et al (1999) reported two experiments designed to provide additional evidence concerning the influence of semantic informativeness on eye movements. The first used the Loftus & Mackworth (1978) methodology. Participants viewed line drawings of scenes under the same viewing instructions and with the same manipulation of semantic informativeness as used by Loftus & Mackworth. In contrast to Loftus & Mackworth but similar to De Graef et al (1990), Henderson et al (1999) found that viewers were no more likely to fixate initially the semantically informative target. Three specific results supported this conclusion. First, participants were equally likely to fixate the semantically informative and uninformative targets after the first (or second) saccade in the scene. Second, participants made the same average number of saccades in the scene prior to the initial fixation on the target object regardless of informativeness. Finally, the magnitude of the initial saccade to the target object was the same (about 3°) regardless of informativeness. These data suggest that the eyes are not initially driven by peripheral semantic analysis of individual objects.

In a second experiment, Henderson et al (1999) used a visual search task to further examine the relationship between semantic informativeness and initial fixation placement. Viewers were given the name of a target object at the beginning of each trial. A line drawing of a scene was then presented, and the participant's task was to determine as quickly as possible whether the target object was present in the scene. The instructions were designed to motivate the participants to find the targets as quickly as possible. If initial eye movements

are drawn to semantically informative objects in the periphery, informative objects should be found more quickly than uninformative objects. Instead, uninformative targets were fixated following fewer fixations (by about 0.5 fixations on average) than informative targets. Thus, there was no evidence that the eyes were drawn to semantically informative objects. Henderson et al (1999) suggested that the eyes reached the uninformative objects sooner because their positions were more spatially constrained by the scenes, not because local scene regions were analyzed for their meaning in the periphery. That is, information about the identity of the scene available during the initial fixation, in combination with a perceptual analysis of large-scale scene properties such as locations and orientations of surfaces, allowed participants to limit their search to likely target locations more easily when the target was semantically consistent with the scene (uninformative) than when it was inconsistent (informative) with the scene and so less spatially constrained.

Recent evidence presented by Mannan et al (1995) also suggests that initial fixation placements are controlled by perceptual features alone. In this study, eye movements were measured while viewers examined gray-scale photographs of real-world scenes that were presented for 3 s each. The photographs were high-pass filtered, low-pass filtered, or unfiltered. Fixation positions were found to be similar on the unfiltered and low-pass filtered scenes, particularly during the first 1.5 s of viewing. This result was found even when viewers were unable to describe the semantic content of the low-pass filtered scene. The direction of the initial saccade in a given scene was also similar for the low-pass and unfiltered versions. Mannan et al (1995) concluded that initial fixations are controlled by local visual rather than semantic features. In a subsequent analysis of these data, Mannan et al (1996) attempted to specify the visual features that determined initial fixation placement. They analyzed local regions of their scenes for seven spatial features: luminance maxima, luminance minima, image contrast, maxima of local positive physiological contrast, minima of local negative physiological contrast, edge density, and high spatial frequency. Only edge density predicted fixation position to any reliable degree, and even this feature produced only a relatively weak effect. Thus, the nature of the visual features that control fixation placement in scenes is still unclear.

Assuming that the Loftus & Mackworth result was not due to statistical error, there are at least two possible explanations for the inconsistency across studies. First, semantic informativeness and visual informativeness may have been correlated in the Loftus & Mackworth experiment (De Graef et al 1990, Rayner & Pollatsek 1992) so that effects that seemed to be due to semantic factors were actually due to visual factors. Second, the scenes used in later studies (De Graef et al 1990, Henderson et al 1999, Mannan et al 1995) may have been more visually complex than those used by Loftus & Mackworth (1978), so that

peripheral semantic analysis would be more difficult in the former cases. Supporting this view, Loftus & Mackworth (1978) observed an average saccadic amplitude of more than 7° in their study, roughly twice the amplitude of the average saccade observed over a large range of scene-viewing experiments (Henderson & Hollingworth 1998). Taken together, then, the data suggest that initial fixations in a scene are controlled by visual rather than semantic features of local regions.

While semantic informativeness does not appear to influence initial fixation placement, qualitative analysis of the figures presented by Buswell (1935) and Yarbus (1967) suggests that it does influence overall fixation density in a scene region. Loftus & Mackworth (1978) also observed that fixation density was greater for semantically informative regions. Similarly, Henderson et al (1999) found that both the number of fixations viewers made in a region when that region was first fixated, and the number of fixations due to looks back to a region from other regions of the scene, were greater for semantically informative objects. In contrast to these results, Friedman (1979, presented in Friedman & Liebelt 1981) found no effect of semantic informativeness on fixation density. In this study, line drawings of scenes containing objects that had been rated for their a priori likelihood in the scene were presented to viewers who examined them in preparation for a difficult recognition memory test. Fixation density was not found to be correlated with rated likelihood. An explanation for the difference in results across studies rests on the strength of the informativeness manipulation. In Loftus & Mackworth (1978) and Henderson et al (1999), semantically informative regions contained semantically anomalous objects (e.g. a microscope in a bar), whereas in Friedman (1979), the manipulation of informativeness was relatively weak, with objects ranging continuously from very likely to somewhat likely in the scenes. Thus, the effect of semantic informativeness on fixation density was probably easier to detect in the former studies.

Together, the available data suggest that fixation placement in a scene is initially based on a combination of visual characteristics of local scene regions, knowledge of the scene category, and global visual properties (large-scale visual features) of the scene. Fixation placement does not seem to depend initially on semantic analysis of peripheral scene regions. However, once a region has been fixated so that semantic analysis is possible based on foveal vision, immediate refixations within the region and later returns to that region can then be based on the semantic informativeness of the region. The extent to which a region is semantically informative is dependent on the viewer's task as well as the nature of the region, leading to changes in fixation density as a function of task. While this basic framework accounts for the majority of available evidence, a large number of questions are yet to be answered. For example, it is not clear what visual features are used to select fixation sites, how spe-

cific sites are weighted during selection, what the selection mechanism is, and how visual and semantic factors trade off over time in controlling fixation placement. It is also not clear how visual features in the scene and cognitive factors related to the goals of the viewer interact in determining fixation sites. These issues are not trivial; while there is some similarity in initial fixation placement across individuals viewing the same scene, this similarity drops rapidly as scene perception unfolds (Mannan et al 1995). Furthermore, the eyes very rarely fixate the same positions in the same order; very few two-fixation sequences are the same across individuals or even within the same individual viewing the same scene a second time (Mannan et al 1997).

Fixation Time During Scene Perception

The *total time* a viewer fixates a given scene region (the sum of the durations of all fixations in a region) varies for different regions in a scene (Buswell 1935, Henderson et al 1999). This finding is not surprising, given that the total time that a region is fixated is correlated with fixation density in that region, and, as discussed above, fixation density tends to be higher for visually and semantically informative regions. At a more fine-grained level of analysis, we can ask whether the durations of individual fixations and temporally contiguous clusters of fixations are also affected by the perceptual and semantic characteristics of particular scene regions. The average fixation duration during scene viewing is about 330 ms, with a significant amount of variability around this mean. Fixation durations range from less than 50 to more than 1000 ms in a skewed distribution with a mode of about 230 ms (Henderson & Hollingworth 1998). The question is whether ongoing perceptual and semantic processing accounts for any of this variability.

