VISUAL COGNITION, 2006, 13 (1), 99–108

Psychology Press
Taylor & Francis Group

Using real-world scenes as contextual cues for search

James R. Brockmole and John M. Henderson

Department of Psychology and Cognitive Science Program, Michigan State University, East Lansing, MI, USA

Research on contextual cueing has demonstrated that with simple arrays of letters and shapes, search for a target increases in efficiency as associations between a search target and its surrounding visual context are learned. We investigated whether the visual context afforded by repeated exposure to real-world scenes can also guide attention when the relationship between the scene and a target position is arbitrary. Observers searched for and identified a target letter embedded in photographs of real-world scenes. Although search time within novel scenes was consistent across trials, search time within repeated scenes decreased across repetitions. Unlike previous demonstrations of contextual cueing, however, memory for scene–target covariation was explicit. In subsequent memory tests, observers recognized repeated contexts more often than those that were presented once and displayed superior recall of target position within the repeated scenes. In addition, repetition of inverted scenes, which made the scene more difficult to identify, produced a markedly reduced rate of learning, suggesting semantic information concerning object and scene identity are used to guide attention.

Philosophically, we never step into the same river twice, but psychologically, that river is very similar each time we encounter it. We recognize the surrounding landscape of trees, rocks, and embankments as being the same objects and features. The tyre swing is in the same spatial location relative to the tree that supports it, which in turn is in the same place relative to the boat-house. Even objects that move or can be translated in space appear in regular spatial arrangements; boats are in the water and clouds are in the air. That is, real-world environments are relatively stable collections of objects that covary with each other in known or predictable ways (Henderson & Hollingworth, 1999).

Please address all correspondence to: James R. Brockmole, now at Psychology Department, University of Edinburgh, Edinburgh, EH8 9JZ, UK. Email: James.Brockmole@ed.ac.uk

This research was supported by the National Science Foundation (BCS-0094433), the Army Research Office (W911NF-04-1-0078), and a Strategic Partnership Grant from the MSU Foundation. We thank Matt Peterson, Todd Kelley, and an anonymous reviewer for comments on a previous version of this manuscript. We also thank Devon Witherell for his help with data collection.

http://www.psypress.com/viscog DOI:10.1080/13506280500165188

One benefit of a stable scene structure is that memory for the configuration of objects can be used to help guide visual attention to behaviourally relevant targets (Chun, 2000). Because spatial relationships among objects in a scene are relatively constant, by virtue of knowing the locations of any set of objects, one also knows (or can at least predict) the location of a single target, thereby reducing or eliminating the need to execute a detailed serial search throughout the entire scene. For example, search for a picnic table is facilitated if we know from past experience that it is located on the river bank under the oak tree south of the boat-house. Thus, the context provided by a scene can cue the location of individual objects.

Chun and his colleagues have conducted a series of elegant experiments in support of this contextual cueing hypothesis (Chua & Chun, 2003; Chun & Jiang, 1998, 1999, 2003; Jiang & Chun, 2001; Olson & Chun, 2002). The general approach of these experiments has been to demonstrate that repeated exposure to a specific arrangement of target and distractor items leads to a progressively more efficient search for the target item. The stimulus arrays generally consisted of a rotated T (target) among a set of rotated Ls (distractors). Intermixed among randomly generated novel displays, a subset of stimuli were consistently repeated where the arrangement of the target and distractor elements was fixed. Thus, in these repeated stimuli, the global structure created by the elements in the display was entirely predictive of the target's location. Consistent with the notion that learned context can be used to guide visual attention, search times for repeated displays were faster than those for novel displays, an effect that increased in magnitude over repetitions. Strikingly, when observers were later asked to discriminate novel and repeated displays, they performed at chance, suggesting the memory mechanism involved in learning target-distractor contingencies was implicit (Chun & Jiang, 1998). Similar effects have also been observed in search tasks where the movement of targets and distractors was correlated, with collections of novel 2-D shapes that predicted target identity (Chun & Jiang, 1999), and with 3-D volumetric shapes (Chua & Chun, 2003).

Although Chun and colleagues acknowledge that the stimuli used in their experiments lack the realism of a natural scene (Chun, 2003), they argue that their stimulus arrays nevertheless contain the kind of structure available in real-world environments. In the light of this argument, our research had two goals. First, we aimed to demonstrate, for the first time, that observers are sensitive to context–target covariation when viewing real-world scenes, and that such sensitivity influences the deployment of attention (Experiment 1). Second, we sought to address new questions regarding contextual cueing specific to situations where real-world scenes serve as the learning context. Specifically, we considered a potential contribution of explicit scene memory (Experiment 1b) and semantic information (Experiment 2) in contextual cueing.

