Memory & Cognition
2009, 37 (1), 90-99
doi:10.3758/MC.37.1.90

Learning scenes from multiple views: Novel views can be recognized more efficiently than learned views

David Waller
Miami University, Oxford, Ohio

Alinda Friedman
University of Alberta, Edmonton, Alberta, Canada

and

Eric Hodgson and Nathan Greenauer
Miami University, Oxford, Ohio

In two experiments, participants were trained to recognize a playground scene from four vantage points and were subsequently asked to recognize the playground from a novel perspective between the four learned viewing perspectives, as well as from the trained perspectives. In both experiments, people recognized the novel view more efficiently than those that they had recently used in order to learn the scene. Additionally, in Experiment 2, participants who viewed a novel stimulus on their very first test trial correctly recognized it more quickly (and also tended to recognize it more accurately) than did participants whose first test trial was a familiar view of the scene. These findings call into question the idea that scenes are recognized by comparing them with single previous experiences, and support a growing body of literature on the existence of psychological mechanisms that combine spatial information from multiple views of a scene.

People are able to recognize previously learned places from perspectives that they have not experienced. For example, after having approached a favorite picnic spot in a park several times from the north and from the east, a person will likely recognize this spot as he or she approaches it from the northeast. The psychological processes that enable this sort of place recognition are currently not completely understood, although two classes of models have been proposed. By one account (which we will call a *normalization approach*), people store a relatively large number of specific examples of their experiences. Recognition from a novel perspective can then occur by matching the current view to a particular stored view (see, e.g., Diwadkar & McNamara, 1997; Tarr & Pinker, 1989). Empirically, normalization processes are indicated when recognition performance declines monotonically with the distance between a given novel view and its nearest learned view. Another class of models that represents a *view combination* (or *view interpolation*) *approach* holds that people do not rely on single instances of their prior experience; rather, novel views of a scene activate multiple stored views (e.g., Bülthoff & Edelman, 1992; Edelman, 1999; Hintzman, 1986; Ullman, 1998). By this approach, recognition is based on the summed activation (modeled proportionately to similarity) of the novel view to the stored views. As a result, some novel views can be recognized at least as well as familiar views.

Friedman and Waller (2008) provided initial evidence for view combination in scene recognition by exposing participants to views of a playground scene that had been taken from two ground-level perspectives (e.g., 48º apart). Subsequent recognition of these trained views during a test phase was not statistically different from that of novel views of the playground taken from an *interpolated* viewpoint—one that was between the two trained perspectives. However, novel *extrapolated* views—those that were outside of the training range—were subsequently recognized less efficiently than were the trained views. Because interpolated views were recognized more efficiently than extrapolated views, yet both were equidistant from the training views, Friedman and Waller concluded that their findings were not consistent with a normalization account of scene recognition. Instead, they concluded that the recognition of coherent real-world scenes was aided by view combination mechanisms that integrated information from the separate trained views.

In the present article, we provide strong additional evidence for view combination mechanisms in human spatial memory by showing that under appropriate circumstances, novel views of a scene can be recognized even more efficiently than can views that have been previously seen and

D. Waller, wallerda@muohio.edu

learned. Such a finding is not predicted by normalization accounts of scene recognition; however, it is possible, in principle, according to most models of view combination (Edelman, 1999; Edelman & Bülthoff, 1992). For example, Edelman (1999) and his colleagues (Edelman, Bülthoff, & Bülthoff, 1999) developed a view combination model for object recognition that we found to be applicable to the recognition of complex scenes (Friedman & Waller, 2008). According to Edelman's model, shape prototypes are mentally represented in a multidimensional "shape space," wherein the distance between shapes is proportional to their structural similarity along the multiple dimensions. When a novel view of a familiar object is presented, all of the stored prototypes that are sufficiently similar in structure to the new stimulus are activated to construct a new "view." The constructed view is then compared with the novel percept. If this constructed view closely matches the structure of the input view (i.e., the match exceeds a threshold of similarity), then the novel view is relatively easy to recognize. Because it is possible that a given novel view may activate several stored prototypes, and because the activation from all of the prototypes is summed, it is, in principle, possible for a novel view of an object to be easier to recognize than views that initially formed the shape space.

In scene recognition, the use of view combination may depend on either the number of experienced views or the degree of similarity in the overlapping information that is available from them, or both (Friedman, Spetch, & Ferrey, 2005; Friedman & Waller, 2008). For example, experiencing multiple views of a scene during learning may allow people to refine the precision of their memory of the angular relations among the objects in the scene. Similarly, learned views of a scene in which the information about the spatial relations among the objects overlaps or is partially redundant should serve to reinforce the salience of those relations. If information from multiple learned views is combined in scene recognition, it is possible that when a novel stimulus that is very similar to a combination of several learned views is presented during testing, it readily activates these stored representations and is thus recognized very efficiently.

In the present article, we examined the possibility that learning a scene from more than the two training views used by Friedman and Waller (2008) could lead to stronger view combination effects than those that they found (i.e., superior—not equivalent—recognition of an unlearned view, in comparison with the learned views). We used four training views of a playground scene that were arrayed around an untrained central view by either 15º (Experiment 1) or 30º (Experiment 2). The training views were created by moving viewpoint locations along an imaginary sphere centered on the playground, differing from the central view in azimuth or elevation (see Figure 1). We designed Experiment 1 to provide evidence that training with these views would result in superior recognition of the untrained interpolated (i.e., central) view, in comparison with recognition of the trained views. The untrained central view depicted the playground from an elevation of 45º, and because it was within the range of views spanned by the training views, we refer to it as the interpolated view.

