Research Report

Psychological Science
21(11) 1551–1556

DOI: 10.1177/0956797610385359
http://pss.sagepub.com

Scene Memory Is More Detailed Than You Think: The Role of Categories in Visual Long-Term Memory

Talia Konkle[1], Timothy F. Brady[1], George A. Alvarez[2], and Aude Oliva[1]
[1]Massachusetts Institute of Technology and [2]Harvard University

Abstract

Observers can store thousands of object images in visual long-term memory with high fidelity, but the fidelity of scene representations in long-term memory is not known. Here, we probed scene-representation fidelity by varying the number of studied exemplars in different scene categories and testing memory using exemplar-level foils. Observers viewed thousands of scenes over 5.5 hr and then completed a series of forced-choice tests. Memory performance was high, even with up to 64 scenes from the same category in memory. Moreover, there was only a 2% decrease in accuracy for each doubling of the number of studied scene exemplars. Surprisingly, this degree of categorical interference was similar to the degree previously demonstrated for object memory. Thus, although scenes have often been defined as a superset of objects, our results suggest that scenes and objects may be entities at a similar level of abstraction in visual long-term memory.

Keywords

visual memory, scene categories, object categories, memory capacity

Received 4/5/10; Revision accepted 6/3/10

Humans have a remarkable capacity to remember pictures in episodic long-term memory, even after only a single exposure to the original image (Shepard, 1967; Standing, 1973; Standing, Conezio, & Haber, 1970). In one seminal study, Standing (1973) presented observers with 10,000 images, and they performed with more than 80% accuracy in a subsequent recognition memory task. However, because previous studies have used conceptually distinctive images, high performance may have been achieved with sparse representations of the images, in which only basic-level category information was stored (Chun, 2003; Simons & Levin, 1997; Wolfe, 1998; although see Standing et al., 1970). Thus, although these landmark studies demonstrate that observers can remember thousands of pictures, open questions remain about the fidelity of the stored scene representations and the infrastructure that supports them.

Recently, large-scale memory studies have demonstrated that observers can retain relatively detailed representations of individual objects. Hundreds to thousands of object representations can be maintained in visual long-term memory with enough fidelity to enable people to succeed at exemplar-level and even state-level discriminations, whether the objects are presented in isolation or embedded in scenes (Brady, Konkle, Alvarez, & Oliva, 2008; Hollingworth, 2004; Konkle, Brady, Alvarez, & Oliva, 2010). For example, we had observers view 2,800 object images, with 1 to 16 exemplars from each basic-level category (Konkle et al., 2010). Afterward, observers were presented with 2 exemplars from the same object category and indicated which one they had seen previously. Memory performance was high when a single exemplar was studied (89%), and there was only a 2% decrease in performance with each doubling of the number of exemplars studied within a category. Given these results with isolated object images, what level of performance might be expected from observers when they are presented with more complex images of real-world scenes and an increased number of scene exemplars?

Intuition suggests that long-term memory representations might be less detailed for individual scenes than for single objects. Scenes may contain more information than individual objects do (Marr, 1982) because scenes contain multiple objects that are shared between exemplars of various scene

Corresponding Authors:
Talia Konkle, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139
E-mail: tkonkle@mit.edu

Aude Oliva, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139
E-mail: oliva@mit.edu

categories. Consider the golf courses or ocean waves in Figure 1, for instance; all of the scene exemplars within each of these categories are conceptually similar and contain the same objects, regions, and color scheme. This similarity could lead to a high degree of interference in visual long-term memory: In others words, recognition performance might decline rapidly as more scene exemplars are studied.

Alternatively, holistic theories of scene representation (Greene & Oliva, 2009b; Oliva & Torralba, 2001) have shown that scenes do not need to be represented as a collection of parts: Global properties are sufficient to enable recognition. Furthermore, scenes have category structure just as objects do (Tversky & Hemenway, 1983), and such conceptual information has been shown to support memory for details (Konkle et al., 2010; Koutstaal et al., 2003; Wiseman & Neisser, 1974). For example, basic-level category schemas may allow for efficient encoding of scenes through either compressive or expanded encoding (Brady, Konkle, & Alvarez, 2009; Schyns, Goldstone, & Thibaut, 1998), or by guiding attention toward distinctive details (Eysenck, 1979; Nosofsky, 1986; Vogt & Magnussen, 2007). Holistic processing of scenes might therefore result in high-fidelity memory representations, leading to only minor interference as the number of studied exemplars increases.

