Memory & Cognition
2008, 36 (3), 467-478
doi: 10.3758/MC.36.3.467

View combination in scene recognition

Alinda Friedman
University of Alberta, Edmonton, Alberta, Canada

and

David Waller
Miami University, Oxford, Ohio

Becoming familiar with an environment requires the ability to integrate spatial information from different views. We provide evidence that *view combination*, a mechanism believed to underlie the ability to recognize novel views of familiar objects, is also used to recognize coherent, real-world scenes. In two experiments, we trained participants to recognize a real-world scene from two perspectives. When the angular difference between the learned views was relatively small, the participants subsequently recognized novel views from locations between the learned views about as well as they recognized the learned views and better than novel views situated outside of the shortest distance between the learned views. In contrast, with large angles between training views, all the novel views were recognized less well than the trained views. These results extend the view combination approach to scenes and are difficult to reconcile with models proposing that scenes are recognized by transforming them to match only the nearest stored view.

As mobile organisms, humans benefit from the ability to recognize places and scenes from vantage points that they have not previously experienced. This ability implicates the existence of psychological mechanisms that compare spatial information from current and previously experienced perspectives. Although the existence of these mechanisms is self-evident, theoretical formulations of how they enable people to recognize scenes have yet to be fully developed, and empirical evidence that would constrain these theories is sparse. In contrast, models of the processes that underlie people's recognition of single objects have undergone significant theoretical development (Biederman, 1987; Biederman & Gerhardstein, 1993; Bülthoff & Edelman, 1992; Bülthoff, Edelman, & Tarr, 1995; Edelman, 1999; Tarr & Pinker, 1989; Ullman, 1989, 1996), and a wealth of results bears on the question of how prior views of objects enable recognition of novel views.

According to a *normalization* account of object recognition, people store a number of *exemplar* views of objects; to recognize a novel view, they transform the novel percept so that it matches the nearest (i.e., most similar) stored exemplar of the object (Tarr, 1995; Tarr & Pinker, 1989; Ullman, 1989). Normalization models assume that even when a novel view is equidistant from two (or more) stored representations, it is normalized with respect to only one of them. Thus, one of the behavioral implications of a normalizing process is that a graceful, roughly monotonic increase in error rate, recognition time, or both will be observed as a function of the distance between the novel view to be recognized and the nearest learned view. It is in this sense that a normalization approach predicts that recognition will be *viewpoint dependent*.

An alternative account of object recognition, known as *view combination*, maintains that the extent to which a novel view of an object can be readily recognized depends on its degree of structural similarity to a set of *multiple* stored views (Bülthoff & Edelman, 1992; Edelman, 1999; Edelman & Bülthoff, 1992; Edelman, Bülthoff, & Bülthoff, 1999). In contrast to the degradation in performance predicted if recognition occurs by normalization, the view combination approach predicts that it is possible for some novel views to be recognized as well as familiar views, because information from two or more structurally similar stored views can be combined to facilitate recognition. However, recognition should not be facilitated if the view to be recognized is too structurally disparate from the familiar (stored) views.

More formally, the predictions of the view combination approach arise because objects are represented as points in a multidimensional *shape space* that is spanned by their parametric similarities to a small number of reference objects, which may be construed as prototypes (Edelman, 1999). Recognizing known objects from novel viewpoints occurs by mathematically interpolating between the shape parameters of two or more prototypes to compute, or "predict," the novel view. The predicted view is then matched to the percept of the novel view.

As has been described in detail by Edelman (1999), the view combination process is functionally analogous to generalization (or activation) from more than one source

A. Friedman, alinda@ualberta.ca

(see also Friedman, Spetch, & Ferrey, 2005). Accordingly, when a novel view is between and relatively close to two or more views stored in the shape space, the predicted view is generated from sources that are structurally similar to the novel view. This leads to relatively easy recognition. The more distant the stored views are from the novel view, the less they can contribute to its recognition, because the relevant generalization functions do not overlap sufficiently; in this situation, recognition is relatively difficult. In other words, the process of generalizing from representations of two or more stored views would be expected to produce superior recognition of novel views only if there is overlap between the generalization functions. Thus, depending on the tuning of its underlying generalization mechanisms, the view combination approach can predict when performance on novel views of an object should be *viewpoint invariant* and when it should be viewpoint dependent (Edelman, 1999; Edelman et al., 1999). On the other hand, the normalization approach predicts that recognition should always be viewpoint dependent, because the normalization process uses information from only one stored representation.

The pattern of results predicted by view combination mechanisms has been demonstrated repeatedly for human object recognition (see, e.g., Bülthoff & Edelman, 1992; Spetch & Friedman, 2003) and, more recently, for birds (Friedman et al., 2005; Spetch, Friedman, & Reid, 2001). Because scenes are composed of objects, it is reasonable to suppose that recognizing scenes may employ some of the same basic psychological mechanisms as object recognition. The present experiments were designed to provide evidence for this idea.

Although there is a large literature that has examined how information across saccades, or from small changes to a single scene, is integrated into working memory (e.g., Henderson & Hollingworth, 2003; Hollingworth & Henderson, 2004), there have been very few attempts to explain how multiple views of the same scene are encoded in long-term memory and used to support the recognition of novel views of the scene (but see Hock & Schmelzkopf, 1980). Similarly, there have been very few attempts to generalize object recognition processes—including either normalization or view combination—to scene recognition. An exception is Diwadkar and McNamara (1997), who explicitly addressed the question and concluded that novel views of scenes are recognized by normalizing them to the nearest stored view, rather than by view combination. In their experiment, participants learned a small-scale (i.e., nonnavigable) layout of unrelated objects from several perspectives and were subsequently asked to recognize the layout from both the learned and the novel orientations. Some of their novel orientations (the *interpolated* views) were between and equally proximate to the learned (training) views, whereas others (the *extrapolated* views) were outside of the shortest distance between the training views (see Figure 1). Diwadkar and McNamara reported that recognition latencies for the novel views were linearly related to their angular distance from the nearest training view, irrespective of whether the novel views were interpolated or extrapolated. That is, there was a similar decline in performance for both interpolated and extrapolated stimuli, relative to the training views. Because a view combination approach predicts facilitation for novel views that are between two learned views, Diwadkar and McNamara concluded that their results provided no support for the existence of view combination mechanisms for scenes.