There is currently some direct evidence that the visual information available in a fixation affects the duration of that fixation. In the study described above, Mannan et al (1995) found that fixation durations were longest during viewing of low-pass filtered scenes, intermediate for high-pass filtered scenes, and shortest for the unfiltered versions, suggesting that individual fixation durations are affected by the nature of the visual information available in the scene. In this study, however, it was not possible to determine if fixation durations were affected by the nature of the visual information available at fixation, the visual information available in the periphery, or both. To separate these possibilities, van Diepen and colleagues (1998) used a moving mask paradigm and directly manipulated the quality of the visual information available at fixation independently of that available beyond fixation. In this paradigm, a mask or other type of visual degradation can be made to move across the scene in spatial and temporal synchrony with the current fixation position (van Diepen et al 1998). Viewers searched for nonobjects in line drawings of scenes, and

the image at fixation was presented normally or was degraded by overlaying a noise mask or by decreasing contrast at the fixated region (van Diepen et al 1995, 1998). When the image was degraded beginning at the onset of a fixation, *first fixation duration* (the duration of the initial fixation in a particular scene region) was longer than in a control condition, suggesting that the duration of the initial fixation was controlled, at least in part, by the acquisition of visual information from the fixated region. This result is similar to that observed when an artificial foveal scotoma is introduced via the moving mask technique during visual analysis of pictures of individual objects (Henderson et al 1997). These studies show that fixation duration is sensitive to the quality of the visual information available during that fixation. However, because stimulus manipulations such as filtering and masking affect both the visual characteristics of the image and the viewer's ability to semantically interpret that image, it is possible that difficulties of semantic analysis rather than visual analysis lead to the longer fixation durations. Contriving manipulations of visual but not semantic characteristics of a given region is a problem that will be difficult to solve with meaningful scene stimuli.

The effect of semantic informativeness on fine-grained measures of fixation time during scene viewing has also been studied. Loftus & Mackworth (1978) found that *first pass gaze duration* (the sum of all fixations from first entry to first exit in a region) was longer for semantically informative objects. Friedman (1979) similarly showed that first pass gaze duration was longer for objects that were less likely to be found in a particular scene. (Loftus & Mackworth and Friedman used the term *duration of the first fixation* to refer to first pass gaze duration.) Using the nonobject counting task, De Graef et al (1990) found that first pass gaze durations were longer for semantically informative objects, though this difference appeared only in the later stages of scene viewing. De Graef et al also found that whereas overall first fixation durations did not differ as a function of the semantic informativeness of the fixated region, first fixation durations on regions that were initially encountered late during scene exploration (following the median number of total fixations) were shorter on semantically uninformative objects. Finally, Henderson et al (1999) found that first pass gaze duration and *second pass gaze duration* (the sum of all fixations from second entry to exit in a region) were longer for semantically informative than uninformative objects. Together, these results show a clear effect of the meaning of a scene region on gaze duration in that region but a less clear effect on first fixation duration.

Conclusions

The results of eye movement studies during scene viewing show that fixation positions are nonrandom, with fixations clustering on both visually and semantically informative regions. The placement of the first few fixations in a

scene seems to be controlled by the visual features in the scene and global semantic characteristics of the scene (e.g. the scene concept) but not by semantic characteristics of local scene regions. As viewing progresses and local regions are fixated and semantically analyzed, positions of later fixations come to be controlled by both the visual and semantic properties of those local regions. The length of time the eyes remain in a given region is immediately affected by both the visual and semantic properties of that region. Thus, although the eyes are not initially drawn to a region based on its meaning, they may remain longer in that region upon first encountering it if it is more semantically informative.

Although there is reasonable consistency in the results of the reviewed studies, there are also some notable discrepancies. It is often difficult to determine the cause of these differences because a number of potentially important factors vary from study to study, including image size, viewing time per scene, image content, and image type (Henderson & Hollingworth 1998). Each factor could produce an independent effect and could also interact with the others in complex ways to influence eye movements. Further investigation of these issues is required before eye movement control in high-level scene perception will be completely understood. Also, another potentially important factor that might exert strong effects on eye movement patterns is the viewing task. Very little systematic work has been conducted to examine the degree to which viewing patterns change as a function of task, but to the extent that eye movement patterns are driven by the goals of the cognitive system (Ballard 1991, Land & Furneaux 1997, Rayner 1978, Yarbus 1967), this will be a critical factor to examine in future studies.

SCENE MEMORY ACROSS SACCADES

In this section, we explore the nature of the representation that is generated across saccades as we view a scene over an extended period of time. Phenomenologically, the visual system seems to construct a complete, veridical perceptual representation of the environment, akin to a high-resolution, full-color photograph. Such a representation could not be based on the information contained in any given fixation, however, because of the rapid drop-off from the current fixation point in both acuity (Anstis 1974, Riggs 1965) and color sensitivity (Mullen 1990). Thus, if our phenomenology reflects reality, the visual system must build up a composite perceptual image over consecutive fixations. Historically, this composite image hypothesis has been instantiated by models in which a perceptual image is generated during each fixation and stored in the brain, with images from consecutive fixations overlapped or spatially aligned in a system that maps a retinal reference frame onto a spatiotopic reference frame (e.g. Brietmeyer et al 1982, Davidson et al 1973, Duhamel et

al 1992, Feldman 1985, Jonides et al 1982, McConkie & Rayner 1975, Pouget et al 1993). In composite image models, the perceptual image formed during two consecutive fixations could be aligned by tracking the extent of the saccade and/or by comparing the similarity of the images themselves.

Although many different models of transsaccadic visual perception based on this basic scheme have been proposed, psychophysical and behavioral data have almost uniformly provided evidence against them (see reviews by Irwin 1992, 1996; O'Regan 1992b; Pollatsek & Rayner 1992). For example, when two dot patterns forming a matrix of dots are presented in rapid succession at the same spatial position within a fixation, a single fused pattern is perceived and performance (e.g. identification of a missing dot from the matrix) can be based upon this percept (Di Lollo 1977, Eriksen & Collins 1967, Irwin 1991). However, when the two patterns are viewed in rapid succession at the same spatial position across a saccade, no such fused percept is experienced and performance is dramatically reduced (Bridgeman & Mayer 1983; Irwin 1991, Irwin et al 1983, 1990; Rayner & Pollatsek 1983; see also O'Regan & Levy-Schoen 1983). Similarly, spatial displacement of a visual stimulus is very difficult to detect when the displacement takes place during a saccade (Bridgeman et al 1975, Henderson 1997, McConkie & Currie 1996). If internal images were being spatially aligned to form a composite image (based, for example, on the distance of the saccade), spatial displacement should be very obvious to the viewer. Other types of image changes, such as enlargements or reductions in object size and changes to object contours, often go unnoticed when they take place during a saccade (Henderson 1997, Henderson et al 1987). Again, if a composite image were being generated via spatial alignment and image overlap, then these kinds of changes should be quite noticeable.

Change Blindness Across Saccades During Scene Viewing

The studies reviewed above strongly suggest that the visual system does not (and, in fact, cannot) retain a detailed perceptual image of the visual input across saccades. Recent research on scene perception lends additional support to this conclusion and further suggests that even the amount of conceptual information that is carried across a saccade is limited. This conclusion comes from a strikingly counterintuitive result in recent scene perception research: Viewers often fail to notice large and seemingly salient changes to scene regions and objects when those changes take place during a saccade (Grimes 1996, McConkie 1990, McConkie & Currie 1996). In a striking demonstration of this effect, Grimes and McConkie (see Grimes 1996) presented viewers with full-color pictures of scenes over extended viewing times. The participants were instructed to view the scenes in preparation for a relatively difficult memory test and were further told that something in a scene would occasion-

ally change and that they should press a button if and when that happened. Participants' eye movements were monitored, and occasionally one region of a scene was changed during the *n*th saccade, where *n* was predetermined. The striking result was that viewers often failed to detect what would seem to be very obvious perceptual and conceptual changes in the scene. For example, 100% of the viewers failed to detect a 25% increase in the size of a prominent building in a city skyline, 100% failed to detect that the hats on the heads of two men who were central in a scene switched one to the other, and 50% failed to notice when the heads were exchanged between two cowboys sitting on a bench (Grimes 1996). Even assuming that only a relatively detailed conceptual representation of a scene (in contrast to a complete perceptual representation) is retained across saccades, these changes should be noticed with relatively high frequency. Thus, these results call into question the idea that a detailed scene representation is carried across saccades in the service of constructing a composite perceptual image.

