VISUAL COGNITION, 2005, 12 (6), 1003–1016

Psychology Press
Taylor & Francis Group

Memory for object position in natural scenes

Andrew Hollingworth

Department of Psychology, The University of Iowa, Iowa City, USA

Memory for the positions 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. Participants estimated the position of the target using the mouse. Three conditions were compared. In the *target present preview* condition, the target object was present in the scene preview. In the *target absent preview* condition, the target object not present in the scene preview. In the *no preview* condition, no preview scene was displayed. 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 specific positions of local objects.

Successful behaviour 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-

Please address all correspondence to: Andrew Hollingworth, The University of Iowa, Department of Psychology, 11 Seashore Hall E, Iowa City, IA, 52242-1407, USA.
Email: andrew-hollingworth@uiowa.edu

This research was support by NIMH grant R03 MH65456.

http://www.tandf.co.uk/journals/pp/13506285.html DOI:10.1080/13506280444000625

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; Irwin & Andrews, 1996; Irwin & Zelinsky, 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 and Henderson (Hollingworth, 2003a, 2004, in press; Hollingworth & Henderson, 2002; Hollingworth, Williams, & Henderson, 2001) have demonstrated significant accumulation of visual information from many individual objects within a scene (see Castelhano & Henderson, this issue, for evidence that such accumulation can occur incidentally). 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 10 objects intervened between target fixation and test (Hollingworth, 2004; Hollingworth & Henderson, 2002), and memory for object form was still well above chance across many intervening scenes (Hollingworth, in press; Hollingworth & Henderson, 2002) and across more than 400 intervening objects (Hollingworth, 2004).

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 and 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 scene location. However, the Irwin and Zelinsky study used the same set of seven discrete locations on every trial and thus can only provide limited evidence regarding memory for the positions of objects in real-world scenes.

The present experiment sought to examine memory for the positions 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 centre 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 to the target present preview condition, 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. Such knowledge of where an object is likely to appear in a scene has been demonstrated to facilitate object localization during search (Henderson, Weeks, & Hollingworth, 1999; Hollingworth, 2003b; Oliva, Torralba, Castelhano, & Henderson, 2003). The critical comparison in the present experiment 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.

METHOD

Participants

Eighteen participants from the University of Iowa community, 18–30 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 centre 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 centre 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 centre of the image. The target probe background was set to a uniform olive green, chosen because none of the target objects contained this colour and thus would not blend into the background.

Scene images subtended $16.9^\circ \times 22.8^\circ$ 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 800×600 pixels $\times$ 24-bit colour on a 17 inch 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