To examine the fidelity of scene representation in visual long-term memory, we conducted a large-scale memory experiment in which observers studied thousands of scene images. We varied the number of exemplars presented per scene category and tested memory using exemplar-level foils.

Method

Participants

Twenty-four adults (age range = 20–35 years) gave informed consent and received compensation for their participation.

Stimuli

The stimuli were 4,672 images from 160 different scene categories gathered using Google Image Search. Stimuli can be downloaded at http://cvcl.mit.edu/MM.

Procedure

The experiment consisted of a study phase and a testing phase, both conducted on a computer. During the study phase, observers viewed thousands of scenes and also performed a repeat detection task that encouraged sustained attention. The study phase was divided into 10 blocks of 20 min each. After a 20-min break, observers completed the testing phase, in which a series of two-alternative, forced-choice test trials was presented. Observers had to discriminate the images they had seen in the study phase from foil images that could either be from a novel scene category or be an exemplar from the same scene category. Before the experiment began, observers were informed about what kinds of test to expect during the testing phase.

Study phase. In the study phase, observers viewed 2,912 images from 128 different scene categories, with 1, 4, 16, or

Fig. 1. Four example scene categories (ocean wave, classroom, golf course, amusement park) from the set of 160 categories used in the experiment.

64 exemplars presented per category. All items that would later be tested were distributed uniformly throughout the stream. Images were presented one at a time (subtending 7.5° × 7.5° visual angle) for 3 s each, alternated with a fixation cross (800 ms; see Fig. 2a).

For the repeat detection task, observers were told to press the space bar any time an image repeated, and they were not informed of the structure or frequency of the repeat conditions. Feedback was given only when participants responded, with the fixation cross turning red if they incorrectly pressed the space bar (false alarm) or green if they correctly detected a repeat (hit). Repeated images were distributed uniformly throughout the study stream and were set to recur after 1, 15, 63, 255, or 1,023 intervening items. These images came from 48 scene categories, in which either 4, 16, or 64 exemplars were presented and 4 exemplars per category were repeated (there were 192 repeats total). Thus, in the study stream, repeats occurred approximately every 1 in 14 items. None of the scene categories with repeated items in the study stream were used in the testing phase.

Testing phase. The testing phase consisted of 224 two-alternative, forced choice trials. Participants proceeded at their own pace and were told to emphasize accuracy, not speed, in making their judgments. On each trial, participants were presented with two items—one previously studied (old) item and one new foil item—and they indicated which item they had seen before (see Fig. 2b). In the exemplar-test conditions, the foil and the studied items were from the same scene category. For all exemplar tests, we varied how many other scene exemplars from that category had been viewed in the study phase: 4, 16, or 64 other scene exemplars. Sixteen scene categories were used for each of these studied-exemplar conditions, with 4 tested images per category, yielding 64 trials per condition.

We also included a novel test condition, in which the studied item and foil item were from different scene categories. In these 32 test trials, the studied item was always a singleton exemplar from the study stream, and the foil item was always a singleton exemplar from a novel scene category.

Critically, all of the two-alternative, forced-choice test trials were the same for all subjects; in these trials, we counterbalanced (a) which one of the two items was studied and which was the foil and (b) how many exemplars each participant studied from a particular category. This complete counterbalancing ensured that any changes in memory performance across conditions could be attributed to a pure impact of additional studied exemplars, and could not be driven by differences in difficulty across scene categories or across specific test-foil comparisons.