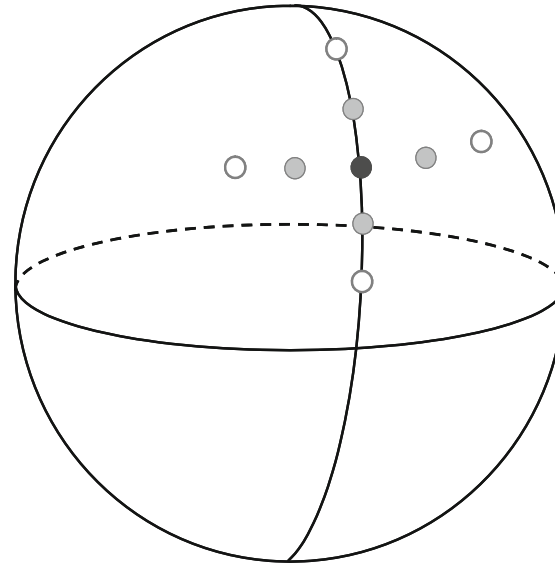

Figure 1. A schematic depiction of the viewing perspectives used in these experiments. Participants viewed an arrangement of playground items (not depicted) at the center of an imaginary sphere from perspectives that varied in azimuth (from −30º, −15º, 0º, 15º, and 30º, arranged left to right in this figure) or elevation (from 15º, 30º, 45º, 60º, and 75º, arranged bottom-to-top in this figure). In Experiment 1, participants trained on perspectives from the four inner viewpoints (gray circles) and were subsequently tested on these, as well as on the interpolated view (black circle) and extrapolated views (white circles). In Experiment 2, participants trained on the previous extrapolated views (white circles) and were tested on these, as well as on the other perspectives that varied in azimuth (arranged left to right).

Experiment 2 began to examine the dynamics of learning by observing the influence of the testing stimuli on recognition performance. This is theoretically important, because testing in these (and similar) experiments typically involves exposure to multiple novel views. Because there may be less variability among some of the novel views than among the trained views (e.g., there may be several training views but only one novel interpolated view), it is possible that averaged across trials, performance on novel test stimuli enables one to measure participants' sensitivity to these types of statistical regularities among their features. For example, Palmeri and Flanery (1999) exposed participants to test stimuli from a previous (unrelated) experiment (Knowlton & Squire, 1993)—dot patterns that had been randomly distorted from an untrained prototype and that had been "old" or "new" items during the test trials of the original experiment. Palmeri and Flanery's participants had not been explicitly trained to classify the patterns, but despite their lack of training, they were able to do so relatively accurately (see also Zaki & Nosofsky, 2007). We address this issue of learning during testing by examining how participants' performance changes across trials in the test phase in Experiments 1 and 2. Furthermore, in Experiment 2, we specifically examined whether a recognition advantage for some novel views occurs on the very first test trial, before participants have had a chance to learn from the set of test items.

EXPERIMENT 1

In the training phase of Experiment 1, participants identified a playground from four different perspectives that surrounded a central untrained (interpolated) view. They were then immediately asked to recognize these trained views, the untrained interpolated view, and additional novel views. If participants recognized the untrained interpolated view more efficiently than they did the trained views, this would provide strong new evidence for view combination processes in scene recognition.

Method

Participants. Sixteen undergraduate students (8 men and 8 women) from Miami University participated in the experiment in return for credit in their introductory psychology course. The mean age of the participants was 18.7 years ($SD = 0.6$).

Materials. The stimuli were color digital renderings of a playground scene consisting of 10 objects randomly arrayed on a flat, grassy field. Grayscale versions of selected stimuli appear in Figure 2. The stimuli were generated from a 3-D computer model using 3-D Studio Max, and we employed lighting effects, shadows, and textures to enhance the realism of the scene.

The training and test stimuli depicted this scene from nine different viewing locations (see Figure 1), all of which were equidistant from the center of the playground. All of the viewing perspectives were above ground level and were oriented toward the exact center of the scene. One of the nine viewing perspectives was the interpolated perspective, and it had an angular elevation above the ground plane of 45º and an arbitrarily assigned azimuth of 0º. Four additional perspectives (training views) were positioned around this interpolated perspective: two at the same azimuth, with elevations of ±15º relative to it (i.e., at 60º and 30º elevations), and two at the same elevation as that of the interpolated perspective and azimuths of ±15º. Finally, four more viewing perspectives (extrapolated views) were positioned around the interpolated perspective: two at the same azimuth as that of the interpolated perspective and elevations of ±30º relative to it (i.e., at 75º and 15º elevations), and two at the same elevation and azimuths of ±30º. The extrapolated perspectives provided novel test views that were each at the same angular distance from one of the training views as the interpolated view was from all of them.

We created either six (for training and extrapolated views) or nine (for the interpolated view) stimulus scenes from each of the viewing perspectives. For each perspective, one of the stimuli was the target and portrayed the correct, to-be-learned arrangement of the playground objects. The additional stimuli were distractors, which portrayed the scene with the positions of two of the playground's five foreground objects (the swing, the merry-go-round, the train, the seesaw, and the jungle gym) switched (see Figure 3). The interpolated view had more distractors created for it than did the other views, because during the test phase, the interpolated view was displayed more frequently than any particular training or extrapolated view. Different objects were switched for the different viewing perspectives (training, interpolated, and extrapolated), and different distractors were used during training and testing. The latter consideration is important because it means that during testing, participants will have never seen either the interpolated, extrapolated, or distractor stimuli; they will only have seen the four training stimuli.

Presentation of the stimuli and collection of the participants' responses were controlled through a computer using E-Prime software from Psychological Software Tools (Pittsburgh, PA). Stimuli were presented on a 32.5 cm × 24 cm CRT monitor (85-Hz refresh rate). Participants responded by pressing buttons on a response box connected to the serial port of the computer.

Procedure. Participants were run individually through the experiment. After being given a brief description of the experiment, the participant sat at the computer and read detailed task instructions. These instructions informed participants that they would be viewing many different arrangements of playground objects and that one par-