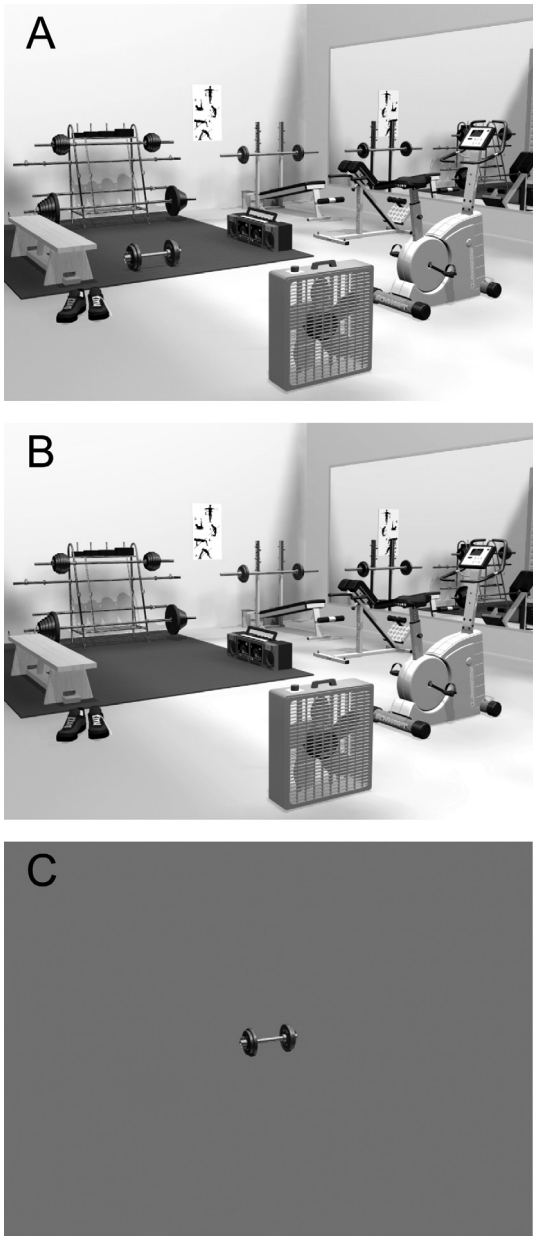


Figure 1. Principal stimuli for one of the 48 scene items. A shows the preview scene with the target object present (barbell). B shows the preview scene with the target object absent. C shows the target probe image.

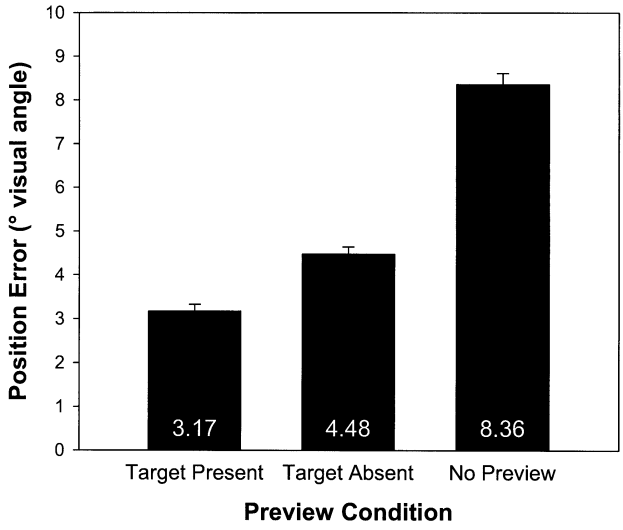
other trials no scene would be displayed. Next, a single target object would be presented in the centre of the screen (the target probe image). On some trials, the target object had been present in the preview scene; on other trials it had not. A label appearing between the scene and the probe image 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 centre 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 of the scene.

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 centre. 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. Trials from the three preview conditions were randomly intermixed. Across participants, each of the 48 scene items appeared in each condition an equal number of times. The entire experiment lasted approximately 40 min.

RESULTS AND DISCUSSION

On each trial, the Euclidean distance between the response position and the centre 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. Mean 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



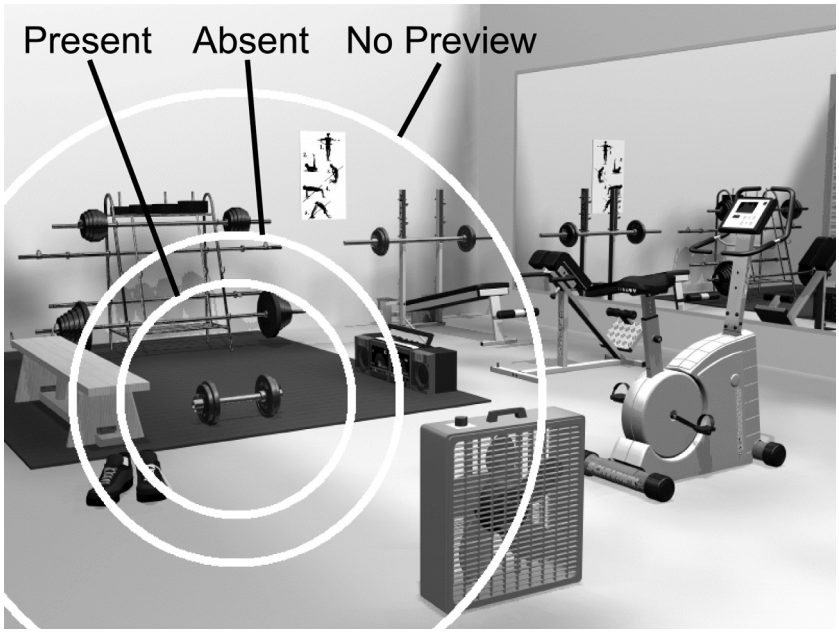


Figure 2. Top panel: Mean position error in degrees of visual angle for the three preview conditions. Error bars are standard errors of the means. Bottom panel: Mean position error (white circles) superimposed over a sample scene and target object (barbell) to illustrate position error relative to the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

