Memory for Object Position in Natural Scenes

Andrew Hollingworth

Department of Psychology
The University of Iowa

Please Address Correspondence to:
Andrew Hollingworth
Department of Psychology
The University of Iowa
11 Seashore Hall E
Iowa City, IA 52242-1407
andrew-hollingworth@uiowa.edu

September 1, 2003
Abstract

Memory for the absolute position of objects in natural scenes was investigated. Participants viewed an image of a real world scene (preview scene), followed by a target object in isolation (target probe), followed by a blank screen with a mouse cursor. In the target present preview condition, the target object was present in the preview scene. Participants moved the cursor to the remembered position of the target object. In the target absent preview condition, the target object was not present in the preview scene. Participants moved the cursor to a position where the target likely would have appeared in the preview scene. In the no preview condition, no preview scene was displayed. Participants imagined a scene and moved the cursor to a position where the target likely would appear. Localization accuracy in the target present preview condition was reliably higher than that in the target absent preview condition, which was reliably higher than localization accuracy in the no preview condition. These data demonstrate that participants can remember both the spatial context of a scene and the absolute positions of local objects.
Successful behavior within an environment depends on knowing the nature of the objects present and their locations. Vision plays the primary role in extracting this information. Vision allows us to perceive the shape of objects and to classify these objects into meaningful categories. Vision also allows us to accurately perceive the spatial location of objects, relative to ourselves and to other objects. If one can perceive an object’s visual form, know what kind of object it is, and know where it is located, then one can typically interact with it intelligently. For example, reaching for a cup of coffee depends on perceiving the shape of an object and recognizing it as belonging to the category of cups that contain coffee; perceiving the perceptual details of the particular cup in view, its orientation, the shape of the handle, and so on, to support grasping; and perceiving the position of the cup in the visual field, to support accurate reaching.

The environments we typically inhabit contain many, often hundreds, of individual objects. The visual system cannot process the details of all these objects simultaneously due to resource limitations and due to the fact that high-resolution sensory information is available for only a very small portion of the visual field (corresponding to the foveal region of the retina). These limitations lead to the sequential selection of objects, with the eyes shifted via saccadic eye movements to fixate individual objects and obtain high-resolution information (see Henderson & Hollingworth, 1998). As a result, visual scene perception is extended over time and space as the eyes are oriented from object-to-object. To construct a scene representation, memory is necessary to integrate information acquired from different eye fixations on local objects. Thus, our ability to represent which objects are where in an environment depends critically on visual memory.

The phenomenon of change blindness, in which participants can be surprisingly insensitive to scene changes when detection requires memory, has been taken by many authors to indicate that scene representations are visually impoverished, with little or no accumulation of
information as the eyes and attention are oriented from object-to-object within a scene (Becker & Pashler, 2002; O’Regan, 1992; Rensink, 2000; Rensink, O’Regan, & Clark, 1997; Scholl, 2000; Simons, 1996; Simons & Levin, 1997; Wolfe, 1999). In contrast to this view, Hollingworth & Henderson (Hollingworth, 2003, submitted a, submitted b; Hollingworth & Henderson, 2002; Hollingworth, Williams, & Henderson, 2001) have demonstrated significant accumulation of visual information from many individual objects within a scene. These studies examined memory for the visual form of individual objects. The typical method involved initial presentation of a complex natural scene, containing many individual objects, for viewing. Memory for a single object was then tested in either a change detection or two-alternative forced choice test. The changed version or distractor was either a different object from the same basic-level category or the same object rotated 90° in depth. Participants could accurately perform these tasks even when as many as ten objects intervened between target fixation and test (Hollingworth, submitted a; Hollingworth & Henderson, 2002), and memory for object form was still well above chance across many intervening scenes (Hollingworth, submitted b; Hollingworth, 2002) and across more than 400 intervening objects (Hollingworth, submitted a).