The study reported by Grimes is important because the results have broad implications for our understanding of perception, cognition, and the nature of consciousness (Dennett 1991). However, it is important to note that the Grimes (1996) report was anecdotal, providing few specific details about the experiment. For example, participants were freely moving their eyes around the scene during the experiment, and the change occurred during a prespecified saccade (i.e. the *n*th saccade) without respect for the position of the fixation prior to or following that saccade. Thus, it is not known whether the change detection performance was related to fixation position in the scene. This factor could be critical, given the evidence reviewed above that semantic analysis of local regions is at least initially constrained to areas of the scene at or near fixation. Thus, it will be important to replicate these results with fixation position controlled.

Change Blindness and Simulated Saccades

In an attempt to determine whether the change blindness phenomenon is a consequence of the execution of a saccade, Rensink et al (1997) introduced a change detection paradigm in which scene changes were decoupled from saccades. A photograph of a scene (A) was presented for 240 ms, followed by a gray field for 80 ms, followed by a changed version of the initial scene (A′), and so on alternating between A, the gray field, and A′. (In some experiments, each version of the scene was repeated before the change, e.g. A, A, A′, A′, to prevent participants from predicting when a change would happen.) The participant was asked to press a button when the change was detected and then to state what the change was. The result was that scene changes were very difficult to detect in this *flicker paradigm*, often requiring tens of seconds of viewing time. Interestingly, once a change had been detected by an observer, it be-

came obvious thereafter. Rensink et al (1997) suggested that when local motion signals are removed from the visual signal (via the intervening gray field), the detection of what would otherwise be highly salient changes becomes extraordinarily difficult, at least until attention is directed to the changing region and perceptual information is explicitly encoded and compared across images.

Because the scene changes in the Rensink et al (1997) study were not synchronized to the viewer's saccades, the researchers concluded that the change blindness effect reported by Grimes (1996) is not tied to the saccadic system. However, given that participants were allowed to move their eyes as they searched for the changing object in the Rensink et al (1997) study, it is possible that a fortuitous relationship between viewers' saccades and the scene changes might still have accounted for their effect. To test this hypothesis, A Hollingworth & JM Henderson (submitted) modified the flicker paradigm so that the first scene image was displayed briefly and one time only, followed by an intervening gray field, followed by a comparison image of the same scene with or without a change to an object in the scene. Because the initial view of the scene was presented only briefly, there was no time for the viewer to execute a saccade. Although better than in the flicker paradigm, change detection performance in this task was still poor. This result suggests that when local motion signals are removed from the input, changes in a scene are difficult to detect, regardless of whether they take place across a saccade or within a fixation. Additional support for this hypothesis was provided by O'Regan et al (1996), who used multiple gray patches (similar to mud splattering on a windshield) presented on a scene simultaneously with the scene change. Although the splatter never covered the changing region, changes were difficult to detect in the splatter condition compared with a control condition without splatter. Similarly, Levin & Simons (1997) showed that visual changes to objects in an ongoing film are difficult to detect across a film cut, where different viewing angles are used before and after the cut. As in the "splatter" condition, a film cut introduces discontinuities across much of the visual field. Together, these results suggest that when local motion signals are eliminated as a result of an intervening blank period (caused by a saccade or a uniform gray field inserted within a fixation), or overwhelmed because of additional motion signals across the visual field (e.g. a splatter or film cut), change blindness results. Thus, change blindness appears to reflect a general and fundamental characteristic of the way in which information is acquired, represented, and retained from a dynamically viewed scene.

The change blindness effect suggests that little of the information that is latent in the retinal image during a fixation is encoded into an enduring form that can be retained across a saccade or other intervening temporal gap. Thus, it becomes important to understand the processes that control the selection of the information to be encoded into a more enduring form. Rensink et al (1997)

proposed that a limited-capacity attentional mechanism must select perceptual information from an iconic store during a fixation and transfer it to a more stable and longer-lasting visual short-term memory (VSTM) representation if it is to be retained. In this hypothesis, scene regions that are more likely to be attended during scene viewing should be more likely to be encoded and stored in a stable format. Supporting this hypothesis, Rensink et al (1997) demonstrated that change detection was facilitated for scene regions that were rated more interesting by a group of viewers who independently judged scene regions in isolation. However, this method is problematic because "interest" was not directly manipulated (see discussion of informativeness ratings in the eye movement section above). Thus, because the interesting and uninteresting regions of the scenes may have differed along many physical dimensions, it is difficult to attribute the change detection differences to interest alone.

In a study designed to direct attention to specific scene regions in a more principled manner, A Hollingworth & JM Henderson (submitted) used semantic consistency to manipulate the semantic informativeness of a scene region. Target objects that were semantically constrained in pairs of scenes were exchanged across scenes to produce images in which a given object was either semantically consistent (e.g. a mixer in a kitchen) or semantically inconsistent (e.g. a live chicken in a kitchen) with the rest of the scene, as described in the eye movement section above. These stimuli were then employed both in the Rensink et al (1997) change detection paradigm and in the simpler version of the paradigm in which a scene was presented only twice rather than alternating back and forth. In both paradigms, the main result was that change detection was better when the changing object was semantically informative. On the assumption that semantic informativeness holds attention (Friedman 1979, Henderson et al 1999, Loftus & Mackworth 1978), these data support the Rensink et al (1997) hypothesis that attention is needed to transfer information to a stable medium (e.g. VSTM; Potter 1976) if that information is to be available to support the detection of changes.

A third set of data supporting the hypothesis that covert attention plays a critical role in the encoding of information from a scene was provided by CB Currie & GW McConkie (submitted), who demonstrated that spatial displacements to objects in scenes that take place during a saccade are much more noticeable when the displaced object is the target of the saccade than when it occupies a position elsewhere in the scene. Given the behavioral and neurophysiological evidence that covert visual-spatial attention tends to be allocated to a saccade target prior to the execution of the saccade (e.g. Deubel & Schneider 1996, Henderson 1992a, Henderson et al 1989; reviewed by Henderson 1996), these data can also be taken to support the view that saccade targets are attended and so are more likely to be retained in memory than other objects in a scene.

Conclusions

The literature reviewed in this section strongly suggests that only a limited amount of information is carried across saccades during complex, natural scene viewing and that this information is coded and stored in a relatively abstract (nonperceptual) format. What, then, accounts for our experience of a complete and integrated visual world? Current evidence suggests that this experience is an illusion or construction based on an abstract conceptual representation coding general information about the scene (e.g. its category) combined with perceptual information derived from the current fixation (e.g. O'Regan 1992b, Grimes 1996; see also Churchland et al 1994, Dennett 1991, but see Deubel et al 1996, for an alternative view).

SCENE CONTEXT AND OBJECT IDENTIFICATION

In this section we review the literature on object identification in scenes. The central question is whether the context established by a scene influences the identification of objects in that scene. In other words, does object identification operate exclusively on bottom-up visual information, as proposed by current theories of object recognition (e.g. Biederman 1987, Bülthoff et al 1995)? Or is object identification sensitive to the meaning of the scene in which an object appears, as proposed by theories of object identification in scenes (e.g. Biederman et al 1982, Friedman 1979, Kosslyn 1994)? First, we review research on scene identification. Second, we review models of the relationship between scene knowledge and object identification. Third, we review the empirical evidence mediating between these models.