EXPERIMENT 1

Experiment 1 tested whether memory for real-world scenes can be used to guide the deployment of attention to search targets. Eight observers searched for a known target (a small grey ''T'' or ''L'') embedded within full-colour photographs of real-world scenes. These stimuli were displayed at a resolution of 600 × 800 pixels with 24-bit colour on a 17-inch CRT with a refresh rate of 100 Hz. Targets were presented in 9-point Arial font. The position of the target within each image was randomly determined. Across all stimuli, targets were uniformly distributed across the visual field.

Two types of trial were created. A novel trial presented a scene that had not been previously shown in the experiment. These trials measured baseline search speed. A repeated trial presented one of eight scenes[1] that were previously shown. Critically, the target's location in each repeated scene was fixed, although the target's identity was randomly selected with each repetition. At the beginning of each trial, a blue dot was centred on a grey background. Observers were instructed to look at this dot and to press a key when ready to view the scene. Upon identifying the target, observers pressed the one of two buttons corresponding to either ''T'' or ''L''. The trial was terminated if a response was not made within 20 s of scene onset.

The sequence of trials was divided into 17 blocks of 16 trials. Each block randomly intermixed each of the eight repeated trials with eight novel trials. Under these constraints, the order of trials was randomly selected for each subject. No information regarding the block structure or the repetition of scenes was given to observers; any learning of scene–target covariation was incidental.

Results and discussion

Trials were excluded from analysis if a response was not made within 20 s (3% of novel trials, < 1% of repeated trials), was incorrect (< 1% of novel and repeated trials), or was greater than 3 standard deviations from the mean as computed on a subject-by-subject basis (2% of novel trials[2]).

Results are illustrated in Figure 1. A 2 (novel vs. repeated) × 17 (block) repeated measures ANOVA demonstrated main effects of trial type, $F(1, 7) = 342$, $p < .001$, and block, $F(1, 7) = 3.56$, $p < .001$. Critically, these factors interacted, $F(16, 112) = 6.01$, $p < .001$. Considering novel trials only, despite a

[1] In a pilot experiment, all scenes were presented exactly once as observers searched for the target letter. The eight repeated scenes were randomly selected from those that fell within the interquartile range. This prevented the easiest and most difficult search scenes from being selected for repetition.

[2] Repeated trials were not subjected to the standard deviation trim because the variance of responses was expected to vary by block. Given only eight repeated trials per block an accurate estimation of the variance within each cell was not possible.

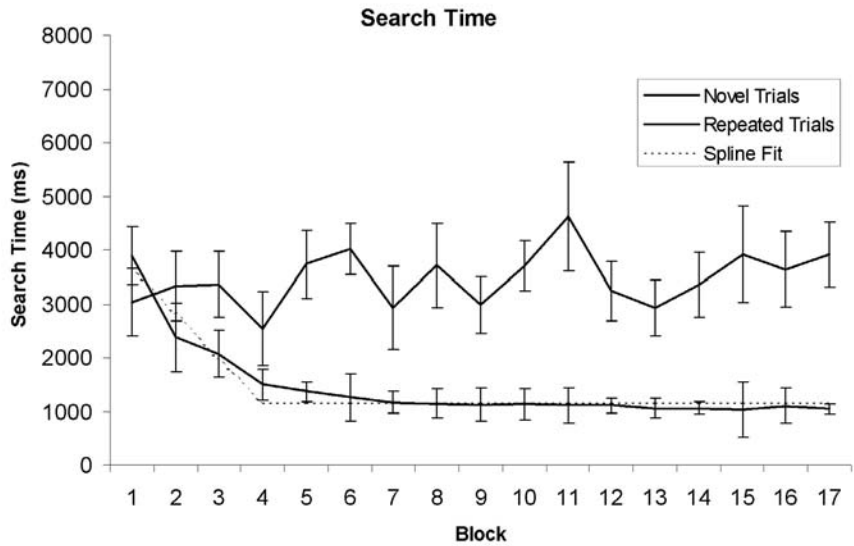

Figure 1. Search times as a function of trial type (novel or repeated) and block in Experiment 1. Error bars represent 95% confidence intervals.

marginal effect of block, $F(16, 112) = 1.68$, $p = .06$, single degree of freedom polynomial tests showed no reliable trends in search time, e.g., linear trend, $F(1, 7) = 2.36$, $p = .17$; quadratic trend, $F(1, 7) < 1$; cubic trend, $F(1, 7) = 1.12$, $p = .32$. On average, search time in novel scenes was 3388 ms. Search time in repeated trials, however, did vary with block, $F(16, 112) = 53.6$, $p < .001$, as described by linear, $F(1, 7) = 139$, $p < .001$, and quadratic trends, $F(1, 7) = 76.2$, $p < .001$.