These studies demonstrate that people can accurately remember and accumulate information about the visual form of many individual objects. But how accurately can participants remember the locations of individual objects in scenes? Simons (1996) showed that participants can detect changes to the global spatial configuration (or layout) of objects, but that study did not require memory for the binding of a particular object to a particular location. In fact, Simons argued that changes were detected based on the retention of spatial layout independently of information about local object properties (i.e., as a configuration of abstract locations). Irwin & Zelinsky (2002) displayed an array of seven objects, followed by a blank screen and a location probe. Participants were able to accurately identify which object had
appeared at the probed location, at least for objects that had been fixated recently. This study provides evidence that visual memory can retain the binding of a particular object to a location. However, the Irwin & Zelinsky study used the same set of seven discrete locations on every trial and thus can only provide limited evidence regarding memory for the absolute position of objects in scenes.

The present experiment sought to examine memory for the absolute position of individual objects in natural scenes. There were three conditions. In the target present preview condition, an image of a realistic, natural scene was presented for 20 s (the preview scene). This was followed by an image of a single target object in the center of the screen (the target probe image). Participants did not know which object in the preview scene would be the target. Next, the target probe was removed, and a blank screen was displayed with a mouse cursor. Participants moved the cursor to indicate the position at which the target object had appeared in the preview scene. The target absent preview condition was identical, except that the target object was not present in the preview scene. Participants were instructed to place the mouse cursor at the position where the target object would likely have appeared if it had been present in the preview scene. In this condition, participants could use memory for the layout of contextual objects and surfaces—combined with information about target identity, size, and other cues—to estimate placement. The critical comparison was between position accuracy in the target present and target absent preview conditions. If position estimates were found to be more accurate in the former condition, this would provide strong evidence that participants remembered the specific position of the target object when it was present in the preview.

Finally, in the no preview condition, no preview scene was displayed. Participants were instructed to imagine a scene that could contain the target object and choose a position where that object might likely appear. Even with no preview scene, target information could guide
placement: A small airplane target object would more plausibly be found in the upper portions of a scene than in the lower portions. A second important comparison was between performance in the no preview and target absent preview conditions. If position estimates in the target absent preview condition were more accurate than in the no preview condition, this would demonstrate that participants remembered information about the spatial context of the scene, such as the arrangement of surfaces and objects.

Participants. Eighteen participants from the University of Iowa community, 18 to 30 years-old, completed the experiment. All participants reported 20/20 uncorrected or corrected vision.

Stimuli. Scene images were created from a set of 48 different 3-D models. Each model depicted a complex, real-world environment. A unique target object was chosen within each model, with the constraint that the target could not appear in the center of the image. To produce the target present preview images, each model was rendered with the target object present in the scene. To produce the target absent preview images, the scene was rendered after the target object had been removed from the model. Sample target present and absent preview images are displayed in panels A and B of Figure 1. Target probe images (panel C of Figure 1) presented the target object in isolation in the center of the screen. For each scene, the same target probe was used in all three preview conditions. The target probe images were created by applying a transparent texture to all objects except the target object. Though not visible in the rendered image, background objects still reflected light and cast shadows within the model, ensuring that the target object appearance was identical to that in the preview images. The target object was then moved to the center of the image. The target probe background was set to a uniform olive green, chosen because none of the target objects contained this color and thus would not blend into the background.
Scene images subtended 16.9° x 22.8° visual angle at a viewing distance of 80 cm. Target objects subtended 3.33° on average along the longest dimension in the picture plane.

**Apparatus.** The stimuli were displayed at a resolution of 800 by 600 pixels by 24-bit color on a 17-in. monitor at a refresh rate of 100 Hz. Position responses were collected using a mouse. The presentation of stimuli and collection of responses was controlled by E-Prime software running on a Pentium IV-based computer. Viewing distance was maintained at 80 cm by a forehead rest. The room was dimly illuminated by a low-intensity light source.