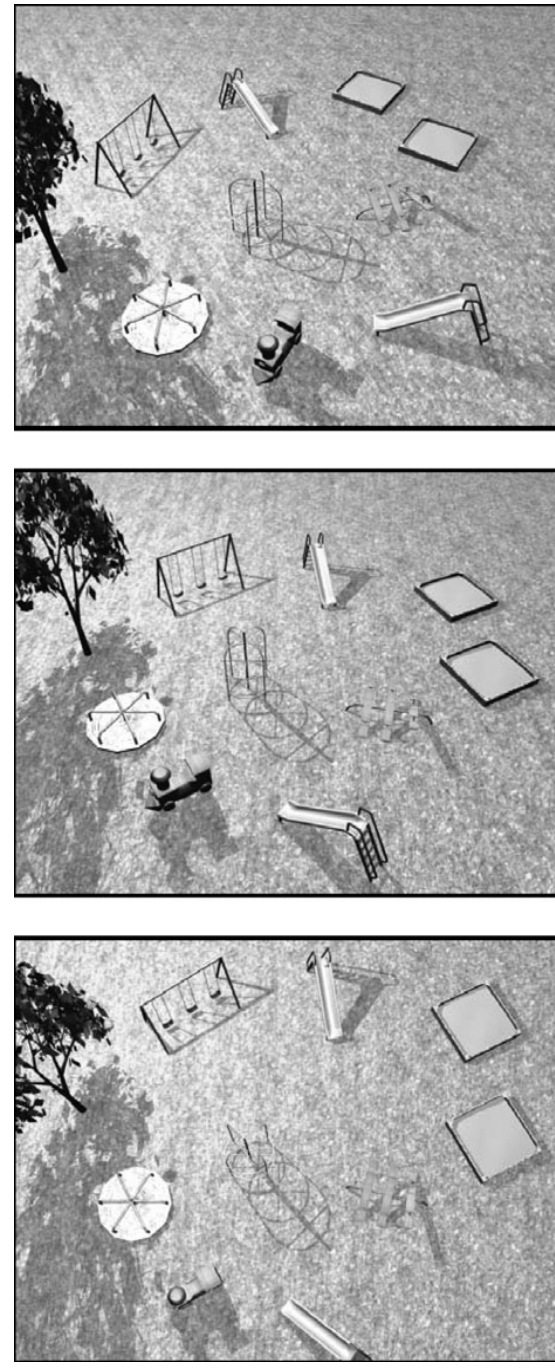

Figure 2. Sample target stimuli (grayscale versions) used in Experiments 1 and 2. Top: The view from −30º azimuth (leftmost white circle in Figure 1). Center: The central interpolated view (black circle in Figure 1). Bottom: The view from a 75º elevation (topmost white circle in Figure 1).

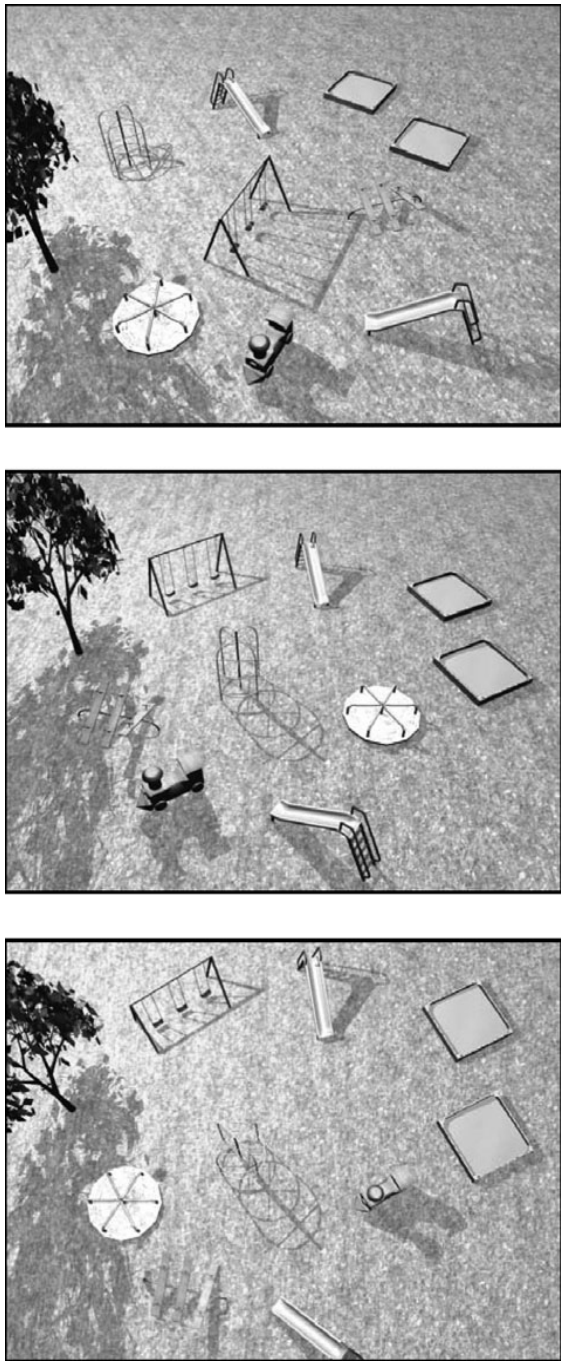

Figure 3. One sample distractor stimulus for each of the three target stimuli depicted in Figure 2. Distractors switched the location of two objects in the scene. For example, in the top panel, the jungle gym and the swing set were switched. Note that the distractor stimuli never switched the sandboxes, the tree, or the slides.

ticular arrangement was "correct." They were instructed to press a green button labeled "Correct" if the arrangement was correct, and a red button labeled "Incorrect" otherwise. Participants were also told that a randomized half of the pictures depicted the correct arrangement and that the other half of the pictures were incorrect.

During training, participants were shown the target and distractor stimuli from only the four training perspectives, and they were required to distinguish the four target stimuli from 12 different distractors (3 for each trained viewpoint). Participants received feedback over headphones during training. The feedback message said "three points" if they were correct and answered in less than 1 sec, "two points" if they were correct and answered in 1 sec or more, or "wrong" if they were incorrect. Participants were told that initially they must guess about which arrangement was the target, but that once they had determined which scene was the correct arrangement, they should respond as quickly and as accurately as possible. They were told that the feedback would stop partway through the experiment, but that they would still receive points for correct responses. This point system had no tangible consequences for participants and was used solely to increase their motivation to perform the task efficiently.

Each training trial began with a warning beep for 1 sec, followed immediately by the presentation of the stimulus. The stimulus was displayed continuously until the participant responded, at which time it disappeared. There was a 1-sec delay before the feedback message was played, then a 250-msec delay before the next trial. Trials for the testing portion of the experiment were identical to training trials, except there was no feedback message.

Training trials. The training trials were administered in blocks of 24. The 12 distractors in a block were each presented once (3 at each training view), and the target was presented 12 times (3 times from each training view). The presentation order of targets and distractors was randomized within each block, and separately for each participant. Participants were required to complete at least two training blocks. If accuracy exceeded 80% in the second or in any subsequent training block, then the participant proceeded to the testing portion of the experiment. Participants were told immediately before testing that they would "still be viewing the playground from several different perspectives." They were reminded that the task was to recognize the correct layout of objects, regardless of the viewing perspective.