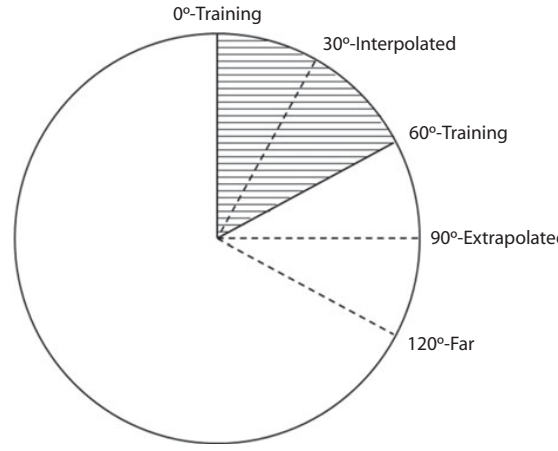

Figure 1. Schematic diagram of the stimulus viewpoints for Experiment 1, as they would be seen from above. The playground is situated at the center of the circle, and the 0º starting point is arbitrary. For one group, the views from 0º and 60º were the training views. For this group, in keeping with Bülthoff and Edelman's (1992) and Diwadkar and McNamara's (1997) terminology, we refer to the novel view at 30º as the *interpolated* view and the novel view at 90º as the *extrapolated* view.

Additional work by McNamara, Diwadkar, Blevins, and Valiquette (2006) offers a suggestion as to why Diwadkar and McNamara's (1997) results did not provide strong evidence for view combination in scene recognition. McNamara et al. asked participants to learn multiple configurations of colored dots on a computer monitor from two depicted vantage points that were separated by 75º. In some conditions, the learning stimuli were presented in a way that induced apparent rotation of the configuration; in others, there was no apparent rotation during learning. Immediately after learning each configuration, the participants were asked to recognize the configuration from interpolated, extrapolated, or trained viewpoints. Consistent with a view combination account, the participants were faster at recognizing interpolated views than they were at recognizing extrapolated views, but importantly, this effect occurred only under conditions designed to induce apparent rotation of the stimulus.

To explain the difference in performance between the apparent and the nonapparent rotation conditions, McNamara et al. (2006), following Marr (1982), suggested that the participants mentally represented the stimulus configurations at two different levels of coding. First, a basic sensory-perceptual level of coding represented the local features of the configuration, such as the locations of the dots on the screen. Second, a more abstract object level of coding represented the spatial structure of the configuration of dots. At this more abstract level of

coding, the mental representation of a scene was likely to include the angles and distances between the depicted objects as they would exist in the real world, rather than as they simply appeared on the display. McNamara et al. suggested that the faster recognition of interpolated versus extrapolated views arose only when the participants were relying on representations at the more abstract level. That is, by emphasizing the structural relations of the configurations, the apparent rotation conditions had facilitated or enhanced the use of more abstract representations. Similar conclusions about the ability of apparent rotation to facilitate the automatic induction of an object's structure and, thus, induce object-level representations were made by Friedman and Harding (1990). Other researchers have evidence that even during active scene perception, visual representations are not necessarily veridical or specific across saccades (e.g., Henderson & Hollingworth, 2003).

If relatively abstract representations are necessary for view combination to facilitate recognition, it is possible that Diwadkar and McNamara (1997) found no evidence for view combination in scene recognition because their participants did not represent the stimuli they used at a sufficiently abstract level. For example, Diwadkar and McNamara's "scenes" consisted of collections of unrelated objects (e.g., a coffee mug, a lightbulb, scissors, a screwdriver, a stapler, and a strainer) that normally would not make up a scene and, thus, might not cohere to form an abstract-level scene representation. Indeed, Mandler and Parker (1976) showed that memory for the locations of objects in a scene could be disrupted if the objects were presented as mere collections, instead of as coherent scenes (e.g., by turning the pictures upside down, removing the horizon line, and rerighting the objects), even when these scenes contained only semantically related objects. Thus, we thought that it was important to reexamine the possibility that view combination processes occur during scene recognition in situations designed to induce more abstract mental representations of the stimuli.

In the present experiments, we had participants learn two views of a real-world complex scene (Experiment 1) or a computer-generated simulation of the scene (Experiment 2). In both experiments, the scenes were clearly recognizable as playgrounds, depicted from a ground-level perspective. Moreover, the participants' task ensured that they had to attend to the spatial relations among objects in the scene to perform successfully. As a result of attending to these relations, we expected that the overall representation of the scene would be relatively abstract. Demonstrating that these scenes promoted view combination (i.e., that interpolated novel views were recognized more quickly and accurately than extrapolated novel views) would provide evidence for McNamara et al.'s (2006) speculation that view combination is supported by relatively abstract representations. However, it would also demonstrate that apparent rotation is not necessary to achieve this level of representation and that view combination may be a relatively common mechanism that underlies the recognition of both single objects and real-world scenes.

In Experiment 1, we trained participants with two views of a real-world scene and subsequently tested their recognition of interpolated, extrapolated, and trained views. In Experiment 2, we tested the prediction of view combination models that facilitation will not occur when the trained views are sufficiently disparate. To preview our results, in Experiment 1, we demonstrated that people recognized novel views of scenes about as readily as they recognized views that they had seen before, provided that the novel views depicted a perspective that was between two familiar views. Views outside this training range were recognized more slowly and less accurately than either the trained or the inside views. In Experiment 2, we replicated this effect and, additionally, showed that it does not occur when training views are sufficiently far apart. We interpret the pattern of results across experiments as implying that view combination processes exist for scenes that function similarly to the view combination processes hypothesized to underlie object recognition.

EXPERIMENT 1

In Experiment 1, participants learned to discriminate two target views of the same scene that were 60º apart from distractors that differed from the targets by subtle changes. They were then tested on the trained views, as well as on novel views of the target scene that were either in between the two training views or outside of the shortest distance between them. If, as is predicted by view combination, participants interpolate between training views but do not extrapolate beyond them, novel views that are between the training views should be recognized about as quickly and accurately as the training views and more quickly and accurately than novel extrapolated views.

We created two kinds of distractors from the original target scenes. For *move* distractors, one of the central objects in the scene was moved from its original location. For *switch* distractors, two of the central objects switched places with each other. We chose these kind of distractors because previous research on scene memory had shown that move and switch distractors are relatively difficult to detect (Mandler & Parker, 1976), even when it is known that the objects involved have been fixated during encoding and participants expect a recognition test (Friedman, 1979). In addition, the switch distractors, in particular, encourage participants to attend to all of the objects in a given scene, as well as to the relations between them; performing correctly on the novel views during testing is virtually impossible otherwise. Finally, performance differences between the two types of distractors could inform us about the degree to which participants generate representations of these scenes that include the depicted real-world interobject spatial relations. For example, whereas switch distractors preserve all of the spatial relations among objects, move distractors alter at least one, and potentially many, of these spatial relations. If interobject spatial relations are encoded and used to form abstract scene-level representations, move distractors, which disrupt these relationships, should be more readily identifiable than switch distractors. Thus, a comparison of performance on the distractor types should also provide evidence for relatively abstract coding of our stimuli. That is, if we obtained evidence for

view combination, the need to encode interobject relations to distinguish the targets from the distractors may have contributed to this result.