**Procedure.** Participants were tested individually. Each participant was given a written description of the experiment along with a set of instructions. Participants were informed that on some trials they would see an image of a scene for 20 s. On other trials no scene would be displayed. Next, a single target object would be presented in the center of the screen (the target probe image). On some trials, the target object had been present in the preview scene; on other trials it had not. A label appearing between the scene and the probe image would inform them whether the target object had been present in the scene, had been absent from the scene, or no preview had been displayed. Participants were instructed that after the target probe image, a mouse cursor would appear on a blank screen. For the target present condition, they were to move the mouse cursor to the position on the blank screen corresponding to the center of the object as it appeared in the scene. For the target absent condition, they were instructed to place the mouse cursor at a position where such an object was likely to have appeared in the preview scene if it had been present. For the no preview condition, participants were instructed to imagine a scene that might contain that object and move the mouse cursor to a position where such an object might likely appear. For this final condition, an example was used in which a fire hydrant probe object might appear in a street scene and would likely be found in the lower half.
Participants pressed a mouse button to initiate each trial. Then, a blank, olive-green screen was displayed for 1000 ms. This was followed by the preview scene for 20 s, except in the case of the no preview condition. Next, an olive-green screen was presented with either “present”, “absent”, or “no preview” displayed in large, white text for 1200 ms. This was followed by the target probe image for 2200 ms. Finally, a blank, olive-green screen was presented with a white, plus-sign cursor initially in the center. Participants moved the cursor with the mouse to their chosen position on the screen and clicked to register the response. A mouse click ended the trial.

Participants first completed a practice session of six trials, two in each of the three preview conditions. Practice scenes were not used in the experimental session. The practice trials were followed by 48 experimental trials, 16 in each of the three preview conditions. Each scene item was displayed once; there was no scene repetition. Across participants, each of the 48 scene items appeared in each condition an equal number of times. The entire experiment lasted approximately 40 minutes.

**Results and Discussion**

On each trial, the Euclidean distance between the response position and the center of the target object was calculated in degrees of visual angle, the *position error*. For each scene item, the same target position was used in all three preview conditions. Mean position error in each of the preview conditions is displayed in the top panel of Figure 2. Means position error is also illustrated in the bottom panel of Figure 2, superimposed over the sample scene. There was a reliable main effect of preview condition, $F(2,34) = 179.52, p < .001$. Mean position error was reliably smaller in the target present preview condition ($3.17^\circ$) than in the target absent preview condition ($4.48^\circ$), $F(1,17) = 31.57, p < .001$. And mean position error was reliably smaller in the
target absent preview condition than in the no preview condition (8.36°), $F(1,17) = 151.32, p < .001$.

These data demonstrate that scene representations maintain both information about the general spatial context of the scene and information about the specific locations of individual objects. The large difference in position error for the target absent preview and no preview conditions demonstrates that participants can remember the spatial structure of a scene. These two conditions were controlled for reasoning based on target object features (such as identity or size), since the same target objects were used in both conditions. The only difference was the presentation of the scene context prior to position estimation. The contextual preview clearly provided a great deal of information. Despite the absence of the target object from the preview scene, participants could quite accurately estimate the target object position. This implies they were able to remember the global spatial structure of the scene, such as the arrangement of surfaces and objects. In addition, it seems unlikely that accurate placement in the target absent preview condition could have been supported solely by memory for abstract spatial structure. Accurate placement implies that participants were able to combine knowledge of contextual features (e.g., that kitchen counters tend to support appliances) with target object information (e.g., that the target object is a toaster). In general, performance in the target absent preview conditions is consistent with previous work demonstrating that people can generate fairly rich and accurate representations of natural scenes (Hollingworth, 2003, submitted a, submitted b; Hollingworth & Henderson, 2002; Hollingworth, Williams, & Henderson, 2001)