Testing. Testing consisted of 96 trials, composed of two blocks of 48 trials. Within each block, the training, interpolated, and extrapolated perspectives were each presented 16 times, and for each of these views, half of the stimuli were targets and the other half were new distractors. Trials depicting training and extrapolated views presented equal numbers of stimuli from each of the four perspectives. The order of the trials was randomized separately for each block and for each participant.

Analysis. In all experiments reported in this article, gender and response assignment (i.e., left button = correct vs. right button = correct) were counterbalanced across participants. After averaging over these factors (neither of which ever had a significant effect or a significant interaction with any other factor), Experiment 1 represents a 2 (testing block: first vs. second) × 3 (view: training, interpolated, extrapolated) within-subjects design. For most of the statistical analyses we present, 95% confidence intervals that exclude between-subjects variation (see Loftus & Masson, 1994) are appended to their parameter estimates.

Results

Learning. The participants required an average of 91.50 ± 22.98 learning trials (ranging between two and seven blocks) before reaching the learning criterion and proceeding to test. Mean percentages correct for the first and last blocks of the learning trials were 61.33 ± 7.53 and 87.24 ± 1.90, respectively. The remaining analyses focused on the performance of participants at test.

Latency. Latency results are depicted in Figure 4. Collapsing over blocks, mean correct response times (RTs) for interpolated, trained, and extrapolated views were 1.74 ± 0.15, 1.95 ± 0.15, and 2.28 ± 0.15 sec, respec-

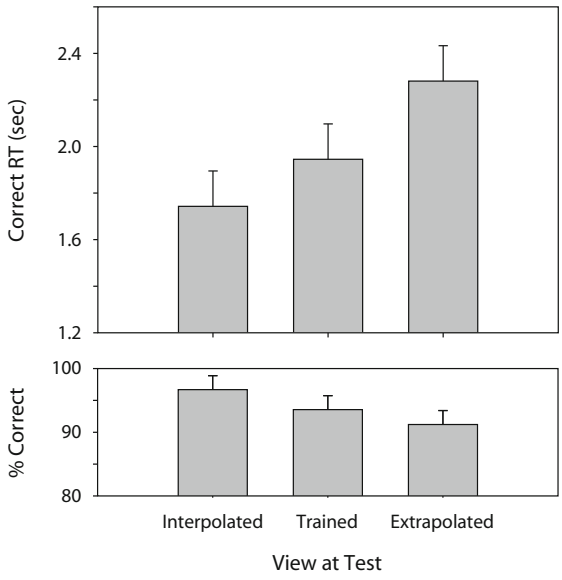

Figure 4. Mean correct response time (RT) and percent correct for recognizing the untrained interpolated view, the trained views, and the extrapolated views in Experiment 1. Error bars represent 95% confidence intervals that do not include between-participants variation.

tively. The difference between correct RTs for the interpolated and the trained perspectives was estimated at 0.20 ± 0.15, indicating significantly faster recognition of the interpolated view. Additionally, the difference between correct RTs for the trained and extrapolated views was estimated at 0.34 ± 0.15. Collapsing the data over blocks was deemed appropriate because the effects of viewing perspective on participants' correct RTs were not moderated by block [$F(2,30) = 0.79$, $MS_e = 88.81$, $p = .46$]. When averaged over the three viewpoints, performance on correctly answered targets was generally faster in the second block of trials ($M = 1.80 \pm 0.46$ sec) than in the first ($M = 2.16 \pm 0.53$ sec).

Graphical analyses suggested that even on the initial testing trials, participants were generally faster to identify the previously unseen interpolated view than the four trained views. For this analysis, we plotted participants' individually normalized latency (i.e., *z* scores computed separately for each participant across trials and view conditions) for correctly answered targets against the number of the testing trial. We then fit the plot with a locally weighted scatterplot smoother (LOWESS). This smoothing procedure works by fitting, for each trial, a predicted latency based on only the 50% of the entire set of test trials that are closest to the given trial. This subset of trials is weighted by their distance from the given trial, and the successive fit values are then connected to form the fit curve (see Cleveland, 1979, for details). These results are depicted in Figure 5.

Accuracy. The effects of testing block and viewing perspective on accuracy were similar to those of latency. Collapsing over blocks, participants' recognition accuracy for the interpolated, trained, and extrapolated views was 96.68% ± 2.20%, 93.55% ± 2.20%, and 91.21% ± 2.20%, respectively. The accuracy difference between the interpolated and the trained viewing perspectives was estimated at 3.13% ± 2.20%, indicating significantly more accurate recognition of the scene from the interpolated than from the trained perspectives. Collapsing over perspectives, participants were generally more accurate on the second block of testing trials ($M = 95.18\% \pm 3.71\%$) than on the first ($M = 92.45\% \pm 2.71\%$). As with latency, collapsing over these factors was deemed appropriate because there was no evidence for an interaction between perspective and block on participants' accuracy [$F(2,30) = 1.84$, $MS_e = 0.19$, $p = .18$].

Discussion

Participants in Experiment 1 recognized a previously unseen view of a scene more quickly and accurately than they recognized the views that they had learned only minutes before. Such a finding is not well explained by the hypothesis that people store a number of experienced views and recognize novel views by matching them to a single, specific stored view (see, e.g., Diwadkar & McNamara, 1997). Even if such a matching process was extremely fast and efficient, it would not be expected to result in faster or more accurate recognition for novel views. Instead, the present finding is consistent with the notion that participants based their recognition on processes that combined information from multiple learned views.