Scene Identification

Scene identification research has focused primarily on two issues: (*a*) the time-course of scene identification and (*b*) the types of information used to identify a scene as a particular scene type. Potter (Potter 1975, 1976; Potter & Levy 1969) conducted a series of studies to investigate the time-course of scene identification and memory encoding. These studies presented a series of photographs of scenes in rapid succession. When a verbal description of a target scene was provided prior to presentation of the series, participants were able to detect the target scene quite reliably, even at a presentation rate of 113 ms per scene. Potter (1976) concluded that a scene can be identified in approximately 100 ms. One concern with these studies is that the scene descriptions did not specify the global identity of the scene but instead described individual objects in the scene (e.g. *a baby reaching for a butterfly*). Thus, detection performance may have been based on the identification of individual objects rather than on identification of the scene as a whole. Schyns & Oliva (1994, Oliva & Schyns 1997) have demonstrated that a photograph of a scene can be identified as a

particular scene type (e.g. highway or living room) from a masked presentation as short as 45–135 ms. This result demonstrates that the information necessary to identify a scene can be extracted quickly, but it does not indicate the precise amount of time required to achieve identification. Future research will be needed to characterize the time-course of scene identification. In particular, the comparative speed of scene versus object identification is important for theories that propose interactions between scene context and the identification of constituent objects.

A second area of research has investigated the scene information used for identification. First, scene identity could be inferred from the identification of one or more key objects (Friedman 1979) and, perhaps, their spatial relations (De Graef et al 1990). Second, a scene could be identified from scene-level information independent of the identities of individual objects (Biederman 1981, 1988; Schyns & Oliva 1994). Most research has supported the latter idea that early scene processing is based on global scene information rather than local object information (Antes et al 1981, Loftus et al 1983, Metzger & Antes 1983, Schyns & Oliva 1994). Schyns & Oliva (1994) demonstrated that scenes can be identified from low-spatial-frequency images that preserve the spatial relations between large-scale structures in the scene but which lack the visual detail needed to identify local objects. In addition, when identifying a scene from a very brief view (50 ms), participants tend to base their interpretation on low-frequency information rather than on high-frequency information (Schyns & Oliva 1994), though this global-to-local bias does not appear to be a hard constraint (Oliva & Schyns 1997).

A related issue concerns the internal representations functional in scene identification. Biederman (1981, 1988) proposed that an arrangement of volumetric primitives (geons), each representing a prominent object in the scene, may allow rapid scene identification independently of local object identification. According to this view, scenes employ the same representational vocabulary as objects, except on a larger spatial scale. This proposal has not been tested empirically; however, there are a number of reasons to think that scenes may not be represented as large objects. Whereas an object tends to have a highly constrained set of component parts and relations between parts, a scene places far less constraint on objects and spatial relations between objects (Henderson 1992b, Hollingworth & Henderson 1998). Evidence from neuropsychology suggests that within- and between-object spatial relations may be represented differently (Humphreys & Riddoch 1994, 1995). In addition, neural imaging results suggest that there may be separate cortical areas supporting object and scene identification (Epstein & Kanwisher 1998). Future research will need to identify more precisely the internal representations constructed from a scene and the processes by which these representations are compared to stored scene representations.

Models of Object Identification in Scenes

Object identification can be assumed to consist of the following component processes. First, the retinal image is translated into a set of visual primitives (e.g. surfaces and edges). Second, these primitives are used to construct structural descriptions of the object tokens in the scene. Third, these constructed descriptions are matched to stored long-term memory descriptions. When a match is found, identification has occurred, and semantic information stored in memory about that object type becomes available. In this view of object identification, the first two stages can be considered perceptual in that the task is to translate retinal stimulation into a structural description that is compatible with stored memory representations. The matching stage, however, can be seen as an interface between perception and cognition, in which perceptual information must make contact with memory representations. Models of object identification in scenes can be divided into three groups based on the stage of object identification at which scene context is proposed to exert an influence. One group of theories proposes that expectations derived from scene knowledge interact with the perceptual analysis of object tokens (i.e. the first two stages of object identification). A second group proposes that the locus of interaction is at the matching stage, when perceptual descriptions are matched to long-term memory representations. A third group proposes that object identification (including the matching stage) is isolated from scene knowledge.

The *perceptual schema model* proposes that expectations derived from knowledge about the composition of a scene type interact with the perceptual analysis of object tokens in that scene (Biederman 1981; Biederman et al 1982, 1983; Boyce et al 1989; Metzger & Antes 1983; Palmer 1975b). According to this view, the memory representation of a scene type (a *schema* or *frame*) contains information about the objects and spatial relations between objects that form that type. The early activation of a scene schema facilitates the subsequent perceptual analysis of schema-consistent objects and, perhaps, inhibits the perceptual analysis of schema-inconsistent objects (Biederman et al 1982). The mechanisms by which schema activation facilitates the perceptual analysis of consistent objects have not been specified in detail. Some researchers (Boyce et al 1989, Metzger & Antes 1983) have suggested that perceptual facilitation could be explained within an interactive activation model, in which partial activation of nodes at the scene level constrains perceptual analysis at the object level. The perceptual schema model predicts that the identification of objects consistent with a scene will be facilitated compared to inconsistent objects. In addition, the constructed description of a consistent object should be more elaborated than that of an inconsistent object.

At the level of the architecture of the visual system, the perceptual schema model assumes that there is no clear distinction between perceptual processing

and cognitive processing. It draws from New Look theories of perception, which propose that cognitively derived hypotheses modulate the encoding of perceptual information (Bruner 1957, 1973; Neisser 1967). In addition, it is consistent with current theories proposing that vision is a constraint-satisfaction problem, in which all available constraints are consulted when interpreting an input pattern (Mumford 1994, Rumelhart et al 1986).

The *priming model* proposes that the locus of the contextual effect is at the stage when a structural description of an object token is matched against long-term memory representations (Bar & Ullman 1996, Friedman 1979, Friedman & Liebelt 1981, Kosslyn 1994, Palmer 1975a, Ullman 1996). According to the priming model, the activation of a scene schema primes the stored representations of schema-consistent object types. This priming can be viewed as a modulation of the criterion amount of perceptual information necessary to select a particular object representation as a match. Relatively less perceptual information will need to be encoded to select a primed object representation compared with an unprimed object representation (Friedman 1979). Similar to the perceptual schema model, the priming model proposes that identification of objects consistent with a scene will be facilitated compared with inconsistent objects. However, the priming model differs from the perceptual schema model because it proposes that scene knowledge influences only the criterion used to determine that a particular object type is present, without directly influencing the perceptual analysis of the object token.

The *functional isolation model* proposes that object identification is isolated from expectations derived from scene knowledge (Hollingworth & Henderson 1998). This model is consistent with current theories of object identification (Biederman 1987, Bülthoff et al 1995; see also Marr & Nishihara 1978) that propose that bottom-up visual analysis is sufficient to discriminate between entry-level object categories. This model is also consistent with theories proposing an architectural division between perceptual processing and cognitive processing (Fodor 1983; Pylyshyn 1980, 1998). The functional isolation model predicts that experiments examining the perceptual analysis of objects should find no effect of the relation between object and scene. However, context effects may arise in experiments that are sensitive to later influences of scene constraint.