Method

Participants. Forty volunteers (18 men and 22 women) from the University of Alberta participated for partial course credit. Of these, 28 participants (13 men and 15 women) met the accuracy criterion of scoring at least 85% correct on each of the two training views during the test trials. However, 4 of the individuals (2 men and 2 women) who met the accuracy criterion had no correct trial for at least one of the other view conditions, and their data were not considered further. It is worth noting that the 70% of the participants who reached the accuracy criterion of 85% correct for the training views during test trials is comparable to the percentage of participants who reached criterion in several of the original studies that provided evidence for view combination mechanisms in object recognition (e.g., Edelman, 1999; Edelman et al., 1999).

Stimuli and Design. Digital color photographs were taken of a playground setting in Oxford, Ohio. Five views of the playground were taken from the circumference of an imaginary circle (radius = 18.29 m) whose center was located at the approximate center of the playground. The starting view was arbitrarily labeled 0º, and subsequent photographs were taken every 30º counterclockwise from that point, resulting in views at 0º, 30º, 60º, 90º, and 120º. Grayscale versions of the stimuli are shown in Figure 2.

Half the participants were trained with the 0º and 60º views and were tested with those views, as well as with novel views that were interpolated (30º), extrapolated (90º), and far (120º), relative to the training stimuli. The other half of the participants were trained with the 60º and 120º views and were tested with those views, as well as with the novel views at 90º (interpolated), 30º (extrapolated), and 0º (far). Note that this design counterbalances interpolated and extrapolated views across groups, thus ensuring that any idiosyncratic differences between the stimuli depicting interpolated and extrapolated views would not be expected to account for performance differences between them. Furthermore, each of the central objects was fully visible from each of the viewpoints.

We used digital imaging software to construct two distractors for each of the five target views. In the first distractor type (*move*), one of the five central objects in the foreground of each target scene shifted its location, and the area where it had been was filled in with grass copied from another section of the picture. For the other distractors (*switch*), two of the foreground objects in each view traded places with each other, and any gaps left by the switch were filled with grass. The appearance of the grass in the five target views was also altered slightly, so that traces left by the editing process could not be used to distinguish targets from distractors. Grayscale versions of the distractors are shown in Figure 3.

A training block consisted of two trials for each training view and one trial for each of its two distractors, for a total of eight trials. The participants completed at least three blocks of training trials. If they achieved 85% correct or better on the third block of training trials, they proceeded to the test trials; otherwise, they continued to receive training blocks until the criterion was reached.

A test block consisted of 2 trials for each of the five target views and 1 trial for each target's two distractors, for a total of 20 trials. The participants received five test blocks. The order of the stimuli was randomized within blocks during both the training and the test phases.

Procedure and Apparatus. After some preliminary instructions, the participants were seated in front of a computer monitor, where they read the following. "In this experiment you're going to be looking at pictures of different layouts of a playground. One arrangement of the objects in the playground is correct, and your job is to learn which is the correct arrangement and to distinguish it from all of the different incorrect arrangements." They were also told that half of the pictures depicted the correct arrangement and the other half were incorrect, and that the presentation order would be randomized between the two.

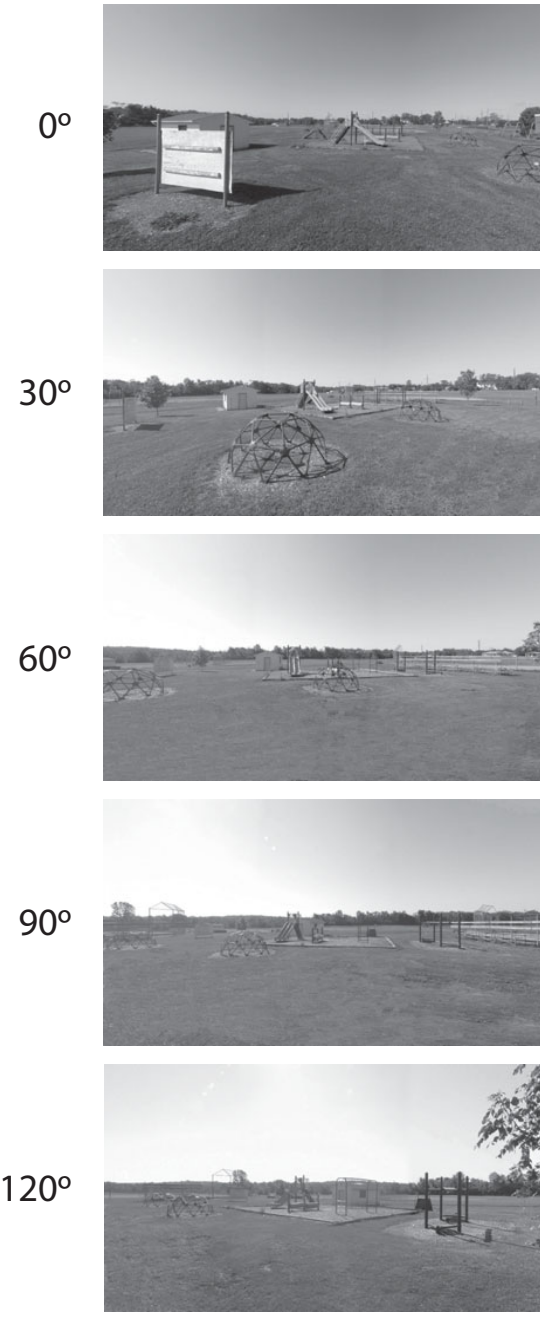

Figure 2. Grayscale versions of the target stimuli used in Experiment 1. Participants learned two target views (either 0º and 60º or 120º and 60º) and subsequently recognized the scene from all of the perspectives shown.

The stimuli were displayed on a 17-in. computer monitor. The participants responded with the index finger of either hand by pressing one of two buttons on a response box connected to the computer through its joystick port. The box was 11 cm wide × 6 cm long × 8 cm high and was situated in front of the computer monitor. The centers of the response switches were approximately 8 cm apart.