These results are reminiscent of—although considerably stronger than—the classic finding from the categorization literature by Posner and Keele (1968, Experiment 3), who trained people to classify stimuli that were distortions of a central, untrained prototype. Subsequent classification performance for the novel prototypes was actually numerically worse than that for previously viewed exemplars; however, the difference was small and not statistically significant. Nevertheless, this lack of a significant difference

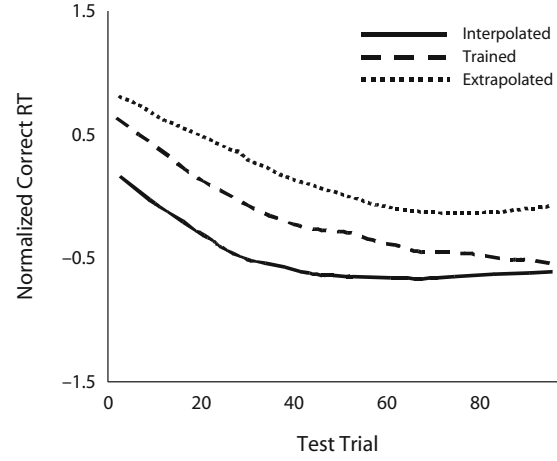

Figure 5. Local scatterplot smoother illustrating the participants' individually normalized correct response times (RTs; i.e., *z*-scores) over testing trials in Experiment 1 for the three types of test view.

between novel and trained stimuli led Posner and Keele to conclude that categorization can depend on the formation of prototypes that code the common elements from a set of learned exemplars. The present results are consistent with similar conclusions about the use of prototype representations in scene recognition; however, they are especially striking in showing a significant facilitation for novel over trained stimuli.

On the other hand, the present results are also consistent with the notion that scene recognition does not rely on prototype representations, but, rather, works by comparing current experience with multiple stored exemplars. Indeed, current theories of view combination in object recognition (see, e.g., Bülthoff & Edelman, 1992; Edelman, 1999) are generally quite consistent with exemplar accounts of categorization. In Edelman's (1999) model, for example, a new view is constructed online from the previously learned views during recognition and is compared for its similarity to the input view.

This experiment also provides initial evidence that view combination mechanisms are engaged throughout learning and do not necessarily represent an ad hoc construction that is formed during the testing phase. For example, there was little indication in our results that the recognition of the interpolated view improved at a faster rate during test than did recognition of previously seen views. Indeed, Figure 5 illustrates that recognition of the previously unseen interpolated view was generally better than that of the trained views from the very beginning of the testing session.

EXPERIMENT 2

Although Experiment 1 provided evidence for the use of view combination in scene recognition, its evidence that view combination occurred throughout learning was mostly suggestive. For example, although Figure 5 indicates that there was superior performance for the untrained view during early test trials, it should be noted that the estimates depicted in the figure involve integrating performance over several subsequent trials (see Cleveland, 1979). This pattern of results thus could have emerged from the participants' sensitivity to the statistical regularities of the testing situation and not to an ongoing process (or enduring representation) that was engaged during training.

Several features of Experiment 2 were designed to determine more precisely when during the experiment participants engage view combination processes. First, to create the training views, we made the simplifying assumption that view combination mechanisms generate a new view of a scene from a viewing perspective whose location is approximated by a weighted average of the locations of the viewing perpectives of the trained views (where the weights are based on the relative frequency of the trained views across the entire training sequence). For example, in the scenario given in the introduction, a person who approaches a picnic spot from the north twice as often as he or she approaches it from the east might be especially good at recognizing the scene from a north–northeasterly perspective, as opposed to a northeasterly perspective. Under this assumption, we examined the malleability of representations formed by view combination by altering the relative frequencies of views of the scene during training and testing. Thus, during training, participants were exposed to stimuli that were hypothesized to result in a particular combined view that was approximated by a testing stimulus that we called the "training interpolated" view. During test, the frequencies of particular test stimuli were altered so that they should lead to a different combined view—one approximated by another stimulus that we called the "testing interpolated" view. Additionally, during testing, we also exposed participants to an "average interpolated" view that depicted the scene from a perspective that was located halfway between the training and testing interpolated views. The average interpolated view was identical to the central interpolated view used in Experiment 1. Participants' relative sensitivity to the training, testing, and average interpolated views can inform us about how and when view combination occurs. In particular, we hypothesized that participants would be better at recognizing the training interpolated view than they would the testing interpolated view early in testing, and that this effect would reverse over the course of testing. Performance on the average interpolated view was expected to be slightly worse than that on the training interpolated view early in testing, and then to improve somewhat across testing trials. Of course, even if there are few performance differences among the three interpolated stimuli, it will still be important to examine differences between their recognition and that of the trained views.

Most importantly, in Experiment 2, we examined a sufficient number of participants in order to have a relatively powerful test of the differences among participants whose first test trial was a novel interpolated view versus those whose first test trial was a trained view. If performance for interpolated views on the very first test trial is better than that for previously viewed stimuli, this will provide evidence for the engagement of view combination mechanisms during learning and subsequent recognition.

An additional contribution of Experiment 2 will be to demonstrate that view combination in scene recognition may be based on training views that are more separated than those that we used in Experiment 1. In particular, in Experiment 1, all training views were separated by 15º from the untrained interpolated view. This meant that the most disparate training views were 30º apart from each other (and adjacent training views were separated by approximately 21º). In Experiment 2, we increased the separation between the training views and the average interpolated view to 30º, making the most disparate training views 60º apart from each other (and adjacent training views separated by approximately 41º). If recognition of the untrained interpolated views under these conditions is more efficient than recognition of the trained views, this will both replicate the major finding of Experiment 1 and extend it to wider training angles.

Method

Participants. Seventy-two undergraduate students (36 men and 36 women) from Miami University participated in the experiment in return for credit in their introductory psychology course. The mean age of the participants was 18.9 years ($SD = 0.9$).

Materials and Procedure. The materials for Experiment 2 were identical to those of Experiment 1. Most of the procedures from Experiment 1 were also used in Experiment 2; the only procedural differences involved changes to the identities and relative frequencies of the training and testing stimuli. In particular, participants were trained on the extrapolated stimuli of Experiment 1. Moreover, for half of the participants, the training stimuli depicting views +30º in azimuth from the central viewpoint were presented four times more frequently than those depicting the −30º azimuth viewpoint. For the other half of the participants, the −30º views were presented four times more frequently than were the +30º views. We called the most frequent training view the "overtrained side" and the least frequent training view the "undertrained side." For all participants, views from +30º elevation/0º azimuth and −30º elevation/0º azimuth were labeled "trained top" and "trained bottom," respectively, and were presented equally often during training.