Although localization performance was quite accurate in the target absent preview condition, it was even more accurate in the target present preview condition. The target present and absent preview conditions were controlled for reasoning based on general contextual
memory and target object features. The only difference was the presence of the target object in the preview. Thus, the advantage for the target present preview condition demonstrates that participants could remember the absolute position of the target object (at least on a significant proportion of trials). Mean position error in the target present preview condition was 3.17° within an image that subtended 16.9° x 22.8°. Note that error was calculated from the center of the target object, and target objects subtended approximately 3° on average, so position estimates in the target present preview condition were typically less than 2° from the nearest contour of the object. This relatively accurate localization was observed despite the fact that participants did not know which object in the preview scene would be the target. As is evident from the sample scene in Figure 1, there were multiple objects in each scene. Accurate position estimates therefore imply that participants remembered the positions of multiple objects within the scene. The precise number of objects is difficult to estimate from these data. One interesting question for future research is whether memory for object position in natural scenes exceeds the 3-4 object capacity of VSTM (see Irwin & Zelinksy, 2002).

Evidence of memory for target location and spatial context has implications for the recent debate over the role of memory in visual search (e.g., Shore & Klein, 2000; Wolfe, 1999). If participants can remember target location and the spatial context of a scene, it seems reasonable that such memory representations could guide visual search. Hollingworth (2003 b) tested this possibility in a search task. As in the present study, participants were shown either a scene preview with the target object present, a scene preview with the target object absent, or no preview. They were then shown the target probe in isolation. This was followed by a search scene, which was the same basic scene as the preview. The target object was always present in the search scene, and it was either the same as the target probe or mirror reversed. Participants’ task was to find the target and determine whether it was in the same orientation as the target probe or not.
Hollingworth (2003b) found the same basic pattern of results as in the present experiment. Search speed (measured both as overall response time and elapsed time until target fixation) was fastest in the target present preview condition, next fastest in the target absent preview condition, and slowest in the no preview condition. These data indicate that the types of spatial memory observed in the present study can be employed dynamically to guide visual search in natural scenes.

In other studies of object position memory, memory for object position displays systematic biases. For example, for an object hidden in a rectangular sand box, children show a systematic position bias toward the center of the left or right half of the rectangle (Huttenlocher, Newcombe, & Sandberg, 1994). Are systematic biases present in memory for the positions of real-world objects in natural scenes? One difficulty in using the present data to answer this question is that on some trials, participants may simply have failed to remember the position of the target object, leading them to guess a position. Guessing itself could lead to what appears to be systematic bias. If an object is on the right-hand side of the screen, the mean position estimate based on guessing should be near the center of the scene, which could appear to be a bias in position estimate toward the center of the image.

To avoid this problem, only relatively accurate position estimates were used to examine possible position biases in the present data. Position estimates in the target present preview condition were examined, with the constraint that the estimate must be less than 2° from the center of the target. Figure 3 illustrates the size of this 2° constraint using the sample scene. Of the total 288 position estimates in the target present preview condition, 110 met this criterion. These position estimates were divided based on the position of the target object in the scene, left-versus right-hand side of the screen and upper versus lower half of the screen. Each position estimate is plotted in Figures 4 and 5 as a function of the horizontal and vertical distance from
the center of the target object, which is plotted at the intersection of the dotted lines. Note that each position estimate appears twice, once for the horizontal division and once for the vertical division.

<< Insert Figure 3 about here >>

For target objects appearing on the left-hand side of the screen, there was no clear horizontal bias in position estimate. Thirty-one of 60 position estimates (51.7%) were to the right of target center, which was not reliably different from chance of 50%, chi-square = .02, \( p = .90 \).

For the target objects appearing on the right hand side of the screen, there was a clear bias to the left. Thirty-seven of the 49 position estimates (75.5%) were to the left of target center, which was reliably higher than chance, chi-square = 12.76, \( p < .001 \).1 For target objects appearing in the upper half of the screen, there was a clear position bias lower. Nineteen of the 23 position estimates (82.6%) were below target center, which was reliably greater than chance, chi-square = 9.78, \( p < .005 \). Finally, for target objects appearing in the lower half of the screen, there was a clear position bias higher. Fifty-seven of the 87 position estimates (65.5%) were above target center, which was reliably greater than chance, chi-square = 8.38, \( p < .005 \). With the exception of target objects on the left-hand side of the screen, estimates were consistently biased toward the center of the image.2

<< Insert Figures 4 and 5 about here >>

What accounts for this general bias toward the center of the image? The effect is consistent with the scene memory phenomenon of boundary extension (see Intraub, 1997, for a

---

1 One position estimate on the left/right dimension was at the precise horizontal center of the target and was dropped from the analysis.