Before turning to the literature on object identification in scenes, it is important to establish the boundary conditions under which scene context could plausibly interact with object perception. First, a scene must be identified early enough to influence the identification of constituent objects. As reviewed above, the information necessary to identify a scene can be extracted quite quickly, possibly from an analysis of global rather than local scene features. Second, scenes must place significant constraints on the objects that can appear in them, and stored knowledge about scene types must include these con-

straints. Supporting this assumption, participants are quite reliable in their judgments about what objects are consistent versus inconsistent with a particular scene (e.g. Friedman 1979, Henderson et al 1999) and exhibit strong response biases as a function of the consistency between object and scene (e.g. Biederman et al 1982, Hollingworth & Henderson 1998, Palmer 1975a). Thus, there seems adequate evidence to suppose that if the architecture of the visual system allows interactions between scene knowledge and object identification, scene-contextual constraint is available early enough and is robust enough to influence the identification of objects.

Studies of Object Identification in Scenes

In this section we review the experimental evidence mediating between these models. The principal difficulty in this literature has been to determine the representational level at which prior scene knowledge interacts with the processing of objects. As an illustrative example, consider a study by Palmer (1975a). Palmer presented a line drawing of a scene for 2 s followed by a brief presentation of an isolated target object that was either semantically consistent with that scene (i.e. likely to appear in the scene) or semantically inconsistent (i.e. unlikely to appear in that scene). In addition, semantically inconsistent target objects could be shaped similarly to the consistent target or not. Palmer found that consistent objects were named more accurately than inconsistent objects and that inconsistent objects shaped similarly to a consistent target were named least accurately. Although this result has been cited as evidence for the influence of scene knowledge on object identification, the effect could arise at a number of different stages of analysis. First, consistent scene context could facilitate the perceptual analysis of consistent objects, as proposed by the perceptual schema model. Second, it could reduce the criterion amount of information needed to reach an identification threshold, as proposed by the priming model and by Palmer. Third, scene context could influence postidentification processing, such as response generation or educated guessing, consistent with the functional isolation model.

Designing experimental paradigms to discriminate between these possibilities has proven difficult. In the remainder of this section, we review experiments that have sought to investigate whether consistent scene context facilitates the identification of objects, with particular focus on the extent to which each experiment is able to discriminate between the models reviewed above. The principal manipulation of object consistency in these studies has been the likelihood of an object appearing in a scene (i.e. the semantic consistency between object and scene), though some studies have manipulated other types of scene relations, including an object's spatial position and size (e.g. Biederman et al 1982, De Graef et al 1990). For further discussion of this literature, see

Boyce & Pollatsek (1992a), De Graef (1992), Henderson (1992b), and Rayner & Pollatsek (1992).

EYE MOVEMENT PARADIGMS In eye movement paradigms, the duration of the fixation(s) on a target object has been taken as a measure of the speed of object identification. Friedman (1979; see eye movement section for detailed discussion of this experiment) found that first pass gaze duration was shorter for semantically consistent versus inconsistent target objects and interpreted the difference in gaze duration as support for the priming model. This interpretation has been questioned, however, because it is unlikely that the difference in gaze duration (more than 300 ms) was due to identification processes alone (Biederman et al 1982, Henderson 1992b, Rayner & Pollatsek 1992). First, the difference may have been caused by the difficulty of integrating an already identified object into a conceptual representation in which it was incongruous (Henderson 1992b). Second, the instructions to prepare for a difficult memory test may have caused participants to dwell longer on objects that were difficult to encode into memory (Hollingworth & Henderson 1998). Third, once identified, inconsistent objects are likely to be more interesting to participants than consistent objects, leading to the longer gaze durations (Biederman et al 1982).

De Graef et al (1990) found shorter first fixation durations on semantically consistent versus inconsistent objects, but this effect arose only when the target object was initially encountered relatively late in scene viewing. The absence of a context effect early in viewing is consistent with the functional isolation model. The context effect obtained later in scene viewing is more difficult to reconcile with this view. However, it is not at all clear why a context effect would develop only late during viewing. One possibility is that participants initially ignored the larger scene, registering scene meaning only after the accumulation of enough local information (Boyce & Pollatsek 1992a, De Graef et al 1990, Rayner & Pollatsek 1992). This explanation, however, runs counter to strong evidence that scenes are identified within the first fixation on the scene and that identification occurs even when such processing is not necessary to perform the task (e.g. Biederman et al 1982, Boyce & Pollatsek 1992b, Hollingworth & Henderson 1998). A more general problem with drawing strong conclusions from this study is that we have no direct evidence to indicate whether first fixation duration reflects object identification alone or later processing as well (Henderson 1992b, Rayner & Pollatsek 1992). Until we know more about the types of object processing reflected in different fixation duration measures, results from eye movement paradigms are unlikely to be able to resolve the question of whether scene context influences the identification of objects.

Boyce & Pollatsek (1992b) developed a variant of the eye movement paradigm in which the naming latency for a fixated object was used as a measure of

object identification performance. In this study, the participant first fixated the center of the screen. A line drawing of a scene then appeared, and 75 ms later, a target object wiggled (shifted about half a degree and then shifted back 50 ms later). The participant's task was to make an eye movement to the wiggled object and, upon completion of the eye movement, to name the object as quickly as possible. Boyce & Pollatsek found that naming latency was shorter for semantically consistent versus inconsistent target objects. As with fixation duration measures, however, we do not know whether differences in naming latency reflect the influence of scene context on object identification or on postidentification processing as well.

OBJECT DETECTION PARADIGMS In object detection paradigms, the accuracy of detecting a target object in a briefly presented scene has been taken as a measure of object identification performance. Biederman (Biederman 1972; Biederman et al 1973, 1974) sought to assess the influence of coherent scene context on object identification by measuring detection performance for target objects presented in normal versus jumbled scenes. The normal images were photographs of common environments, and the jumbled images were created by cutting the photographs into six rectangles and rearranging them (though the rectangle containing the target object remained in its original position). Scenes were presented briefly (20–700 ms) followed by a mask and a cue marking the position where the target object had appeared. Participants more accurately discriminated the target object from distractors when the scene was normal versus jumbled. Similar results were found in a search paradigm (Biederman et al 1973); participants took less time to find the target object when it appeared in a normal versus in a jumbled scene. These results have been widely cited as support for the perceptual schema model. However, this paradigm has been criticized because the jumbling manipulation introduced new contours to the jumbled scenes and thus did not control the visual complexity of the normal versus jumbled images (Bar & Ullman 1996, Henderson 1992b). In addition, the normal scene advantage may not have reflected differences in the perceptual analysis of objects. Compared to the jumbled condition, participants may have more successfully encoded the spatial relation between the cued region and the rest of the scene when the scene was normal. They could then choose the test object that was likely to have appeared in that position (Biederman 1972).

More recent object detection experiments have tested detection performance for consistent versus inconsistent objects presented in the same scene context (Biederman et al 1982, 1983; Boyce et al 1989; Hollingworth & Henderson 1998; Masson 1991). These experiments employed signal detection measures to discriminate contextual influence at the level of perceptual analysis from influence at later levels of analysis. The logic behind signal de-

tection methodology is that effects of context on perceptual processing will be reflected in measures of sensitivity, whereas later effects of context (e.g. at the matching stage or at postidentification stages) will be reflected in measures of bias (but see Norris 1995).