During testing, the relative frequencies of the overtrained side and the undertrained side views were switched, so that, for example, the participants who trained predominantly on the −30º azimuth views were presented four times as many views from the +30º azimuth view at test. During testing, the average interpolated view, as well as the novel views from 0º elevation/±15º azimuth—were shown an equal number of times each. For participants whose overtrained side was the +30º azimuth view, the +15º and the −15º testing views were referred to as the "training interpolated" and "testing interpolated views," respectively. These labels were switched for participants whose overtrained side was the −30º azimuth view.

Participants were trained in blocks of 18 trials. In each block, views from the overtrained side, the trained top, the trained bottom, and the undertrained side were presented eight, four, four, and two times, respectively, with half of the trials being targets and half distractors. As in Experiment 1, participants proceeded to the test trials after scoring at least 80% correct on the second or any subsequent training block.

During testing, 90 trials were presented in three blocks of 30 trials. Each block presented four average interpolated views, four training interpolated views, four testing interpolated views, four trained top, four trained bottom, two overtrained side, and eight undertrained side views, with half of the trials being targets and the other half being distractors. All trials were randomized separately for each block and for each participant. Of particular interest were differences over the course of testing between the training interpolated and testing interpolated views, as well as differences between all of the interpolations and the trained views.

Analysis. After averaging over gender and response assignment, the test trials in Experiment 2 represent a 3 (block: first, second, third) × 7 (view: average interpolated view, training interpolated view, testing interpolated view, trained top, trained bottom, overtrained side, and undertrained side) within-subjects design. As with Experiment 1, analyses were conducted on RTs for correctly answered target items and on accuracy.

Results

Learning. The participants required an average of 82.65 ± 14.54 learning trials (ranging from 3 to 28 blocks) before reaching the learning criterion and proceeding to test. The mean percentages correct for the first and last blocks of the learning trials were 65.02 ± 3.53 and 89.75 ± 1.58, respectively. The remaining analyses focused on the performance of participants at test.

Latency. Contrary to our expectations, the difference between correct recognition RTs for the training interpolated and testing interpolated views was not moderated by test block [$F(12,708) = 0.89$, $MS_e = 296.32$, $p = .56$]. Collapsing over testing blocks, there were significant differences among the views [$F(6,426) = 8.38$, $MS_e = 102.43$, $p < .01$], as depicted in Figure 6. Only 3.08% of the variance of the main effect of view was accounted for by differences among the three novel interpolated views (average, 1.46 ± 0.075 sec; training, 1.52 ± 0.075 sec; and testing, 1.48 ± 0.075 sec), and, accordingly, there were no statistical differences among the three interpolated views [$F(2,142) = 0.88$, $MS_e = 90.57$, $p = .42$]. Similarly, only 16.12% of the variance of the main effect of view was accounted for by differences among the four trained views (top, bottom, overtrained, and undertrained), and there were no statistical differences among these four views [$F(3,213) = 2.55$, $MS_e = 108.46$, $p = .06$]. Importantly, 80.80% of the variance of the main effect of view was accounted for by the contrast comparing the three interpolated views with the four trained views. This contrast (scaled so that the sum of the squared weights was 1) was estimated at 0.22 ± 0.01, indicating significantly faster performance with the interpolated views than with the trained views.

We next examined the 31 participants whose very first test trial was a correctly recognized target view. Because relatively few of these participants first viewed a training ($n = 3$), testing ($n = 4$), or average ($n = 3$) interpolated view, and because no differences were found among the three interpolated views, we collapsed participants whose first view depicted any of the three interpolations and compared their RTs with those of participants whose first test trial was a correctly recognized trained view ($n = 21$). Recognition time was faster for participants judging an interpolated view ($M = 1.57 \pm 0.70$ sec) than for those judging a previously trained view ($M = 3.05 \pm 0.86$ sec). The difference between these groups was estimated at 1.48 ± 1.31 sec.

Finally, as in Experiment 1, we plotted participants' individually normalized latency (i.e., z score) for correctly answered targets against the number of the testing trial, and fitted the plot with a local scatterplot smoother (again using a 50% fitting window; see Cleveland, 1979). The results, depicted in Figure 7, suggest that participants responded more quickly to the training interpolated view than to the testing interpolated view early in testing, but that by around Trial 20, people were faster to recognize the testing interpolated view than they were the training interpolated view.

Accuracy. As depicted in Figure 6, there was a significant effect of viewpoint [$F(6,426) = 10.69$, $MS_e = 1.58$, $p < .01$] on accuracy. As with the RTs, a large (45.71%) percentage of the variance of the main effect of viewpoint was accounted for by the contrast examining the difference between the three novel interpolated views and the four previously learned trained views. This contrast was estimated at 4.24% ± 0.26%. A much smaller (6.3%) percentage of the variance of the main effect of viewpoint was accounted for by the difference between the novel training and the testing interpolated views, which was estimated as 2.43% ± 1.49% and indicated greater accuracy with the testing interpolated view than with the training interpolated view. As with previous analyses, there was no evidence for a block × view interaction [$F(12,852) = 1.09$, $MS_e = 1.22$, $p = .36$].

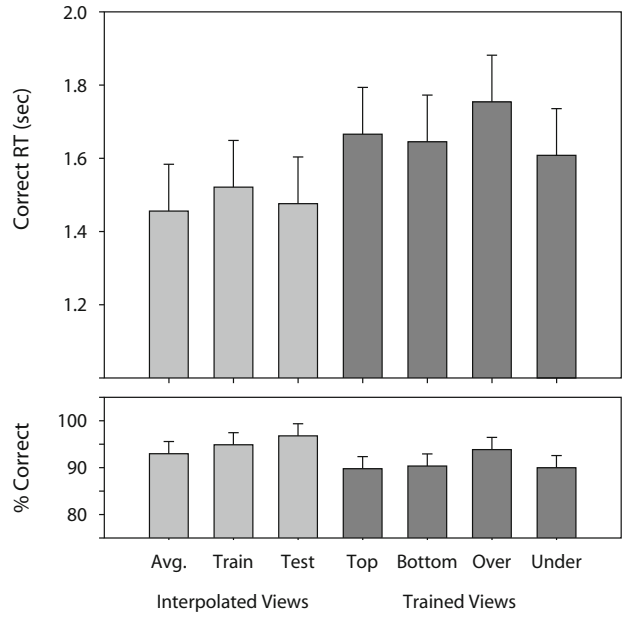

Figure 6. Mean correct response time (RT) and percent correct for recognizing the different test views in Experiment 2. The leftmost (light gray) bars represent performance on the three types of interpolated test views; the rightmost (dark gray) represent performance on the training views during testing. Error bars represent 95% confidence intervals that do not include between-participants variation.