position error was reliably smaller in the target present preview condition (3.17°) than in the target absent preview condition (4.48°), $F(1, 17) = 31.57$, $p < .001$. And mean position error was reliably smaller in the target absent preview condition than in the no preview condition (8.36°), $F(1, 17) = 151.32$, $p < .001$. This pattern of results has been replicated using a modified set of scenes and a preview duration of 10 s (Hollingworth, 2003b).

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 global 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 condition is consistent with previous work demonstrating that people can generate fairly rich and accurate representations of natural scenes (Hollingworth, 2003a, 2004, in press; Hollingworth & Henderson, 2002; Hollingworth et al., 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 specific 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° × 22.8°. Note that error was calculated from the centre 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 three or four object capacity of VSTM (Hollingworth, 2004; 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 (2003b) tested this possibility in a search task. Participants were shown a scene preview for 10 s. As in the present study, the target object was either present in the preview or absent from the preview. A no preview control condition provided baseline search data. After the scene preview, participants saw 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 in the search scene and determine whether it was in the same orientation as the target probe or not. Hollingworth found a similar 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 addition, preview benefits were still observed when the preview duration was reduced to 2 s and to 500 ms, demonstrating that spatial information sufficient to guide search can be acquired relatively rapidly from a scene.

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, older children show a systematic position bias toward the centre 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 centre of the scene, which could appear to be a bias in position estimate toward the centre 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 centre of the target. Figure 3 illustrates

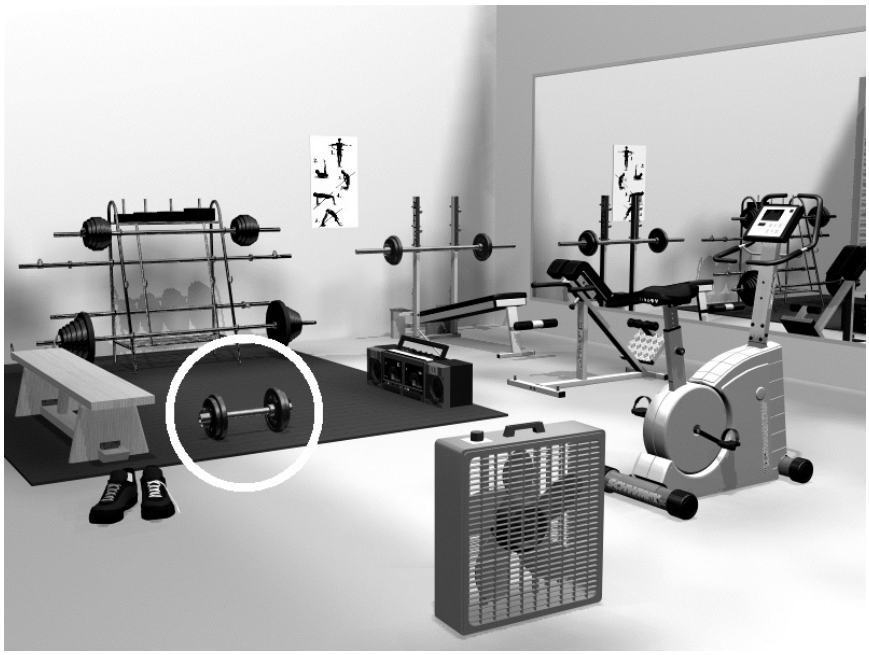


Figure 3. The size of the response region used to examine position biases.

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 centre 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.