Biederman et al (1982) asked participants to decide whether a target object had appeared within a briefly presented scene at a particular location. During each trial, a label naming a target object was presented until the participant was ready to continue, followed by a line drawing of a scene for 150 ms, followed by a pattern mask with an embedded location cue. Participants indicated whether the target had appeared in the scene at the cued location. The object appearing at the cued location either could be consistent with the scene or could violate scene expectations along one or more dimensions, including probability (semantic consistency), position, size, support, and interposition (whether the object occluded objects behind it or was transparent). Biederman et al found that detection sensitivity (d') was best when the cued object did not violate any of the constraints imposed by scene meaning. Performance was poorer across all violation dimensions, with compound violations (e.g. semantically inconsistent and unsupported) producing even greater performance decrements. Biederman et al (1982, Biederman 1981) interpreted these results as supporting a perceptual schema model. They argued that because semantic violations were no less disruptive than structural violations, the locus of semantic contextual influence must be during the perceptual analysis of object tokens (but see Henderson 1992b, De Graef et al 1990).

Boyce et al (1989) explored whether the detection advantage observed for semantically consistent versus inconsistent objects was due to the global meaning of the scene or to the presence of other semantically related objects within the scene, as had been suggested by Henderson et al (1987). Boyce et al manipulated the consistency of the cued object with both the global scene and with other cohort objects appearing in the scene. For example, a doll could appear in a bedroom with other bedroom objects, in a bedroom with objects more likely to be found in a refrigerator, in a refrigerator scene with other bedroom objects, or in a refrigerator with other refrigerator objects. Detection sensitivity was facilitated when the cued object was semantically consistent with the global scene in which it appeared compared with when it was inconsistent with the global scene. In contrast, there was no effect of the consistency of the cued object with the cohort objects in the scene. Boyce et al concluded that the global meaning of the scene, rather than the specific objects present in the scene, is functional in facilitating object identification.

The results of Biederman et al (1982) and Boyce et al (1989) provide the strongest evidence to date that consistent scene context facilitates object identification and provide the core support for the perceptual schema model. However, a number of methodological concerns have been raised regarding these

paradigms (De Graef et al 1990, De Graef 1992, Henderson 1992b, Hollingworth & Henderson 1998). First, there is reason to believe that the signal detection methodology did not adequately eliminate response bias from sensitivity measures. These object detection studies did not compute sensitivity using the correct detection of a particular signal when it was present and the false detection of the same signal when it was absent, as required by signal detection theory. Catch trials presented the same scene (and cued object) as in target-present trials but merely changed the label appearing before the scene. In addition, the Biederman et al studies (1982, 1983) did not control the semantic consistency between the target label and the scene on catch trials: False alarms were computed in both consistent and inconsistent cued object conditions by averaging across catch trials on which the target label was semantically consistent and semantically inconsistent with the scene. Hollingworth & Henderson (1998) replicated the Biederman et al (1982) study first using the original signal detection design and then using a corrected design in which participants attempted to detect the same object on corresponding target-present and catch trials. The experiment using the original design replicated the consistent object advantage found by Biederman et al and Boyce et al (1989). However, the experiment using the corrected design showed no advantage for the detection of semantically consistent versus semantically inconsistent objects. These results suggest that the consistent object advantage in previous object detection experiments likely arose from the inadequate control of response bias and not from the influence of scene context on the perceptual analysis of objects.

The second concern with previous object detection paradigms (Biederman et al 1982, 1983; Boyce et al 1989) is that participants may have searched areas of the scene where the target object was likely to be found. If the spatial positions of semantically consistent objects were more predictable than those of inconsistent objects, detection of the former would have been facilitated compared to the latter, even if there were no differences in the perceptibility of each type of object (Hollingworth & Henderson 1998). Supporting this idea, Henderson et al (1997) demonstrated that semantically consistent objects are indeed easier to locate in scenes than inconsistent objects (as described in the above section on eye movements in scenes). A similar advantage may have been afforded to consistent objects in object detection paradigms, leading to an apparent advantage for the perceptual processing of these objects. Hollingworth & Henderson (1998) tested whether differences in search efficiency influence performance in the object detection paradigm. They presented the target object label after the scene so that participants could not form a search strategy. Contrary to earlier studies, Hollingworth & Henderson found a reliable advantage for the detection of semantically inconsistent objects (see discussion in above section on change detection, and Hollingworth & Henderson 1998).

To investigate the identification of objects in scenes independently of response bias, Hollingworth & Henderson (1998) introduced a post-scene, forced-choice discrimination procedure. This procedure is a variant of the Reicher-Wheeler paradigm (Reicher 1969), which has proven the best means to assess the identification of letters in words. A scene was presented for a short time (250 ms) and could contain either one of two semantically consistent target objects or one of two semantically inconsistent target objects. For example, a farm scene could contain either a chicken or a pig in the consistent condition, and it could contain either a mixer or a coffee maker in the inconsistent condition. The scene was masked for 30 ms, and the mask was followed immediately by a forced-choice screen displaying two labels either corresponding to the two consistent targets or to the two inconsistent targets. Under these conditions, response bias should be eliminated because contextual information will not assist in discriminating between two consistent object alternatives and it will not assist in discriminating between two inconsistent object alternatives. In addition, this paradigm provides a stronger test of the priming model: Effects of criterion modulation should be reflected in discrimination performance, but such effects may not be reflected in detection sensitivity (Farah 1989, but see Norris 1995). Using this procedure, Hollingworth & Henderson found no advantage for the discrimination of consistent versus inconsistent objects: The nonreliable trend was in the direction of better inconsistent object discrimination. Masson (1991) has reported a similar effect for the discrimination of object tokens using a post-scene, forced-choice procedure.

Conclusions

The majority of studies investigating object identification in scenes have found advantages for consistent versus inconsistent objects. It could be argued that despite the existence of methodological problems in each of these studies, there is sufficient converging evidence to support the general conclusion that consistent scene context facilitates the identification of objects (Rayner & Pollatsek 1992, Boyce & Pollatsek 1992a). Such a conclusion would be plausible if it were not for the fact that the same methodological problem seems to be present in all studies to date that have found advantages for the identification of consistent versus inconsistent objects. Namely, these paradigms do not appear to have adequately discriminated between effects of context on object identification and postidentification effects. Recent experiments indicate that when later effects of context are eliminated from measures of object identification, no consistent object advantage is obtained (Hollingworth & Henderson 1998). Thus, we believe that the functional isolation model currently provides the best explanation of the relation between scene knowledge and object iden-

tification. This conclusion must be viewed as preliminary, however, given the relatively small set of studies that have investigated object identification in scenes.

CONCLUSION

The topics discussed in this chapter include some of the most important outstanding questions remaining for high-level vision. How are the eyes controlled during active scene exploration? What types of representations are constructed and retained as scene viewing unfolds over time? How does the stored knowledge that is accessed during ongoing scene perception interact with incoming perceptual information? The ultimate answers to these questions will have important implications for our understanding of the functional and architectural properties of the human visual and cognitive systems, and so for the fundamental nature of the human mind.

ACKNOWLEDGMENTS

Preparation of this chapter was supported by a grant from the National Science Foundation (SBR 9617274) to JMH and by a National Science Foundation graduate fellowship to AH. We would like to thank Fernanda Ferreira for her insightful comments and discussions.