To estimate the rate at which the context–target association was learned, a spline regression procedure fit two contiguous line segments to the search time data for repeated trials by varying the slope of the first segment and the joint between the lines. The slope of the second line was fixed at 0, reflecting asymptotic performance after a maximal benefit of repetition was achieved. The slope of the first function reflects the rate at which the context–target association was learned, the joint of these functions indicates the number of repetitions required for learning to be complete, and the intercept of the second function indicates the maximally efficient search speed. The best fitting function (multiple $R = .90$) is illustrated in Figure 1; search times decreased at a rate of 829 ms per block (i.e., per repetition) through block 4 after which an asymptote in search speed of 1152 ms was achieved.

Consistent with the contextual cueing hypothesis, search times within repeated scenes were faster than those within novel scenes and this difference was

magnified, to a point, over repetitions. Thus, observers are sensitive to covariation that exists between search targets and real-world scenes even when the search target is not predictable a priori from the meaning of the scene and when the relationship between the location of the target and the meaning of the scene is arbitrary. This learned covariation can be used to efficiently guide attention to task-relevant scene regions. It is intriguing, however, that the observed cueing effect is much larger than that found using simple stimulus arrays. For example, in Chun and Jiang's (1998) seminal paper on contextual cueing using rotated Ts and Ls, search time in the repeated displays continued to decrease through 15–20 repetitions, yielding a benefit of between 60 and 80 ms over novel displays. In the present study, however, only four repetitions were required to achieve the maximal cueing benefit to an advantage of over 2s. This informal comparison suggests that the strength of the contextual cueing effect is sensitive to the type of information provided in a stimulus.[3] We examined two such factors: Explicit memory (Experiment 1b) and semantic memory (Experiment 2).

EXPERIMENT 1B

Prior demonstrations of contextual cueing have shown that observers were unable to explicitly recognize repeated contexts more often than novel contexts, linking the phenomenon to implicit memory. Unlike memory for random arrays of letters, however, observers are able to explicitly recognize thousands of previously novel scene images after a single exposure to them (Shepard, 1967; Standing, 1973), raising the possibility that memory for scene–target associations could be explicitly, rather than implicitly, encoded. In addition, more variability exists across scenes than simple letter arrays, which could also increase one's ability to explicitly discriminate two scenes compared to two letter arrays.

To test the possibility that explicit memory plays a role in cueing when real-world scenes constitute the visual environment, following the completion of Experiment 1, the same observers' memory for the search scenes was tested. The tested scenes included the eight scenes that were repeated throughout the search task, eight scenes that were shown once in the search task, and eight scenes that were never shown during the search task. Observers classified each scene as ''old'' (included in the search task) or ''new'' (not included in the search task).

[3] The baseline search rate (as measured by novel trials) was longer in Experiment 1 than has been found in previous studies using nonscene stimuli, potentially enabling a greater contextual cueing effect. However, the magnitude of the difference in baseline search rates did not equal the magnitude of the difference in the cueing effect. While baseline search time in Experiment 1 was 3–4 times slower, the cueing effect was 20–25 times greater. Baseline difference cannot entirely account for the larger cueing effect observed here.

Search targets were removed from the old scenes, and for those that were reported as old, observers indicated where the target had been located. Using a mouse trackball, observers moved a blue dot equivalent in size to the target letter to the location they remembered the target occupying. The analyses of interest were whether observers displayed explicit recognition of repeated scenes and the associated target locations.

Results and discussion

Results are summarized in Table 1. Of the old trials, observers reported remembering 97% of the scenes that were repeated and 38% of the scenes that were presented once.[4] Nine percent of new scenes were remembered (i.e., false alarms). A one-way repeated measures ANOVA demonstrated that these memory rates reliably differed, $F(2, 14) = 95.3$, $p < .001$, with all pairwise comparisons reliable. The accuracy of observers' memory for target position was operationalized as the distance between the actual location of the target and the observer's placement of the target. The average placement error was 1.7 cm for repeated scenes and 9.3 cm for novel scenes, $t(7) = 4.13$, $p < .001$.