For target objects appearing on the left-hand side of the screen, there was no clear horizontal bias in position estimate. Thirty-one of sixty position estimates (51.7%) were to the right of target centre, which was not reliably different from chance of 50%, $\chi^2 = 0.02$, $p = .90$. 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 forty-nine position estimates (75.5%) were to the left of target centre, which was reliably higher than chance, $\chi^2 = 12.76$, $p < .001$. (One position estimate on the left/right dimension was at the precise horizontal centre of the target and was dropped from the analysis.) For target objects appearing in the upper half of the screen, there was a clear position bias lower. Nineteen of the twenty-three position estimates (82.6%) were below target centre, which was reliably greater

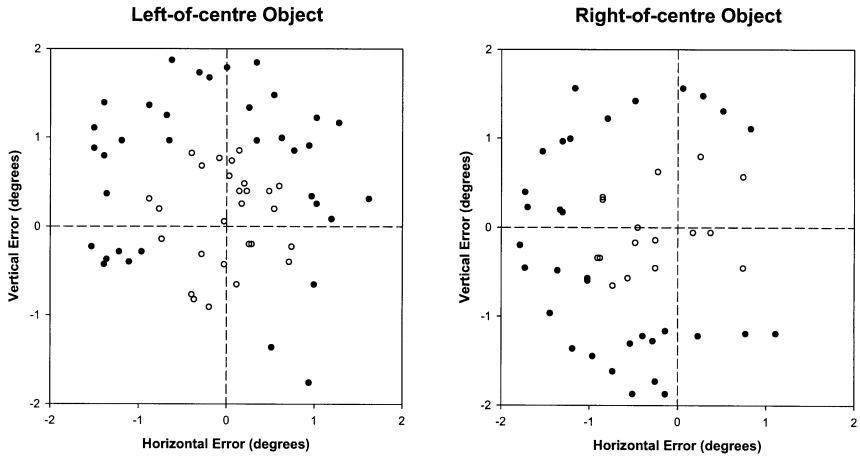


Figure 4. Individual position estimates in the target present preview condition for objects appearing on the left and right sides of the screen. Each estimate is plotted relative to the centre of the target object on that trial (the intersection of the dotted lines). Only estimates within 2° of target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

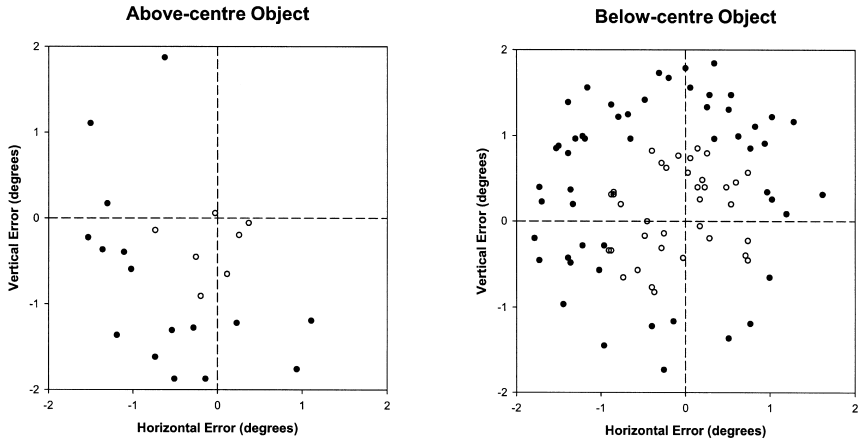


Figure 5. Individual position estimates in the target present preview condition for objects appearing on the top and bottom halves of the screen. Each estimate is plotted relative to the centre of the target object on that trial (the intersection of the dotted lines). Only estimates within 2° of target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

than chance, $\chi^2 = 9.78$, $p < .005$. Finally, for target objects appearing in the lower half of the screen, there was a clear position bias higher. Fifty-seven of the eighty-seven position estimates (65.5%) were above target centre, which was reliably greater than chance, $\chi^2 = 8.38$, $p < .005$. With the exception of target objects on the left-hand side of the screen, estimates were consistently biased toward the centre of the image.