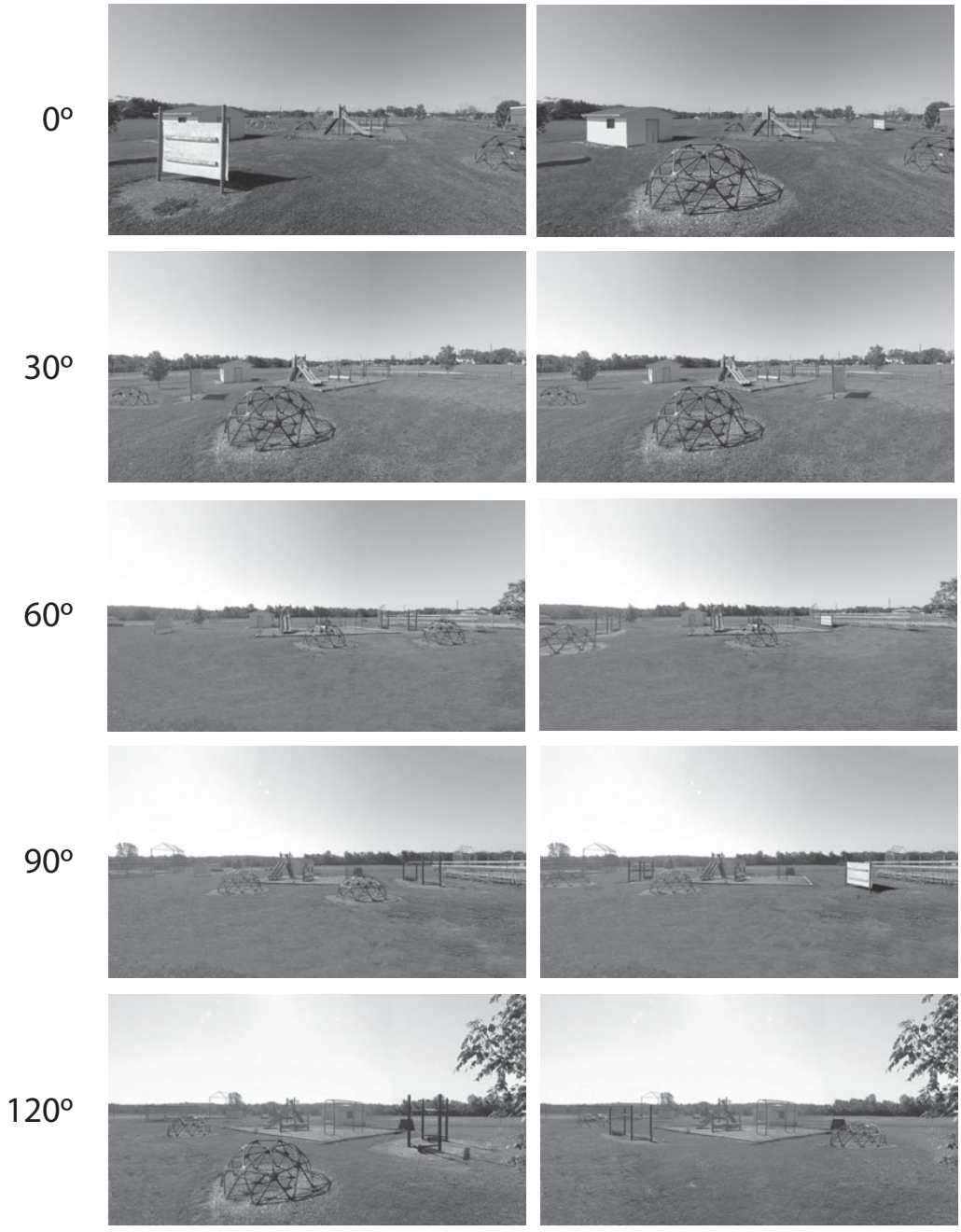

Figure 3. Grayscale versions of the distractor stimuli used in Experiment 1. Each distractor either moved (left column) or switched (right column) selected objects with respect to the corresponding target scene (see Figure 1).

The labels above the switches were "+" and "−," and the participants were asked to respond by pressing the "+" key if the current stimulus was a picture of the target scene and the "−" key if it was a picture that had been changed in some way from the target scene. For half the participants, the "+" key was on the left, and for the other half, it was on the right.

The participants received auditory feedback over headphones during the training trials. If the participants were correct on a training trial, the feedback informed them that they had received 1 point. If they were wrong, the feedback said "wrong." The participants were told 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 1 point for each correct response and no points for each wrong response, so they should still try to respond quickly and accurately. This point system had no

tangible consequences for the participants and was used solely to increase their motivation to perform the task efficiently. Finally, the participants were told the following. "During the part of the experiment in which you don't get feedback, many of the pictures that you see will have been taken from other locations than what you learned. Remember that your job is to identify the correct arrangement, regardless of where you view it from."

On each training trial, there was a warning beep for 1 sec, followed immediately by the stimulus. When the participant responded, there was a 1-sec delay, and then he or she received the feedback message over the headphones for approximately 1 sec. There was a 2-sec delay before the next trial. The procedure on test trials was identical, except that there was no feedback.

Results and Discussion

In all the tests reported, we adopted an alpha level of .05. We trimmed correct response times (RTs) longer than three standard deviations above the means computed individually for each participant separately over the target and distractor test conditions. The omitted trials represented 1.8% and 1.3% of the targets and distractors, respectively. For all the analyses, group (i.e., participants who learned the 0º and 60º views vs. those who learned the 60º and 120º views) was entered as a factor but exerted no effects or interactions and was dropped from the analyses we will report.

Response times. The mean RTs are shown in Figure 4, top panel. There was a main effect of viewing angle [training, interpolated, extrapolated, and far; $F(3,69) = 5.88$, $MS_e = 347{,}682.30$]. Planned comparisons showed that the difference between the training and the interpolated conditions was not significant [$F(1,23) < 1$] and accounted for just 0.3% of the variance in the main effect of viewpoint. In contrast, there was a significant difference between the interpolated and the extrapolated conditions [$F(1,23) = 4.87$, $MS_e = 726{,}822.99$], and this difference accounted for 28.9% of the main effect of viewpoint. In addition, correct detection of the switch distractors was slower ($M = 2{,}907$ msec) than detection of the move distractors ($M = 2{,}174$ msec), and this difference was significant [$F(1,23) = 9.52$, $MS_e = 678{,}918.07$].

Response accuracy. For the 24 participants whose RT data were included in the analyses above, the effect of view was also reliable for percent correct [$F(3,69) = 4.73$, $MS_e = 163.85$; see Figure 4, bottom panel]. The percent correct on training trials did not differ from that on interpolated trials [$F(1,23) = 2.59$, $MS_e = 116.26$, $p = .12$]. Although the difference in accuracy between interpolated and extrapolated views also failed to reach significance [$F(1,23) = 2.59$, $MS_e = 465.04$, $p = .12$], the means were in the direction predicted by a view combination account (92.9% vs. 85.8%, respectively). Furthermore, similar to the pattern observed for the RTs, for percent correct, the training versus interpolated contrast accounted for just 6.5% of the variance in the view main effect, whereas the interpolated versus extrapolated contrast accounted for 25.9% of the variance. The accuracy difference between the move and the switch distractors failed to reach significance [$F(1,23) = 1.50$, $MS_e = 256.64$, $p = .23$], although the means were also in the predicted direction of relative difficulty (74.0% and 68.3%, respectively).