Results and Discussion

Forced-choice memory performance

Performance in the forced-choice memory task is plotted in Figure 2c. When the foil was a scene from a novel category, observers correctly identified 96% (*SEM* = 1%) of the studied images. Thus, even with approximately 3,000 scenes in memory, observers could distinguish which item they had studied,

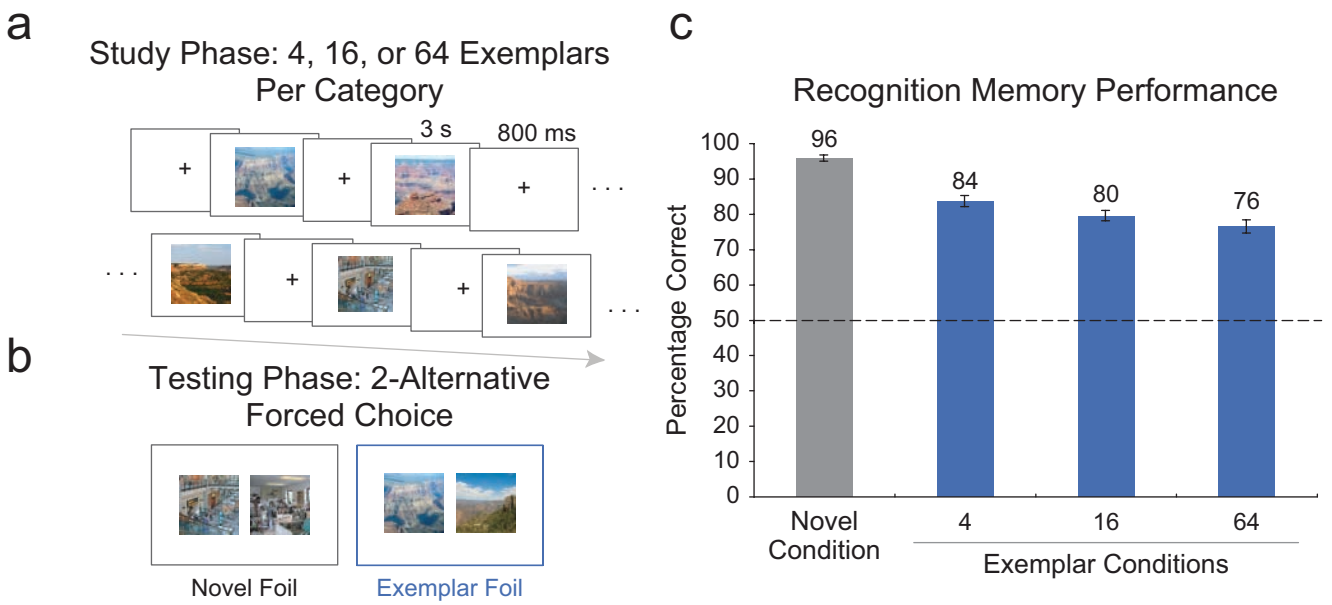

Fig. 2. Experimental procedure and results. During the study phase (a), observers were presented with 2,912 scenes for 3 s each, with a variable number of exemplars presented from each scene category. Each image was followed by a fixation cross (800 ms). During the testing phase (b), memory was tested with two-alternative, forced-choice tests. In the novel condition, one studied scene was presented alongside a new scene from a novel category. In the exemplar conditions, a studied scene was presented alongside a new exemplar from the same scene category. Performance on the recognition memory task was quantified as percentage correct (mean hit rate) and is plotted (c) for the novel condition and for the exemplar conditions, in which the number of studied exemplars was 4, 16, or 64. Chance performance is indicated by the dashed line. Error bars represent ±1 *SEM*.

with performance near ceiling. This finding is compatible with previous results demonstrating a massive visual long-term memory capacity for scenes when the foils are categorically distinct (e.g., Standing, 1973).

We next examined memory performance with exemplar-level foils. There was a significant drop in performance when memory for a scene was tested against a foil from a different category, compared with when it was tested against a foil from the same category, $t(23) = 10.0$, $p < .001$ (see also Brady et al., 2008). Further, increasing the number of studied exemplars from 4 to 64 reliably reduced performance from 84% to 76%, $F(2, 46) = 14.7$, $p < .001$, $\eta^2 = .39$. However, each doubling of the number of studied exemplars decreased forced-choice performance by only 1.8% (*SEM* = 0.3%). This result demonstrates that the information retained when observers remember a particular image is sufficient to distinguish that image from many other exemplars of the same category in memory. This suggests that a significant amount of detail in addition to the category is stored.