Literature Cited

Akins K, ed. 1996. *Perception: Vancouver Studies in Cognitive Science,* Vol. 5. Oxford: Oxford Univ. Press

Anstis SM. 1974. A chart demonstrating variations in acuity with retinal position. *Vis. Res.* 14:589–92

Antes JR. 1974. The time course of picture viewing. *J. Exp. Psychol.* 103:62–70

Antes JR, Penland JG, Metzger RL. 1981. Processing global information in briefly presented scenes. *Psychol. Res.* 43:277–92

Ballard DH. 1991. Animate vision. *Artif. Intell.* 48:57–86

Bar M, Ullman S. 1996. Spatial context in recognition. *Perception* 25:343–52

Biederman I. 1972. Perceiving real-world scenes. *Science* 177:77–80

Biederman I. 1981. On the semantics of a glance at a scene. In *Perceptual Organization*, ed. M Kubovy, JR Pomerantz, pp. 213–53. Hillsdale, NJ: Erlbaum

Biederman I. 1987. Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94:115–47

Biederman I. 1988. Aspects and extensions of a theory of human image understanding. In *Computational Processes in Human Vision: An Interdisciplinary Perspective*, ed. ZW Pylyshyn, pp. 370–428. Norwood, NJ: Ablex

Biederman I, Glass AL, Stacy EW Jr. 1973. Searching for objects in real-world scenes. *J. Exp. Psychol.* 97:22–27

Biederman I, Mezzanotte RJ, Rabinowitz JC. 1982. Scene perception: detecting and judging objects undergoing relational violations. *Cogn. Psychol.* 14:143–77

Biederman I, Rabinowitz JC, Glass AL, Stacy

EW Jr. 1974. On the information extracted from a glance at a scene. *J. Exp. Psychol.* 103:597–600
Biederman I, Teitelbaum RC, Mezzanotte RJ. 1983. Scene perception: a failure to find a benefit from prior expectancy or familiarity. *J. Exp. Psychol.: Learn. Mem. Cogn.* 9:411–29
Boyce SJ, Pollatsek A. 1992a. An exploration of the effects of scene context on object identification. See Rayner 1992, pp. 227–42
Boyce SJ, Pollatsek A. 1992b. Identification of objects in scenes: the role of scene background in object naming. *J. Exp. Psychol.: Learn. Mem. Cogn.* 18:531–43
Boyce SJ, Pollatsek A, Rayner K. 1989. Effect of background information on object identification. *J. Exp. Psychol.: Hum. Percept. Perform.* 15:556–66
Bridgeman B, Hendry D, Stark L. 1975. Failure to detect displacements of the visual world during saccadic eye movements. *Vis. Res.* 15:719–22
Bridgeman B, Mayer M. 1983. Failure to integrate visual information from successive fixations. *Bull. Psychon. Soc.* 21:285–86
Brietmeyer BG, Kropfl W, Julesz B. 1982. The existence and role of retinotopic and spatiotopic forms of visual persistence. *Acta Psychol.* 52:175–96
Bruner JS. 1957. On perceptual readiness. *Psychol. Rev.* 64:123–52
Bruner JS. 1973. *Beyond the Information Given*. New York: Norton
Bülthoff HH, Edelman SY, Tarr MJ. 1995. How are three-dimensional objects represented in the brain? *Cereb. Cortex* 3: 247–60
Buswell GT. 1935. *How People Look at Pictures*. Chicago: Univ. Chicago Press
Carpenter RHS. 1988. *Movements of the Eyes.* London: Pion
Churchland PS, Ramachandran VS, Sejnowski TJ. 1994. A critique of pure vision. See Koch & Davis 1994, pp. 23–60
Davidson ML, Fox MJ, Dick AO. 1973. Effect of eye movements on backward masking and perceived location. *Percept. Psychophys.* 14:110–16
De Graef P. 1992. Scene-context effects and models of real-world perception. See Rayner 1992, pp. 243–59
De Graef P, Christiaens D, d'Ydewalle G. 1990. Perceptual effects of scene context on object identification. *Psychol. Res.* 52: 317–29
Dennett DC. 1991. *Consciousness Explained*. Boston: Little Brown
Deubel H, Schneider WX. 1996. Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vis. Res.* 36:1827–37
Deubel H, Schneider WX, Bridgeman B. 1996. Postsaccadic target blanking prevents saccadic suppression of image displacement. *Vis. Res.* 36:985–96
Di Lollo V. 1977. Temporal characteristics of iconic memory. *Nature* 267:241–43
Duhamel JR, Colby CL, Goldberg ME. 1992. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* 255:90–92
Epstein R, Kanwisher N. 1998. A cortical representation of the local visual environment. *Nature* 392:598–601
Eriksen CW, Collins JF. 1967. Some temporal characteristics of visual pattern recognition. *J. Exp. Psychol.* 74:476–84
Farah MJ. 1989. Semantic and perceptual priming: How similar are the underlying mechanisms? *J. Exp. Psychol.: Hum. Percept. Perform.* 15:188–94
Feldman JA. 1985. Four frames suffice: a provisional model of vision and space. *Behav. Brain Sci.* 8:265–89
Findlay JM. 1982. Global processing for saccadic eye movements. *Vis. Res.* 22: 1033–45
Fodor JA. 1983. *Modularity of Mind*. Cambridge, MA: MIT Press
Friedman A. 1979. Framing pictures: the role of knowledge in automatized encoding and memory for gist. *J. Exp. Psychol.: Gen.* 108:316–55
Friedman A, Liebelt LS. 1981. On the time course of viewing pictures with a view towards remembering. In *Eye Movements: Cognition and Visual Perception*, ed. DF Fisher, RA Monty, JW Senders, pp. 137–55. Hillsdale, NJ: Erlbaum
Grimes J. 1996. On the failure to detect changes in scenes across saccades. See Akins 1996, pp. 89–110
Henderson JM. 1992a. Visual attention and eye movement control during reading and picture viewing. See Rayner 1992, pp. 260–83
Henderson JM. 1992b. Object identification in context: the visual processing of natural scenes. *Can. J. Psychol.* 46:319–41
Henderson JM. 1996. Visual attention and the attention-action interface. See Akins 1996, pp. 290–316
Henderson JM. 1997. Transsaccadic memory and integration during real-world object perception. *Psychol. Sci.* 8:51–55
Henderson JM, Hollingworth A. 1998. Eye movements during scene viewing: an overview. See Underwood 1998, pp. 269–95
Henderson JM, McClure KK, Pierce S, Schrock G. 1997. Object identification