The 12 participants who did not meet the accuracy criterion for inclusion in the group data analysis of RTs also displayed a pattern of performance that was predicted on the basis of the view combination approach: Their accuracy on the training, interpolated, extrapolated, and far trials was 62.9%, 71.7%, 46.7%, and 43.3%, respectively [$F(3,33) = 3.80$, $MS_e = 568.70$]. The percent correct for the training and interpolated conditions was significantly better than chance [$t(11) = 2.87$ and 2.66, respectively]. In addition, the switch distractors ($M = 46.3\%$) were also more difficult than the move distractors ($M = 61.3\%$) for this group of individuals [$F(1,11) = 4.44$, $MS_e = 304.18$, $p = .06$]. Although the low accuracy of these individuals made further examination of their RT data unwarranted, the difference between the participants who reached the criterion and those who did not is principally quantitative; both groups showed the pattern predicted by the view combination approach. The group that did not reach criterion during testing may thus represent participants who responded correctly on the final training block by chance and had not actually learned the stimuli to the same degree as had the participants who reached criterion.

In sum, the participants in Experiment 1 who met the training criterion responded to a novel interpolated view about as quickly and as accurately as they responded to the trained views. However, they were slower and less accurate in responding to novel views outside of the train-

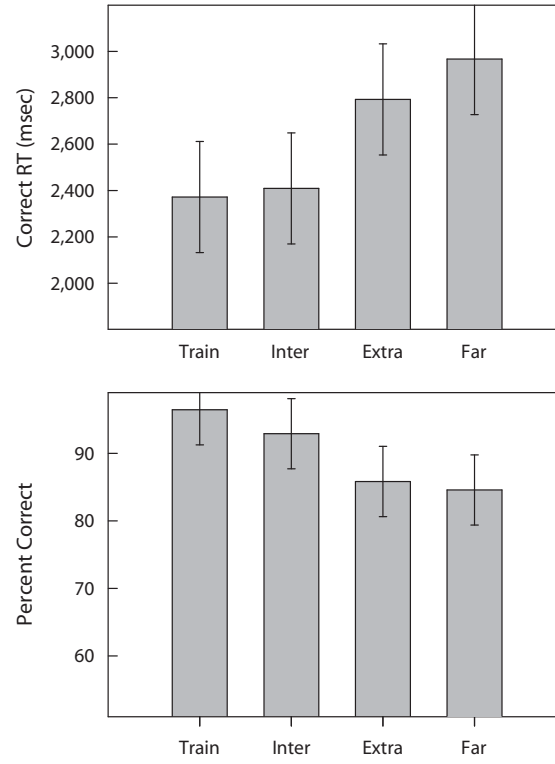

Figure 4. Response times (RTs; top) and accuracy (bottom) for the four perspectives from which the scene was viewed during testing in Experiment 1. Error bars are Loftus and Masson (1994) 95% confidence limits computed for within-subjects effects. Train, training; Inter, interpolated; Extra, extrapolated.

ing range than they were in responding to the interpolated novel views. The participants who did not meet the criterion showed the same pattern of accuracy as the people who did. Importantly, because of the manner in which the distractors were constructed for Experiment 1, there was no way to respond consistently correctly on the basis of the appearance of only one object in the scene. It is thus quite likely that the participants coded and used the spatial structure of the scene—particularly the relationships among objects—as a basis for their responses. Further evidence for the use of interobject relations comes from the observation that switch distractors (which do not disrupt spatial relationships) were generally more difficult to correctly identify than move distractors (which do disrupt spatial relationships). According to McNamara et al.'s (2006) speculation, the relatively abstract mental representations of the scenes (i.e., representations that coded relations in the depicted real-world scene, as opposed to a view of the stimulus itself) may have enabled the use of view combination mechanisms.

EXPERIMENT 2

The view combination approach to scene recognition predicts, as we found in Experiment 1, that some novel views may be recognized as efficiently as previously learned views. However, it also predicts that not all interpolated views necessarily receive this advantage. In particular, for a view combination process to support superior recognition of a novel viewpoint, it is critical that activation (or generalization) from the functional representations of the training views overlap to produce a predicted view that is similar to the to-be-recognized novel view. If the representations of the training views do not overlap in this sense, there should be relatively little advantage for a novel view that is between the two training views.

Superior recognition of a novel interpolated view under some conditions, but not under others, was demonstrated by Friedman et al. (2005), who conducted a comparative investigation of how pigeons and humans recognize objects and their photographs learned from two different viewpoints. The objects in their experiment were made of unique 3-D geometric shapes attached end-to-end at different angles. When the training stimuli were photographs taken from 60º apart, the pigeons showed evidence of having combined information from the two training views, insofar as their performance on the novel interpolated view was facilitated, relative to their performance on the novel extrapolated views. In contrast, when the training views were 90º apart, the pigeons showed no facilitation for interpolated views. Friedman et al. interpreted these data as implying that the process of generalizing between the training views can be expected to produce superior recognition of novel views only when there is overlap between the generalization functions of the training views; the *radial basis functions* that perform the mathematical interpolation between views in Edelman's (1999) model are functionally generalization gradients.

To date, no one has explicitly tested whether the difference between learning from viewing perspectives that are relatively close together or far apart changes the ability of people to use view combination mechanisms for recognizing either objects or scenes. As has previously been noted, Diwadkar and McNamara (1997) concluded that their data did not support view combination; however, an inspection of the data in their Figure 5 suggests that there was a difference in recognition performance to novel views that were between either close or far learned views. In particular, RTs to novel views that were between a relatively narrow range of studied views (i.e., novel views between 0º, 45º, and 90º) were shorter, on average, than those to novel views that were equally distant from studied views that had been further apart (i.e., novel views between 270º and 360º; see their Figure 5). This difference would be predicted by a view combination approach, but Diwadkar and McNamara did not directly analyze the difference between the two study ranges, nor was their experiment designed specifically to test this type of hypothesis.

In Experiment 2, we tested the hypothesis that if view combination processes work in scene recognition the way they do for objects, there should be a training angle far

Figure 5. Grayscale versions of the computer-generated target scene (top) from 0º and one of its two associated distractors (bottom) used in Experiment 2. The positions of the swings and the seesaw have been switched in the distractor, relative to their positions in the target.

enough apart that there would be insufficient activation from the learned views to support enhanced recognition of an interpolated view of either an object or a scene. To examine this prediction, in Experiment 2, we had two groups of participants learn a scene from two viewpoints. For the *narrow* group, we used a slightly smaller training angle than we did in Experiment 1: 48º. For the *wide* group, we used an angle of 96º. If the larger training angle we used in Experiment 2 does not support view combination, the pattern of responding across the two groups should differ. In particular, the narrow group should replicate the pattern observed in Experiment 1, but the wide group should show little difference in recognition performance between the interpolated and the extrapolated test views, and their performance on these views should be worse than performance on the training views.