We also examined accuracy differences between interpolations and trained views for the participants' first test trial. Accuracy was better for participants whose first view was one of the three interpolated views ($M = 92.86\% \pm 10.17\%$) than for those whose first view was a previously trained one ($M = 86.36\% \pm 4.06\%$). The difference between these groups was estimated at $6.49\% \pm 15.29\%$.

Discussion

In addition to replicating the main result of Experiment 1—that participants recognized a novel view more quickly and accurately than they did familiar views—the present experiment offers strong additional evidence that the psychological structures (or processes) that support this effect are in place (or occur) before testing. Participants who viewed a novel interpolated stimulus on their very first test trial recognized it more quickly and more accurately than did participants whose first test trial was a familiar view of the scene.

Despite the strength of this effect, little of it was accounted for by differences between recognition of the novel training interpolated and testing interpolated views. Although Figure 7 depicts relatively good performance with the training interpolated view early in testing, and with the testing interpolated view after around Trial 20, this difference was not corroborated by a test of the interaction between testing block and viewing perspective. One possible reason for the lack of this difference may be that the representations that result from view combination both are formed rapidly and endure long enough not to lose their efficacy for recognition. The rapid formation of these representations could explain why the testing interpolation was well recognized during the first testing block and suggests that despite the relatively low frequency of the undertrained view during training, the undertrained view may have been influential in the view combination process. The idea that a representation formed by view combination endures after others are established could explain why the training interpolation was also well recognized during testing. The possibility that these representations are rapidly formed and relatively long lived may be especially likely when one considers that in the present experiment, the mental representations corresponding to the training and testing interpolations may have been formed within a few moments of each other, and that there was no necessary reason for one to overshadow the other.

GENERAL DISCUSSION

The present experiments demonstrate that recognizing scenes can involve more than merely comparing one's current view of a scene with a single previously experienced view. The repeated finding that our participants were able to recognize novel views of a scene more efficiently than they were able to recognize previously learned views suggests instead that people combine the information from different views of a scene during (or before) the process of recognizing a novel view.

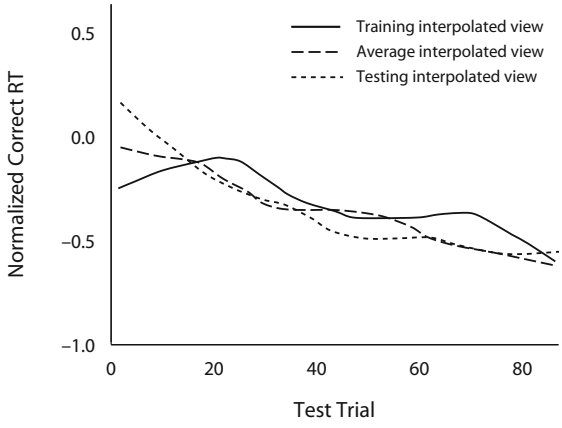

Figure 7. Local scatterplot smoother illustrating the participants' individually normalized correct response times (RTs; i.e., *z* scores) over testing trials in Experiment 2 for the different interpolated views.

The literature on object recognition may provide clues about how a view combination mechanism can be used to bring about successful scene recognition. One of the best-specified accounts of object recognition is Edelman's (1999) view combination model, which posits the existence of shape prototypes that are arranged in a multidimensional "shape space." Distance in this space is specified by object similarity so that objects that are similar in structure are closer together in the space. When a novel view of a familiar object is presented, or when a completely new object is presented, all of the stored prototypes that are similar to the test stimulus (above a threshold parameter) along one or more dimensions provide activation to construct a new "view" that is then compared with the input. To the extent that the constructed view matches the structure of the input view, recognizing the input will be easy, even if it is a novel view or a completely new object. In this conception, view combination occurs relatively rapidly, during the input of the novel view, because it activates structurally similar object prototypes created during prior experiences with the objects.

It is worth noting that one way to conceptualize view combination is as a form of generalization (Friedman et al., 2005). In particular, for object recognition, Edelman (1999) proposed that the prototype shapes in the multidimensional space were excitatory radial basis functions (modeled as Gaussians) that become activated to the extent that they overlap in structural similarity with the to-be-recognized input. In principle, the summation of generalized excitation could cause a higher peak than any single function that contributes to the summation, which may explain the present results.

Although scenes are clearly more complex than single objects, the present work suggests that under certain circumstances, scene recognition may also rely on view combination mechanisms. The present data do not allow us to determine whether there are scene prototypes that are based on structure (e.g., the relative angles and distances between objects) or possibly on content (e.g., commonly associated objects), but our results are consistent with the idea that the untrained interpolated viewpoint in the present study was highly similar to a mental representation that resulted from the combination of multiple training views. Moreover, the present data suggest that the view combination process for scenes is dynamic, flexible, and ongoing throughout both training and testing.

Despite providing solid evidence for the existence of view combination mechanisms that operate in scene recognition, our findings do not resolve the issue of whether view combination occurs primarily during encoding or retrieval. On one hand, it is possible that as stimuli are encoded, they are combined with representations of other recently viewed scenes, or with an existing scene representation. By this account, mental processes are engaged at encoding to ensure that the stored representation is both flexibly functional and efficiently organized. This representation does not necessarily contain information identical to specific individual experiences; rather, it contains relatively abstract information about the scene in general (e.g., spatial relations that are invariant over viewpoint changes). It is thus possible for the information in memory to be more similar to an object or event that has not been experienced than to one that has. On the other hand, the present data are also consistent with the notion that people remember a set of specific instances and then combine them during retrieval. Indeed, Edelman's (1999) view combination model explains recognition as involving retrieval-based processes that construct a predicted view from activated prototypes.