2 The same trends were evident even when considering only estimates within 1° of target center (open circles in Figures 4 and 5). For target objects on the left-hand side of the screen, 15 of 27 estimates were to the right of target center (55.6%). For target objects on the right-hand side of the screen, 11 of 16 estimates were to the left of target center (68.8%). For target objects in the upper half of the screen, 6 of 7 estimates were below target center (85.7%). For target objects in the lower half of the screen, 20 of 35 estimates were above target center (57.1%).
review) and could derive from the same source. Intraub has found that when participants are asked to draw a complex picture from memory, they tend to expand the scope of the picture, suggesting that scene memory is expanded to include information beyond the boundaries of the original image. As a result, objects that were originally visible in the image will be spatially compressed to some extent. If, in the present experiment, scene boundaries were extended in memory, then one would expect position estimates to be biased toward the center of the image. Further analysis of this effect must await research specifically designed to examine position biases.

**Conclusion**

After viewing a preview of a natural scene, participants were able to fairly accurately estimate the position at which a target object would have appeared in the scene if it had been present (target absent preview condition). This result demonstrates memory for the general spatial and contextual features of a scene. In addition, position estimates were even more accurate when the target object was actually present in the preview scene (target present preview condition), demonstrating memory for the absolute positions of individual objects. These findings inform our basic understanding of the nature of scene representations. Contrary to claims that scene representations are visually impoverished (e.g., O’Regan, 1992; Rensink et al., 1997), the present data are consistent with the claim that visual memory supports scene representations containing information about the visual form and locations of multiple individual objects (Hollingworth & Henderson, 2002). In addition, these data suggest that memory for the positions of objects in a scene could support important visual tasks, such as visual search.
References


Hollingworth, A. (submitted b). The relationship between online visual representation of a scene and long-term scene memory.


Acknowledgements

This research was supported by NIMH grant R03 MH65456. Correspondence concerning this article should be addressed to Andrew Hollingworth, The University of Iowa, Department of Psychology, 11 Seashore Hall E, Iowa City, IA, 52242-1407. Email should be sent to andrew-hollingworth@uiowa.edu.
**Figure Captions**

*Figure 1.* Principal stimuli for one of the 48 scene items. Panel A shows the preview scene with the target object present (barbell). Panel B shows the preview scene with the target object absent. Panel C shows the target probe image.

*Figure 2.* Top panel: Mean position error in degrees of visual angle for the three preview conditions. Error bars are standard errors of the means. Bottom panel: Mean position error (white circles) superimposed over a sample scene and target object (barbell) to illustrate position error relative to the size of the scenes. Note that the means are collapsed across all 48 scene items.

*Figure 3.* The size of the response region used to examine position biases.

*Figure 4.* Individual position estimates in the target present preview condition for objects appearing on the left and right sides of the screen. Each estimate is plotted relative to the center of the target object on that trial (the intersection of the dotted lines). Only estimates within 2° of target center are displayed. Open circles depict estimates within 1° of target center.

*Figure 5.* Individual position estimates in the target present preview condition for objects appearing on the top and bottom halves of the screen. Each estimate is plotted relative to the center of the target object on that trial (the intersection of the dotted lines). Only estimates within 2° of target center are displayed. Open circles depict estimates within 1° of target center.
Hollingworth Object Position

Target Present | Target Absent | No Preview
--- | --- | ---
Position Error (° visual angle)

3.17 | 4.48 | 8.36

Preview Condition