In addition to comparing performance when participants learned a scene from relatively narrow or wide viewpoints, we also used Experiment 2 to examine whether the predictions of view combination would apply to stimuli that were computer-generated representations of a 3-D environment. The rationale for the new stimuli was both pragmatic and theoretical. Pragmatically, the use of computer-generated stimuli gave us increased precision and control over stimulus features, such as viewing angles and distances. We used shadows, lighting, and texture effects to enhance the realism of the scene (see Figure 5); however, the stimuli were clearly not as photorealistic as the stimuli used in Experiment 1. Because the work of McNamara et al. (2006) and Henderson and Hollingworth (2003) and the data from Experiment 1 support the idea that a relatively abstract representation is necessary for view combination, it would be interesting to determine whether computer-generated depictions of 3-D environments could be represented sufficiently abstractly to enable view combination. Generalization of the findings to these stimuli would thus demonstrate the robustness of view combination in scene recognition.

Finally, the results of Experiment 1 indicated that participants correctly rejected the move distractors more easily than the switches. Thus, to maximize the opportunity for the participants to code the spatial relations between objects in both the narrow and the wide conditions in Experiment 2, we used only switch distractors.

Method

Participants. Thirty volunteers (16 men and 14 women) from the University of Alberta pool participated. Six of these (2 men and 4 women) did not reach the accuracy criterion during testing, and their data were not considered further. The remaining 24 participants were randomly assigned to one of four groups. Two of the groups received their training with views that were 48º apart from each other (the *narrow* group), and two received their training with views that were 96º apart (the *wide* group).

Half the participants in the narrow group received training views at 0º and 48º and had interpolated, extrapolated, and far views at 24º, 72º, and 96º, respectively; the other half was trained with views at 48º and 96º and was tested with views at 72º (interpolated), 24º (extrapolated), and 0º (far). The same scheme was used for the wide groups, except that the views were −24º, 24º, 72º, 120º, and 168º (using the same arbitrary 0º reference point). One of these groups was trained with −24º and 72º, so their novel views were 24º (interpolated), 120º (extrapolated), and 168º (far); the other group was trained with 72º and 168º, so their novel views were 120º (interpolated), 24º (extrapolated), and −24º (far). Because of this counterbalancing scheme, the 24º view was both the interpolated and the extrapolated view for both the narrow and the wide angle training groups. As in Experiment 1, this controlled for the possibility that any idiosyncrasies across these two views would be balanced across the critical interpolated, extrapolated, narrow, and wide conditions and could not be responsible for any observed differences in performance.

Stimuli. The stimuli for Experiment 2 consisted of 24 color digital images (eight targets and 16 distractors) of a playground scene that was generally similar to the scene used in Experiment 1. The stimuli were generated from a 3-D computer model, using 3-D Studio Max, and employed lighting effects, shadows, and textures to enhance the realism of the scene. The eight target stimuli depicted the scene from viewing angles of −24º, 0º, 24º, 48º, 72º, 96º, 120º, and 168º. For each of the targets, we constructed two switch distractors, each of which exchanged the positions (but not the orientations) of two of the central objects in the foreground of the scene. Across all distractors, each of the six central objects (swings, merry-go-round, jungle gym, teeter-totter, slide, and sandbox) was switched approximately an equal number of times. An example of one of the targets and distractors is shown in Figure 5. Again, the central objects were fully visible from each of the viewpoints. We used different learning angles than we had used in Experiment 1 because basing the narrow condition stimuli on 60º differences between views would have yielded a "far" view in the wide angle condition that was actually closer—in the other direction—to one of the two learned views than it was to the extrapolated view.

Procedure. The procedure was similar to that used in Experiment 1, with the following exceptions. First, we changed the feedback during training so that the participants received 1 point for correct responding and 2 points for correct responding that was faster than a 2-sec criterion (although the exact value of the criterion was unknown to the participants). The point system again had no tangible consequences for the participants and was used merely to encourage accurate and fast responses. Second, the participants received a minimum of five training blocks before they were moved to the test phase. Third, we increased the accuracy criterion to 90% correct on the training views during the test trials. These three changes were made in order to try to ensure that the training views were well learned; we used the same 90% criterion as Edelman et al. (1999) and approximately the same number of training trials. Fourth, to better equate the pace of the training and test trials, we inserted a 2-sec intertrial interval during both training and testing.

Results and Discussion

As in Experiment 1, we trimmed the correct RTs that were longer than three standard deviations above the means computed individually for each participant separately for the target and distractor test conditions. The omitted trials constituted 0.8% of the targets and 0.3% of the distractors.

Response times. The correct RTs during the test trials for the individuals who reached the 90% criterion on the training views were submitted to a viewpoint (training, interpolated, extrapolated, or far) × training condition (narrow or wide) ANOVA. The main effect of training condition [$F(1,22) = 18.92$, $MS_e = 4{,}597{,}367.47$] reflected the fact that the group who received the wide training angle took more than twice as long to respond as the group who received the narrow training angle (3,796 vs. 1,892 msec, respectively). There was also a main effect of viewpoint [$F(3,66) = 8.33$, $MS_e = 309{,}867.21$] and, importantly, a training condition × viewpoint interaction [$F(3,66) =$

3.05, MS_e = 309,867.21]. Figure 6, top panel, illustrates this interaction.

Planned contrasts using the error variance from ANOVAs conducted separately on each training condition showed that, for the narrow group, there was no significant difference in RT between the training and the interpolated conditions (43 msec; 1.2% of the variance in the main effect of view; $F < 1$) but that there was a significant difference between the interpolated and the extrapolated conditions [251 msec; 40.7% of the variance in the main effect of view; $F(1,11) = 7.83$, MS_e = 96,502.82]. Thus, for this group, the results of Experiment 1 were replicated. In contrast, the group that received the wide training angle had the opposite results: There was a relatively large difference between the training and the interpolated conditions [814 msec; $F(1,11) = 7.15$, MS_e = 1,112,609.99], accounting for 41.2% of the variance in the main effect of view for that group, but no significant difference between the interpolated and the extrapolated views [168 msec; 0.17% of the variance; $F(1,11) < 1$]. Thus, the pattern of means across training groups extends Friedman et al.'s (2005) findings for object recognition by pigeons to the domain of human scene recognition.

Accuracy. Figure 6, bottom panel, shows the mean percentage correct as a function of viewpoint and training condition. There was a main effect of viewpoint [$F(3,66) = 3.82$, MS_e = 176.09], but there was no main effect of training condition [$F(1,22) < 1$] and no interaction between training condition and viewpoint [$F(3,66) < 1$]. The planned contrasts between the training views and the interpolated view approached significance for the narrow condition [$F(1,11) = 3.28$, MS_e = 533.90, $p = .10$] and for the wide condition [$F(1,11) = 3.64$, MS_e = 206.63, $p = .09$]. However, the contrasts between the interpolated and the extrapolated views were not significant for either training condition [$F(1,11) < 1$ and $F(1,11) = 1.21$, respectively]. Both groups performed sufficiently accurately that there does not appear to have been a speed–accuracy trade-off among the different test conditions.