Repeat detection performance

Repeat detection performance was high, with hit rates at 96% with 1 intervening item (*SEM* = 2%), 91% with 15 intervening items (*SEM* = 2%), 80% with 63 intervening items (*SEM* = 3%), 74% with 255 intervening items (*SEM* = 3%), and 56% with 1,023 intervening items (*SEM* = 4%). The false alarm rate was near floor at 3% (*SEM* = 0.4%). Repeat detection sensitivity decreased systematically with an increase in intervening items, $F(4, 92) = 104.1$, $p < .001$, $\eta^2 = .82$ (Fig. 3b).

Comparison Between Objects and Scenes

The results demonstrate that memory representations of scenes are of relatively high fidelity, as observers are capable of distinguishing between scene exemplars with only a modest impairment in performance as the number of studied exemplars increases. A similar large-scale memory study with object exemplars (Konkle et al., 2010) provides an opportunity to compare the impact that the number of studied exemplars has on remembering visual stimuli of different complexity. It is important to note that the two studies used a similar procedure to select images from the basic-level categories: Images spanned the category, and the target the target and the foil were randomly selected from the images within a category, such that natural variability between exemplars within a category determined the relationship between the tested image and the foil image. If categories are an important structure supporting visual long-term memory, similar results might be expected for objects and scenes, despite the obvious differences between these two stimulus types (e.g., scenes contain many objects).

The comparison between the object and scene memory studies is shown in Figure 3 (also see the Supplemental Material available online). Although performance in the novel foil condition was significantly higher for scenes than for objects (scenes: 96%; objects: 93%), $t(40) = 2.07$, $p < .05$, the slope of interference by number of exemplars is strikingly similar for the two stimulus types. In both cases, the impact of the number of studied exemplars on memory performance was well fit by a linear relationship between memory performance and the

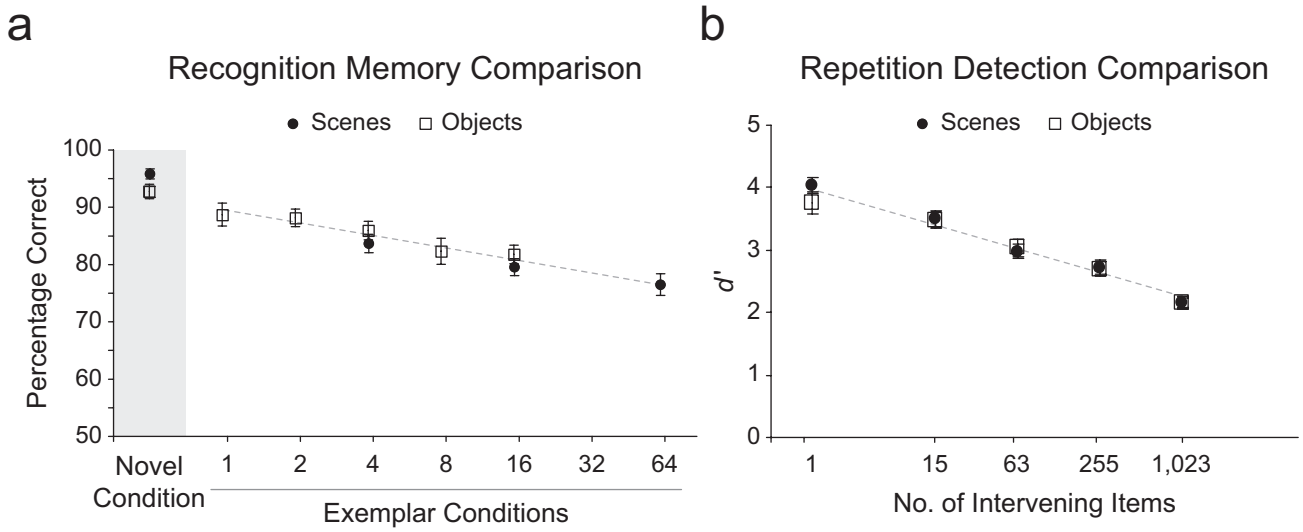

Fig. 3. Comparison of categorical interference in visual long-term memory for scenes (experiment reported here) and for objects (data from Konkle, Brady, Alvarez, & Oliva, 2010). Percentage correct (mean hit rate) in the two-alternative, forced-choice memory tasks of both experiments is plotted in (a). In the novel condition, the foil item was from a different basic-level category. In the exemplar conditions, the foil item was from the same basic-level category from which a variable number of exemplars had been viewed in the study stream. Repeat detection performance during the study phase is shown for both experiments in (b). Performance was quantified using *d′*, which takes into account hits, corrected for false alarms, and is plotted as a function of the number of intervening items between initial and repeat presentation. Regression lines are shown. Error bars represent ±1 *SEM*.