Although the relative efficiency of recognition for some novel scenes as compared with trained views supports the existence of representations formed by view combination, our data also indicate that these representations are not so abstract as to be orientation-free, or "allocentric." Indeed, our results clearly indicate that participants had a preferred orientation in spatial memory: the orientation of the interpolated stimulus scene. Thus, unlike much of the contemporary literature on spatial memory for layouts (e.g., Shelton & McNamara, 2001), the preferred orientation in memory was not coincident with a view that participants had seen during the training session, but would be better described as "the average" of the trained views.

This finding is reminiscent of Mou and McNamara's (2002) results, in which people exhibited a preferred direction in spatial memory that was based not on their personal experience, but on the geometrical structure of an array of objects. Interestingly, because the playground objects in the present scene were randomly arrayed (see Figure 2), they likely did not provide a salient geometric structure (such as an axis of symmetry) that could be leveraged as a preferred axis in spatial memory. Alternatively, it is conceivable that the interpolated views in these experiments, for an unknown reason, represented a more "canonical" view of a playground scene than did the trained views (see Palmer, Rosch, & Chase, 1981, for a discussion of canonical views in object recognition). However, to the extent that canonical views of objects or scenes correspond to those that are frequently encountered, the fact that the best-recognized interpolated views in the present experiments depicted the

scene from an aerial perspective renders this possibility unlikely. Indeed, there is some evidence that the recognition of an object from perspectives that differ in elevation from the trained views may be more difficult than recognition of the object from perspectives that differ in azimuth (Edelman & Bülthoff, 1992).[1] Thus, we think that in the present case, the use of the unseen interpolated view as a privileged direction in spatial memory was guided by the geometrical arrangement of the viewpoint locations of the training stimuli, and not by the intrinsic or canonical properties of the object array.

To our knowledge, the finding of superior recognition of previously unseen naturalistic scenes is novel and begs the question of what types of stimuli give rise to this effect. We speculate that this finding depends critically on the number of training stimuli, the number of viewing perspectives, and the relative overlap of the spatial information they contain. With enough training stimuli—especially those with redundant or overlapping relational information—it seems quite adaptive to extract, store, or retrieve the common elements rather than, or in addition to, the myriad individual views. In the present case, the viewpoints from which participants learned about the scene provided a great deal of overlapping spatial information through variations in both their elevation and azimuth. However, stimuli that vary in elevation may be less critical than those varying in azimuth for generating superior recognition performance of novel versus trained views, because people routinely view and learn about scenes from the fixed elevation provided by their eye height.

Finally, two additional aspects of the present results are worthy of notice. First, in Experiment 1, recognition of novel stimuli from perspectives that were not centrally located, but were equally distant from trained views (i.e., extrapolated views), was worse than recognition of either trained or interpolated views. This finding demonstrates that the ease with which the interpolated views were recognized (as compared with the trained views) was not simply a function of the distance between training and testing views, but that, instead, the facilitation for novel views relies critically on the geometric arrangement of the training views in comparison with the untrained interpolated view. Second, the demonstration of this interpolation effect from training views that were separated by up to 60º in Experiment 2 suggests that view combination processes may be fairly robust. Taken together, the results of these experiments indicate that—at least under some circumstances—recognition of a familiar scene from novel points of view relies on a comparison with combinations of our prior experiences, rather than comparison with a single discrete experience.

AUTHOR NOTE

We thank Geoff Hollis, Bernd Kohler, and Douglas Robertson for assistance with preparing the stimuli, programming, and conducting the experiments. Portions of the research were supported by a grant to A.F. from the Natural Sciences and Engineering Research Council of Canada. Work by E.H. on this project was supported by a gift from Ted Smith to Miami University. Correspondence should be addressed to D. Waller, Department of Psychology, Miami University, Oxford, Ohio 45056 (e-mail: wallerda@muohio.edu).

REFERENCES

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Science*, **89**, 60-64.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association,* **74**, 829-836.

Diwadkar, V. A., & McNamara, T. P. (1997). Viewpoint dependence in scene recognition. *Psychological Science*, **8**, 302-307.

Edelman, S. (1999). *Representation and recognition in vision*. Cambridge, MA: MIT Press.

Edelman, S., & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, **32**, 2385-2400.

Edelman, S., Bülthoff, H. H., & Bülthoff, I. (1999). Effects of parametric manipulation of inter-stimulus similarity on 3D object categorization. *Spatial Vision*, **12**, 107-123.

Friedman, A., Spetch, M. L., & Ferrey, A. (2005). Recognition by humans and pigeons of novel views of 3-D objects and their photographs. *Journal of Experimental Psychology: General*, **134**, 149-162.

Friedman, A., & Waller, D. (2008). View combination in scene recognition. *Memory & Cognition*, **36**, 467-478.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **93**, 411-428.

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, **262**, 1747-1749.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, **1**, 476-490.

Mou, W., & McNamara, T. P. (2002). Intrinsic frames of reference in spatial memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 162-170.

Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. [B.] Long & A. [D.] Baddeley (Eds.), *Attention and performance IX* (pp. 135-151). Hillsdale, NJ: Erlbaum.

Palmeri, T. J., & Flanery, M. A. (1999). Learning about categories in the absence of training: Profound amnesia and the relationship between perceptual categorization and recognition memory. *Psychological Science*, **10**, 526-530.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353-363.

Shelton, A. L., & McNamara, T. P. (2001). Systems of spatial reference in human memory. *Cognitive Psychology*, **43**, 274-310.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, **21**, 233-282.

Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. *Cognition*, **67**, 21-44.

Zaki, S. R., & Nosofsky, R. M. (2007). A high-distortion enhancement effect in the prototype-learning paradigm: Dramatic effects of category learning during test. *Memory & Cognition*, **35**, 2088-2096.

NOTE

1. Notably, Bülthoff and Edelman (1992) trained with viewpoints that had the same elevation; thus, they expected (and found) that novel viewpoints that differed in elevation from the trained viewpoints would be more difficult to recognize than novel views at the same elevation. This result did not occur in the present experiments. For both experiments, we collapsed the two training stimuli that depicted the scene from the same elevation (e.g., the over- and undertrained perspectives) and compared their recognition at test with the two training stimuli that had the same azimuth (e.g., the top and bottom perspectives). In no case were estimates of the differences between these two types of stimuli different from zero. Given Bülthoff and Edelman's results, it is likely that the differences were absent because we trained at different elevations.

(Manuscript received June 2, 2008;
revision accepted for publication September 7, 2008.)