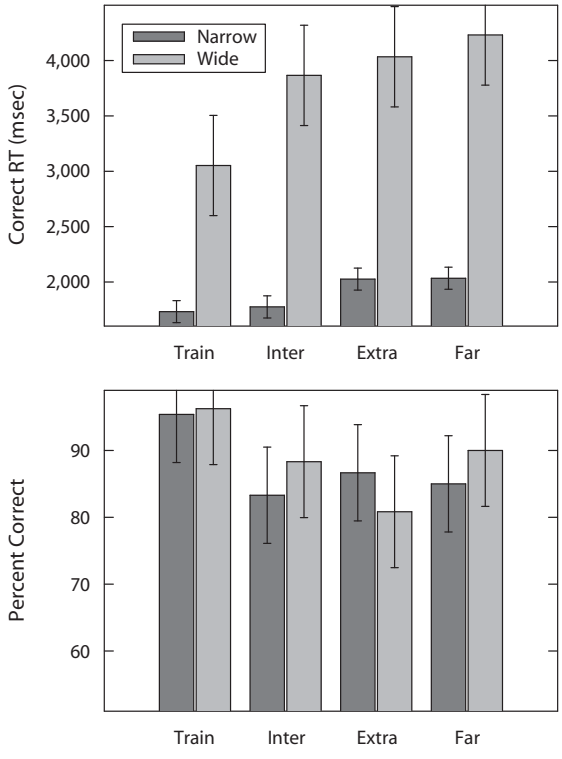

Figure 6. Response times (RTs; top) and accuracy (bottom) for the two training conditions, across the four perspectives from which the scene was viewed during testing in Experiment 2. Error bars represent Loftus and Masson (1994) 95% confidence limits computed for within-subjects effects.

For the 6 participants who did not reach the 90% accuracy criterion, the mean percent correct for the training, interpolated, extrapolated, and far conditions were 73.3%, 75.0%, 66.7%, and 63.3%, respectively. There were too few participants (i.e., only 4 in the narrow condition and 2 in the wide condition) to conduct a meaningful analysis of these means.

In sum, the participants in Experiment 2 who were trained with two views of a scene that were relatively close together recognized novel interpolated views about as quickly as they recognized the training views and were significantly slower to recognize novel extrapolated views. In contrast, the participants in the wide condition responded significantly more slowly to both the interpolated and the extrapolated views than to the training views. That view combination was evident in the narrow angle group of Experiment 2 with computer-generated stimuli holds promise for the use of such stimuli in future investigations of these effects.

GENERAL DISCUSSION

The overall pattern of results obtained across our experiments is consistent with a growing body of evidence that people and other species recognize single objects by combining information from two or more stored representations (Bülthoff & Edelman, 1992; Edelman, 1999; Edelman et al., 1999; Friedman et al., 2005; Spetch & Friedman, 2003; Spetch et al., 2001). Our results extend these findings about object recognition by providing evidence that view combination mechanisms also underlie the recognition of scenes. In particular, our participants recognized novel views of a scene that were between two previously learned views better than they recognized views that were outside of the training range but equally distant from at least one of the trained views. Notably, the results for the wide training angle group in Experiment 2, in which the two training views did not produce superior recognition of an interpolated novel view, are also consistent with the view combination approach, because they reinforce the idea that for activation from more than one view to be effective, there has to be some degree of overlap in the functional representations of the learned views. The absolute amount that is "too far" for view combination mechanisms to contribute to superior recognition is likely to vary as a function of the complexity of the objects or

scenes involved; indeed, the type of object and scene parameters that affect this mechanism requires systematic, and likely parametric, investigation.

It is clear that not all the participants found our recognition task to be easy. This is true of both the present study and of previous studies of the ability to combine views of single objects (e.g., Bülthoff et al., 1995) or of scenes (e.g., Hollingworth & Henderson, 2004; Hock & Schmelzkopf, 1980). Even so, the participants in the present study who did not reach the accuracy criterion that would enable meaningful examination of their RT data showed the same overall pattern of accuracy as the participants who did reach the criterion; it is thus possible that the people who did not reach criterion had been moved out of the learning phase by chance, before they had learned the targets as well as the remaining participants had.

Given the present results, there are several possible explanations for why Diwadkar and McNamara's (1997) experiments failed to exhibit effects that were entirely consistent with view combination. One possibility involves methodological differences between our experiments and theirs. In particular, Diwadkar and McNamara's procedure involved a sequential set of training blocks in which the participants first viewed the scene repeatedly from only one perspective and were required to name the objects as a list to demonstrate their learning. They subsequently learned to discriminate between that original training view and three additional training views before proceeding to the test phase. It is possible that the requirement to list the objects reduced the participants' tendency to encode the relative angles and distances between them. It is also possible that the relative primacy of one particular view during training resulted in an overreliance on it at test (see also Shelton & McNamara, 1997, who used a similar procedure to similar effect). If one view was better learned than the others, this could explain the monotonic pattern of recognition performance that Diwadkar and McNamara interpreted as normalization. By contrast, the procedures that we used required the participants to learn and discriminate between two target views and several very similar distractors throughout the training phase. Thus, this uniform distribution of viewing angles throughout training, as well as the subtleties of the discrimination between the targets and the distractors, may have led to encoding and recognition processes that relied on more than one training view (i.e., view combination).

A theoretically more interesting set of reasons for the observed evidence for view combination obtained in the present study follows from McNamara et al.'s (2006) speculation about a dual-level representation for scenes. Several aspects of our stimuli may have encouraged the participants to form relatively more abstract mental representations than participants were able to with the stimuli used by Diwadkar and McNamara (1997), and as was discussed previously, abstract representations may foster the use of view combination. First, our distractors encouraged the participants to encode the spatial relations between objects; it was virtually impossible to respond correctly otherwise. Representations that directly encode interobject relations contain information that is invariant over changes in viewpoint and are thus naturally conceptualized as being relatively abstract. Second, our stimuli depicted the scenes from a ground-level perspective. A ground-level viewpoint provides different kinds of depth cues (e.g., occlusion) that may not have been as available in the bird's-eye views used by Diwadkar and McNamara. It is possible that view combination mechanisms are especially sensitive to such depth cues. Third, in complex real-world scenes, the background changes across viewing perspectives, thus providing additional relational and contextual cues to use in forming the original memory representations and in generalizing to novel views. Fourth, there were more objects in the real (and computer-simulated) playground scenes than in Diwadkar and McNamara's stimulus array, both in the central tableaux itself and in the background; background objects provide distal cues from different perspectives that may be important in combining views of scenes. Fifth, the objects in real-world scenes may be very similar (e.g., trees, buildings) or even identical (e.g., some of the playground equipment) to one another, which should reduce the ability to remember the scene by using verbal cues but should enhance (or enforce) the ability to remember it via visual cues and interobject spatial relations.