Old scenes were recognized at rates that exceeded false alarms, with scenes that were repeated during search virtually always recognized. Additionally, memory for target position was over five times more accurate for the repeated scenes compared to the once-viewed scenes. Thus, repeated exposure to scene stimuli led to explicit memory for the repeated scenes and the associated target location. Unlike previous demonstrations of contextual cueing in which obser-

TABLE 1
Mean performance (with standard deviation) on tests of memory for scenes and associated target locations

	Amount of exposure in search task		
Tasks	*Repeated*	*Viewed once*	*Never viewed*
Percentage of scenes recognized as ''old''	97% (9%)	38% (26%)	9% (9%)
Error in target localization	1.7 cm (0.52 cm)	9.3 cm (4.8 cm)	

[4] It is perhaps surprising that only 38% of the scenes that were presented once during learning were recognized. However, observers were searching for a target rather than trying to memorize the scenes, as they were in the original Shepard (1973) and Standing (1967) studies. As such, recognition rates in this study likely represent incidental memory rather than intentional memory for scenes.

vers were unable to distinguish repeated and novel letter arrays, the present results point to a role for explicit memory in contextual cueing in real-world scenes. Depending on the nature of the stimuli, contextual cueing may be driven by both implicit and explicit memory processes. Experiment 2 investigated whether semantic interpretability, another potential difference between letter arrays and real scenes, can also be recruited to aid in learning arbitrary scene–target contingencies.

EXPERIMENT 2

Unlike letter arrays, scenes convey semantic information. What role, if any, does meaning play in contextual cueing? If contextual cueing simply involves learning associations between consistently mapped visual features, then object/scene identity and meaning would be completely irrelevant. If identity and meaning can facilitate contextual cueing, then any manipulation that reduces an observer's ability to identify objects or scenes should likewise reduce the contextual cueing effect. To contrast these possibilities, eight new observers participated in a replication of Experiment 1 with inverted scenes. Inversion interferes with the extraction of meaning from scenes while not affecting the quality of the image. For example, inverted depictions of scenes are more difficult to recognize (Rock, 1974), produce less conceptual masking (Intraub, 1984), interfere with similarity ratings between stimuli (Klein, 1982), and eliminate detection advantages observed for changes that occur to objects of central interest in a scene (Shore & Klein, 2000).

Results and discussion

Trials were excluded if a response was not made within 20 s (6% of novel trials, < 1% of repeated trials), was incorrect (< 1% of novel and repeated trials), or was greater than 3 standard deviations from the mean as computed on a subject-by-subject basis (3% of novel trials).

Results are illustrated in Figure 2. In general, the pattern of results paralleled that in Experiment 1. A reliable interaction between trial type and block was observed in a two-way repeated measures ANOVA, $F(16, 112) = 3.27$, $p < .001$. Search time for novel trials did not vary as a function of block, $F(16, 112) < 1$, and averaged 3620 ms. A reliable effect of block was observed for the repeated trials, $F(16, 112) = 13.8$, $p < .001$, characterized by linear, $F(1, 7) = 226$, $p < .001$, and quadratic trends, $F(1, 7) = 28.2$, $p < .01$.

Critical differences, however, were observed between the results obtained using inverted scenes (Experiment 2) and those using upright scenes (Experiment 1). A mixed-model ANOVA crossing the within-subject factors of trial type and block, and the between subjects factor of scene type (normal vs. inverted) indicated a main effect of scene type, $F(1, 14) = 7.47$, $p < .05$, and an interaction between scene type and trial type, $F(1, 14) = 6.51$, $p < .05$ (no other effects

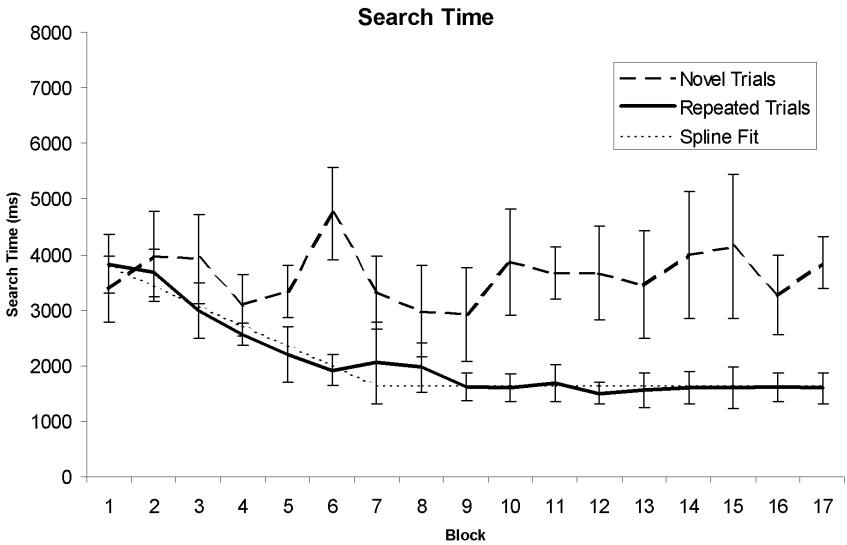

Figure 2. Search times as a function of trial type (novel or repeated) and block in Experiment 2. Error bars represent 95% confidence intervals.