without foveal vision: evidence from an artificial scotoma paradigm. *Percept. Psychophys.* 59:323–46
Henderson JM, Pollatsek A, Rayner K. 1987. The effects of foveal priming and extrafoveal preview on object identification. *J. Exp. Psychol.: Hum. Percept. Perform.* 13: 449–63
Henderson JM, Pollatsek A, Rayner K. 1989. Covert visual attention and extrafoveal information use during object identification. *Percept. Psychophys.* 45:196–208
Henderson JM, Weeks PA Jr, Hollingworth A. 1999. The effects of semantic consistency on eye movements during scene viewing. *J. Exp. Psychol.: Hum. Percept. Perform.* In press
Hollingworth A, Henderson JM. 1998. Does consistent scene context facilitate object perception? *J. Exp. Psychol.: Gen.* In press
Humphreys GW, Riddoch MJ. 1994. Attention to within-object and between-object spatial representations: multiple sites for visual selection. *Cogn. Neuropsychol.* 11: 207–41
Humphreys GW, Riddoch MJ. 1995. Separate coding of space within and between perceptual objects: evidence from unilateral visual neglect. *Cogn. Neuropsychol.* 12: 283–311
Irwin DE. 1991. Information integration across saccadic eye movements. *Cogn. Psychol.* 23:420–56
Irwin DE. 1992. Perceiving an integrated visual world. In *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*, ed. DE Meyer, S Kornblum, pp. 121–42. Cambridge, MA: MIT Press
Irwin DE. 1996. Integrating information across saccadic eye movements. *Curr. Dir. Psychol. Sci.* 5:94–100
Irwin DE, Yantis S, Jonides J. 1983. Evidence against visual integration across saccadic eye movements. *Percept. Pychophys.* 34: 35–46
Irwin DE, Zacks JL, Brown JS. 1990. Visual memory and the perception of a stable environment. *Percept. Pychophys.* 47:35–46
Jonides J, Irwin DE, Yantis S. 1982. Integrating visual information from successive fixations. *Science* 215:192–94
Koch C, Davis JL, eds. 1994. *Large-Scale Neuronal Theories of the Brain*. Cambridge, MA: MIT Press
Kosslyn SM. 1994. *Image and Brain*. Cambridge, MA: MIT Press
Land MF, Furneaux S. 1997. The knowledge base of the oculomotor system. *Philos. Trans. R. Soc. London Ser. B* 352:1231–39
Levin DT, Simons DJ. 1997. Failure to detect changes to attended objects in motion pictures. *Pschonom. Bull. Rev.* 4:501–6
Loftus GR, Mackworth NH. 1978. Cognitive determinants of fixation location during picture viewing. *J. Exp. Psychol.: Hum. Percept. Perform.* 4:565–72
Loftus GR, Nelson WW, Kallman HJ. 1983. Differential acquisition rates for different types of information from pictures. *Q. J. Exp. Psychol.* 35A:187–98
Mackworth NH, Morandi AJ. 1967. The gaze selects informative details within pictures. *Percept. Psychophys.* 2:547–52
Mannan S, Ruddock KH, Wooding DS. 1995. Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. *Spat. Vis.* 9:363–86
Mannan SK, Ruddock KH, Wooding DS. 1996. The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spat. Vis.* 10:165–88
Mannan SK, Ruddock KH, Wooding DS. 1997. Fixation sequences made during visual examination of briefly presented 2D images. *Spat. Vis.* 11:157–78
Marr D. 1982. *Vision.* San Francisco: Freeman
Marr D, Nishihara HK. 1978. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. London Ser. B* 200:269–94
Masson MEJ. 1991. Constraints on the interaction between context and stimulus information. *Proc. Conf. Cogn. Sci. Soc., 13th, Chicago,* ed. KJ Hammond, D Gentner, pp. 540–45. Hillsdale, NJ: Erlbaum
Matin E. 1974. Saccadic suppression: a review and an analysis. *Psychol. Bull.* 81:899–917
McConkie GW. 1990. *Where vision and cognition meet.* Presented at Hum. Front. Sci. Program Workshop Object Scene Percept., Leuven, Belgium
McConkie GW, Currie CB. 1996. Visual stability across saccades while viewing complex pictures. *J. Exp. Psychol.: Hum. Percept. Perform.* 22:563–81
McConkie GW, Rayner K. 1975. The span of the effective stimulus during a fixation in reading. *Percept. Psychophys.* 17:578–86
Metzger RL, Antes JR. 1983. The nature of processing early in picture perception. *Psychol. Res.* 45:267–74
Mullen KT. 1990. The chromatic coding of space. In *Vision: Coding and Efficiency,* ed. C Blakemore, pp. 150–58. Cambridge: Cambridge Univ. Press
Mumford D. 1994. Neuronal architectures for pattern-theoretic problems. See Koch & Davis 1994, pp. 125–52
Neisser U. 1967. *Cognitive Psychology.* Englewood Cliffs, NJ: Prentice Hall

Norris D. 1995. Signal detection theory and modularity: on being sensitive to the power of bias models of semantic priming. *J. Exp. Psychol.: Hum. Percept. Perform.* 21:935–39

Oliva A, Schyns PG. 1997. Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cogn. Psychol.* 34: 72–107

O'Regan JK. 1992a. Optimal viewing position in words and the strategy-tactics theory of eye movements in reading. See Rayner 1992, pp. 333–54

O'Regan JK. 1992b. Solving the "real" mysteries of visual perception: the world as an outside memory. *Can. J. Psychol.* 46: 461–88

O'Regan JK, Levy-Schoen A. 1983. Integrating visual information from successive fixations: does trans-saccadic fusion exist? *Vis. Res.* 23:765–68

O'Regan JK, Rensink RA, Clark JJ. 1996. "Mud splashes" render picture changes invisible. *Invest. Opthamol. Vis. Sci.* 37:979

Palmer SE. 1975a. The effects of contextual scenes on the identification of objects. *Mem. Cogn.* 3:519–26

Palmer SE. 1975b. Visual perception and world knowledge: notes on a model of sensory-cognitive interaction. In *Explorations in Cognition,* ed. DA Norman, DE Rumelhart, LNR Res. Group, pp. 279–307. San Francisco: Freeman

Pollatsek A, Rayner K. 1992. What is integrated across fixations? See Rayner 1992, pp. 166–91

Potter MC. 1975. Meaning in visual search. *Science* 187:965–66

Potter MC. 1976. Short-term conceptual memory for pictures. *J. Exp. Psychol.: Hum. Learn. Mem.* 2:509–22

Potter MC, Levy EI. 1969. Recognition memory for a rapid sequence of pictures. *J. Exp. Psychol.* 81:10–15

Pouget A, Fisher SA, Sejnowski TJ. 1993. Egocentric spatial representation in early vision. *J. Cogn. Neurosci.* 5:150–61

Pylyshyn Z. 1980. Computation and cognition: issues in the foundations of cognitive science. *Behav. Brain Sci.* 3:111–32

Pylyshyn Z. 1998. Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behav. Brain Sci.* In press

Rayner K. 1978. Eye movements in reading and information processing. *Psychol. Bull.* 85:618–60

Rayner K, ed. 1992. *Eye Movements and Visual Cognition: Scene Perception and Reading*. New York: Springer-Verlag

Rayner K, Pollatsek A. 1983. Is visual information integrated across saccades? *Percept. Psychophys.* 34:39–48

Rayner K, Pollatsek A. 1989. *The Psychology of Reading*. Englewood Cliffs, NJ: Prentice-Hall

Rayner K, Pollatsek A. 1992. Eye movements and scene perception. *Can. J. Psychol.* 46: 342–76

Reicher GM. 1969. Perceptual recognition as a function of meaningfulness of stimulus material. *J. Exp. Psychol.* 81:275–80

Rensink RA, O'Regan JK, Clark JJ. 1997. To see or not to see: the need for attention to perceive changes in scenes. *Psychol. Sci.* 8:368–73

Riggs LA. 1965. Visual acuity. In *Vision and Visual Perception,* ed. CH Graham, pp. 321–49. New York: Wiley

Rumelhart DE, Smolensky P, McClelland JL, Hinton GE. 1986. Schemata and sequential thought processes in PDP models. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Psychological and Biological Models*, ed. JL McClelland, DE Rumelhart, PDP Res. Group, 2:7–57. Cambridge, MA: MIT Press

Schyns PG, Oliva A. 1994. From blobs to boundary edges: evidence for time and spatial scale dependent scene recognition. *Psychol. Sci.* 5:195–200

Ullman S. 1996. *High-Level Vision: Object Recognition and Visual Cognition.* Cambridge, MA: MIT Press

Underwood G, ed. 1998. *Eye Guidance in Reading and Scene Perception*. Oxford: Elsevier

van Diepen PMJ, De Graef P, d'Ydewalle G. 1995. Chronometry of foveal information extraction during scene perception. In *Eye Movement Research: Mechanisms, Processes and Applications,* ed. JM Findlay, R Walker, RW Kentridge, pp. 349–62. Amsterdam: Elsevier. In press

van Diepen PMJ, Wampers M, d'Ydewalle G. 1998. Functional division of the visual field: moving masks and moving windows. See Underwood 1998. In press

Volkmann FC. 1986. Human visual suppression. *Vis. Res.* 26:1401–16

Yarbus AL. 1967. *Eye Movements and Vision.* New York: Plenum