$\log_2$ number of studied exemplars. There was no reliable difference between the slope or the intercepts between these two studies—slope: $t(40) = 0.4$, $p = .71$; intercept: $t(40) = 0.8$, $p = .40$ (Fig. 3a). For scenes, the average slope of memory performance was –1.8% (*SEM* = 0.3%), with an intercept of 87% (*SEM* = 1.8%). For objects, the slope of memory performance was –2.0% (*SEM* = 0.4%), with an intercept of 89.2% (*SEM* = 1.8%). Performance in the repeat detection task was also very similar across object and scene stimuli (Fig. 3b)—slope: $t(40) = 1.7$, $p = .09$; intercept: $t(40) = 1.1$, $p = .26$.

General Discussion

We examined the fidelity of memory for scene exemplars in a large-scale memory experiment by using exemplar-level tests and varying the number of exemplars per category. We found that observers presented with two exemplars were able to successfully choose which one they had previously studied, even after encoding thousands of scene images and up to 64 exemplars from that category. Although it has typically been assumed that in previous large-scale memory experiments the stored scene representations were sparse (e.g., Chun, 2003; Simons & Levin, 1997; Wolfe, 1998), the current data demonstrate that memory representations for scenes can contain significant detail beyond the scenes' basic-level category (see also Standing et al., 1970).

Additionally, we observed that memory performance showed a minor but systematic decline as more exemplars were studied from a scene category. This degree of interference from scene exemplars was remarkably similar to the interference caused by additional object exemplars—approximately a 2% decline in memory performance with each doubling of the number of studied exemplars. Moreover, the repeat detection results show that the impact of intervening items (or elapsed time) is similar for scenes and objects, and this further supports the generality of long-term memory processes across different kinds of visual content.

Although the minimal impact of exemplars may suggest that semantic structure is largely irrelevant, previous work has shown that abstract images that do not relate to preexisting knowledge are remembered very poorly (Koutstaal et al., 2003; Wiseman & Neisser, 1974). Furthermore, in this study, we observed an 8% impairment in performance by adding 60 additional exemplars within the same category to the study stream. By contrast, in the work of Standing (1973), an 8% drop in performance resulted from the addition of nearly 7,500 categorically distinct items. Thus, visual categories may actually provide the conceptual scaffolding for supporting detailed visual long-term memory for scenes, just as they do for objects (Konkle et al., 2010).

The fidelity of memory for scenes

Although recognition memory was high, these results do not imply that visual long-term memory retains a photographic representation of scenes. Rather, our results suggest that memory representations contain the kind of details that would be most likely to distinguish between meaningful exemplar-level changes (e.g., your bedroom compared with someone else's bedroom). It is possible that preexisting knowledge of scene categories, such as bedrooms and beaches, helps support high-fidelity long-term memory for any particular scene exemplar.

According to this account, a first glance at a scene gives the observer a statistical summary representation of the global image structure and its basic-level category (Greene & Oliva, 2009a; Oliva, 2005). This category knowledge helps guide attention to details that are the most relevant for distinguishing this scene from other scenes. Shifting attention to different aspects of the scene over time allows for the accumulation of visual details (Brady, Konkle, Oliva, & Alvarez, 2009; Hollingworth, 2004, 2008; Melcher, 2006). Thus, given sufficient preexisting knowledge and time, diagnostic visual details are accumulated and stored in long-term memory (Brady et al., 2008; Brady, Konkle, Oliva, & Alvarez, 2009; Vogt & Magnussen, 2007).

Relationship between scene and object categories

We found that the impact of scene exemplars on memory performance and the impact of object exemplars on memory performance were of a similar magnitude. How should we account for this, given the obvious differences between real-world scenes and isolated individual objects? In both the object and the scene experiments, we sampled exemplars that spanned the breadth of their category. However, this sampling procedure does not by itself predict that the interference of object exemplars and scene exemplars should be the same. To make this prediction, we also have to assume that the infrastructure of long-term memory is such that any two object exemplars and any two scene exemplars are, on average, equally likely to interfere with each other. Thus, these results potentially make a strong statement about the general structure of visual categories—namely, that the richness of scene categories is similar to the richness of object categories. This proposal fits with the notion that basic-level categories are situated at an information-theoretic optimum, maximizing within-category similarity and minimizing between-category similarity (Corter & Gluck, 1992; Rosch, 1978).