involving scene type were reliable, all Fs < 1). This interaction was driven by differences in search time for repeated trials as search time for novel trials did not differ between the experiments, $F(1, 14) < 1$. To more completely characterize this interaction, a spline regression procedure was used to determine the learning rate for context–target associations in inverted scenes. The best fitting function (multiple $R = .73$) is depicted in Figure 2; search time decreased at a rate of 359 ms through block 7, attaining an asymptote of 1649 ms.

Maximal learning of the context–target associations within the inverted depictions of scenes required approximately twice the number of repetitions as the normal scene depictions. Given that the visual features in each version were identical, this result indicates that the ability to extract and retain semantic information speeds the acquisition of learning that underlies contextual cueing.

CONCLUSION

The amount of information available in real-world scenes dwarfs our cognitive processing abilities. For example, observers often fail to notice colour changes, object deletions, and object-token substitutions introduced to scenes from one view to the next (Simons & Levin, 2003), and only the last 3–4 attended items are stored in visual short-term memory (Irwin & Andrews, 1996; Luck & Vogel, 1997; McCarley, Wang, Kramer, Irwin, & Peterson, 2003; Peterson, Kramer, Wang, Irwin, & McCarley, 2001; Pylyshyn & Storm, 1988). Despite these

limitations, we rarely notice any processing difficulty because our attention system is extremely efficient at identifying and selecting goal-relevant scene regions.

Research using simple arrays of letters and shapes has demonstrated that the efficient detection of goal-relevant information is guided, at least in part, by memory for stable visual contexts that in turn predict the location of task-relevant visual information. For the first time, the experiments reported here expanded consideration of these contextual cueing effects to real-world scenes. Experiment 1 demonstrated that repeated exposure to photographs of real-world scenes led to a decrease in search time for a consistently but arbitrarily located target, indicating that observers learn associations between scenes and target locations. Compared to previous demonstrations of cueing in nonscene displays, however, scene–target associations were learned up to five times faster and led to a search time advantage twenty times greater. Experiments 1b and 2 demonstrated that scenes recruited at least two memory mechanisms not available to support the learning of context–target associations in nonscene stimuli. First, scene–target associations were explicitly encoded in memory. Observers recognized repeated scenes and recalled their associated target positions far better than for novel scenes. Second, cueing was facilitated by semantic memory for scene content. Approximately twice the number of repetitions were required to observe a maximal learning benefit when scenes were inverted (making them harder to interpret) compared to upright scenes. In summary, while the use of simple arrays of letters and shapes has increased our understanding of how memory for visual contexts develops and how it can in turn be used to guide behaviour, such paradigms have lead to only a partial characterization of such processes when it comes to action within the real world.

REFERENCES

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

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

Chun, M. M. (2003). Scene perception and memory. In D. E. Irwin & B. H. Ross (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 42, pp. 79–108). San Diego, CA: Academic Press.

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

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

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

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

Intraub, H. (1984). Conceptual masking: The effects of subsequent visual events on memory for pictures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 115–125.

Irwin, D. E., & Andrews, R. V. (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*. Cambridge, MA: MIT Press.

Jiang, Y., & Chun, M. M. (2001). The spatial gradient of visual masking by object substitution. *Vision Research*, *41*, 3121–3131.

Klein, R. M. (1982). Patterns of perceived similarity cannot be generalized from long to short exposure durations and vice versa. *Perception and Psychophysics*, *32*, 15–18.

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

McCarley, J. S., M. S. (2003). How much memory does oculomotor search have? *Psychological Science*, *14*, 422–426.

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

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

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*, 179–197.

Rock, I. (1974). The perception of disoriented figures. *Scientific American*, *230*, 78–85.

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

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

Simons, D. J., & Levin, D. T. (2003). What makes change blindness interesting? In D. E. Irwin & B. H. Ross (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 42, pp. 295–322). San Diego, CA: Academic Press.

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

Manuscript received January 2005
Manuscript accepted March 2005