The same trends were evident even when considering only estimates within 1° of target centre (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 centre (55.6%). For target object on the right-hand side of the screen, 11 of 16 estimates were to the left of target centre (68.8%). For target objects in the upper half of the screen, 6 of 7 estimates were below target centre (85.7%). For target objects in the lower half of the screen, 20 of 35 estimates were above target centre (57.1%).

This position bias toward the centre of the scene is consistent with studies examining spatial biases in memory for groups of objects in two-dimensional configurations (Hund & Plumert, 2002, 2003; Plumert & Hund, 2001). Plumert and Hund (2001) presented children and adults with a box divided into four quadrants. A set of five objects was placed in each quadrant in a random configuration. The objects were removed, and participants then attempted to place the objects back in the original positions. Plumert and Hund found that position estimates were consistently biased toward the centre of each quadrant. The present study shows that such spatial biases toward the centre of a geometric region generalize to adult position memory for objects in natural scenes. In addition, these biases were observed despite the fact that natural scenes contain significant spatial contextual information to constrain object position memory.

Finally, position memory biases toward scene centre may provide an explanation for the scene memory phenomenon of boundary extension. Intraub and colleagues (see Intraub, 1997 for a review) have 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. If, as suggested by the present results, participants remember the locations of objects as closer to scene centre than they actually were, one would expect precisely the boundary extension effect observed by Intraub: When drawing a scene from memory, participants should compress objects toward the centre of the drawing, resulting in a depiction that extends beyond the original scene boundaries.

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 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

Becker, M. W. & Pashler, H. (2002). Volatile visual representations: Failing to detect changes in recently processed information. *Psychonomic Bulletin and Review*, *9*, 744–750.

Castelhano, M. S., & Henderson, J. M. (this issue). Incidental visual memory for objects in scenes. *Visual Cognition*, *12*, 1017–1040.

Henderson, J. M., & Hollingworth, A. (1998). Eye movements during scene viewing: An overview. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 269–283). Oxford, UK: Elsevier.

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

Hollingworth, A. (2003a). Failures of retrieval and comparison constrain change detection in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *29*, 388–403.

Hollingworth, A. (2003b). *Visual memory and the online representation of complex scenes*. Paper presented at the Munich Visual Search symposium, Munich, Germany.

Hollingworth, A. (2004). Constructing visual representations of natural scenes: The roles of short- and long-term visual memory. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 519–537.

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

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

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

Hund, A. M., & Plumert, J. M. (2002). Delay-induced bias in children's memory for location. *Child Development*, *73*, 829–840.

Hund, A. M., & Plumert, J. M. (2003). Does information about what things are influence children's memory for where things are? *Developmental Psychology*, *39*, 939–948.

Huttenlocher, J., Newcombe, N., & Sandberg, E. H. (1994). The coding of spatial location in young children. *Cognitive Psychology*, *27*, 115–147.

Intraub, H. (1997). The representation of visual scenes. *Trends in Cognitive Sciences*, *1*, 217–222.

Irwin, D. E., & Andrews, R. (1996). Integration and accumulation of information across saccadic eye movements. In T. Inui & J. L. McClelland (Eds.), *Attention and performance XVI: Information integration in perception and communication* (pp. 125–155). Cambridge, MA: MIT Press.

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

Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003). Top down control of visual attention in object detection. In *IEEE proceedings of the international conference on Image Processing* (Vol. I, pp. 253–256).

O'Regan, J. K. (1992). Solving the ''real'' mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology*, *46*, 461–488.

Plumert, J. M., & Hund, A. M. (2001). The development of memory for location: What role to spatial prototypes play? *Child Development*, *72*, 370–384.

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

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

Scholl, B. J. (2000). Attenuated change blindness for exogenously attended items in a flicker paradigm. *Visual Cognition*, *7*, 377–396.

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

Simons, D. J. (1996). In sight, out of mind: When object representations fail. *Psychological Science*, *7*, 301–305.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*, 261–267.

Wolfe, J. M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting memories* (pp. 71–94). Cambridge, MA: MIT Press.