Conclusion

Beyond the capacity and fidelity of long-term memory, our results speak to a fundamental principle of the representation of visual events. Scene and object representations have largely been considered separate visual entities: Scenes are often thought to be a superset of objects. However, the data not only from the current study but also from studies as diverse as those involving recognition at a glance (i.e., Greene & Oliva, 2009a; Potter, 1976; Thorpe, Fize, & Marlot, 1996) and categorical statistical learning (Brady & Oliva, 2008) suggest a strong

alternative: Scene and object categories may be best treated as entities at a similar level of conceptual abstraction, providing the semantic structure necessary to support recognition and memory of visual details.

Acknowledgments

T.K. and T.F.B. contributed equally to this work.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This work was partly supported by the National Science Foundation (Career Award IIS- 0546262) and by a Google Research Award to A.O.

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

References

Brady, T.F., Konkle, T., & Alvarez, G.A. (2009). Compression in visual short-term memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*, 487–502.

Brady, T.F., Konkle, T., Alvarez, G.A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, USA*, *105*, 14325–14329.

Brady, T.F., Konkle, T., Oliva, A., & Alvarez, G.A. (2009). Detecting changes in real-world objects: The relationship between visual long-term memory and change blindness. *Communicative & Integrative Biology*, *2*, 1–3.

Brady, T.F., & Oliva, A. (2008). Statistical learning using real-world scenes: Extracting categorical regularities without conscious intent. *Psychological Science*, *19*, 678–685.

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

Corter, J.E., & Gluck, M.A. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, *111*, 291–303.

Eysenck, M.W. (1979). Depth, elaboration, and distinctiveness. In L.S. Cermak & F.I.M. Craik (Eds.), *Levels of processing in human memory* (pp. 89–118). Hillsdale, NJ: Erlbaum.

Greene, M.R., & Oliva, A. (2009a). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*, 464–472.

Greene, M.R., & Oliva, A. (2009b). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*, 137–176.

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. (2008). Visual memory for natural scenes. In S.J. Luck & A. Hollingworth (Eds.), *Visual memory* (pp. 123–162). New York, NY: Oxford University Press.

Konkle, T., Brady, T.F., Alvarez, G.A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*, 558–578.

Koutstaal, W., Reddy, C., Jackson, E.M., Prince, S., Cendan, D.L., & Schacter, D.L. (2003). False recognition of abstract versus common objects in older and younger adults: Testing the semantic categorization account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 499–510.

Marr, D. (1982). *Vision*. New York, NY: W.H. Freeman.

Melcher, D. (2006). Accumulation and persistence of memory for natural scenes. *Journal of Vision*, *6*(1), Article 2. Retrieved August 23, 2010, from http://www.journalofvision.org/content/6/1/2

Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Oliva, A. (2005). Gist of the scene. In L. Itti, G. Rees, & J.K. Tsotsos (Eds.), *The encyclopedia of neurobiology of attention* (pp. 251–256). San Diego, CA: Elsevier.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*, 145–175.

Potter, M.C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 509–522.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B.L. Lloyd (Eds.), *Cognition and categorization* (pp. 189–206). Hillsdale, NJ: Erlbaum.

Schyns, P.G., Goldstone, R.L., & Thibaut, J.P. (1998). The development of features in object concepts. *Behavioral & Brain Sciences*, *21*, 1–17.

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

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

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

Standing, L., Conezio, J., & Haber, R.N. (1970). Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science*, *19*, 73–74.

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

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology*, *15*, 121–149.

Vogt, S., & Magnussen, S. (2007). Long-term memory for 400 pictures on a common theme. *Experimental Psychology*, *54*, 298–303.

Wiseman, S., & Neisser, U. (1974). Perceptual organization as a determinant of visual recognition memory. *The American Journal of Psychology*, *87*, 675–681.

Wolfe, J.M. (1998). Visual memory: What do you know about what you saw? *Current Biology*, *8*, R303–R304.