Finally, the stimuli in the present experiments depicted a semantically coherent and, thus, generally familiar scene, as contrasted with an array of unrelated objects. Unrelated objects, by definition, have few prior associations in memory and may, therefore, require additional mental resources to memorize. These resources may have detracted from those needed to integrate between training views. Furthermore, there is evidence that the presence of semantic and functional relations among objects in a scene leads to relatively abstract (object-level) coding (Carlson-Radvansky & Radvansky, 1996) and that the presence of semantically related objects, together with scene features such as a horizon line and other distal information, may activate abstract scene schemas that facilitate comprehension and provide expectations about what objects should be in the scene and where they might be located (Friedman, 1979; Mandler & Parker, 1976). Thus, even though the individual objects in the playground scenes of the present experiments were somewhat haphazardly arranged, similar to those in Diwadkar and McNamara's (1997) tabletop display, the semantic and functional relatedness of the objects in the playground scenes may have contributed to the ease of encoding the interobject spatial relations between them and, thus, to the effective use of view combination mechanisms. To summarize, the quantity and nature of the interitem relations available from different views of a real-world scene may very well contribute to the quality and abstractness of its representation and, thus, to how well novel views of it can be recognized. Although the present data do not allow us to determine which one or more of the possibilities we have described was responsible for the obtained results, each possibility has theoretical consequences, for both the representation and the processing of scenes, that are worthy of further investigation.

These considerations about the role of interobject, functional, and semantic relations in forming abstract scene

representations beg the question of what a *scene* really is. At one extreme, it is conceivable that a scene is anything in front of one's eyes. According to this conceptualization, inventories of unrelated objects and well-organized pictures of coherent real-world scenes would be psychologically equivalent. However, this equivalence has little empirical support. For example, even with real-world scenes, not all objects are equally well encoded or remembered (Friedman, 1979; Loftus & Mackworth, 1978; Mandler & Parker, 1976; Mandler & Stein, 1974), and as was noted earlier, memory for a group of semantically related objects can be disrupted by removing features (the horizon or their relative locations) that make the objects cohere as a scene. The stimuli in the present study were closer to the semantically well-organized end of the continuum, whereas Diwadkar and McNamara's (1997) were closer to an inventory. It is possible that our conclusions apply solely or primarily to well-organized scenes; yet any theory of scene processing will undoubtedly have to explain performance at both extremes.

Our general conclusion that scene recognition is supported by view combination mechanisms is reminiscent of several contemporary findings about how people perceive and attend to single views of a scene. For example, Henderson and Hollingworth (2003; see also Hollingworth & Henderson, 2004) have shown that people are surprisingly insensitive to viewpoint rotations in depth if the rotations are introduced gradually and incrementally. These authors have suggested that visual memory is implicitly updated, on the basis of the most recent view of the scene. Similarly, Intraub and her colleagues (Gottesman & Intraub, 2003; Intraub, Gottesman, & Bills, 1998; Intraub, Gottesman, Willey, & Zuk, 1996; Intraub & Richardson, 1989) have explained the phenomenon of *boundary extension*—the tendency to remember a more expansive view of a scene than what one has actually been shown—in terms of the activation of perceptual schemata that are based on the anticipation of what a scene's perceptual structure will look like from one saccade to the next. In both of these cases—as with view combination—perceptual or memorial information is actively transformed into a representation that is more abstract than what is actually experienced.

View combination mechanisms that underlie scene recognition are consistent with the long-standing view that extended experience with an environment enables one to form a mental representation of it that is more general than one's egocentric experiences during learning (Newcombe & Huttenlocher, 2000; O'Keefe & Nadel, 1978; Shemyakin, 1962; Siegel & White, 1975; Tolman, 1948). By demonstrating that information from different encounters with an environment is combined to facilitate efficient scene recognition, the present results may provide a substantial boost toward understanding how flexible mental representations of space are stitched together from separate experiences.

AUTHOR NOTE

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. We thank Eric Hodgson, Geoff Hollis, and Bernd Kohler for assistance with preparing the stimuli, programming, and conducting the experiments. We also thank Yvonne Lippa for helpful comments on previous drafts and Laura Carlson for her insightful comments. Correspondence should be addressed to A. Friedman, Department of Psychology, University of Alberta, Edmonton, AB, T6G 2E9 Canada (e-mail: alinda@ualberta.ca) or D. Waller, Department of Psychology, Miami University, Oxford, OH 45056 (e-mail: wallerda@muohio.edu).

REFERENCES

Biederman, I. (1987). Recognition by components: A theory of human image understanding. *Psychological Review*, **94**, 115-147.

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception & Performance*, **19**, 1162-1182.

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

Bülthoff, H. H., Edelman, S., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex*, **5**, 247-260.

Carlson-Radvansky, L. A., & Radvansky, G. A. (1996). The influence of functional relations on spatial term selection. *Psychological Science*, **7**, 56-60.

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. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, **108**, 316-355.

Friedman, A., & Harding, C. A. (1990). Seeing versus imagining movement in depth. *Canadian Journal of Psychology*, **44**, 371-383.

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.

Gottesman, C. V., & Intraub, H. (2003). Constraints on spatial extrapolation in the mental representation of scenes: View-boundaries vs. object-boundaries. *Visual Cognition*, **10**, 875-893.

Henderson, J. M., & Hollingworth, A. (2003). Global transsaccadic change blindness during scene perception. *Psychological Science*, **14**, 493-497.

Hock, H. S., & Schmelzkopf, K. F. (1980). The abstraction of schematic representations from photographs of real-world scenes. *Memory & Cognition*, **8**, 543-554.

Hollingworth, A., & Henderson, J. M. (2004). Sustained change blindness to incremental scene rotation: A dissociation between explicit change detection and visual memory. *Perception & Psychophysics*, **66**, 800-807.

Intraub, H., Gottesman, C. V., & Bills, A. (1998). Effects of perceiving and imagining scenes on memory for pictures. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 186-201.

Intraub, H., Gottesman, C. V., Willey, E. V., & Zuk, I. J. (1996). Boundary extension for briefly glimpsed photographs: Do common perceptual processes result in unexpected memory distortions? *Journal of Memory & Language*, **35**, 118-134.

Intraub, H., & Richardson, M. (1989). Wide-angle memories of close-up scenes. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 179-187.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception & Performance*, **4**, 565-572.

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

Mandler, J. M., & Parker, R. E. (1976). Memory for descriptive and spatial information in complex pictures. *Journal of Experimental Psychology: Human Learning & Memory*, **2**, 38-48.

MANDLER, J. M., & STEIN, N. L. (1974). Recall and recognition of pictures by children as a function of organization and distractor similarity. *Journal of Experimental Psychology*, **102**, 657-669.

MARR, D. (1982). *Vision*. San Francisco: Freeman.

MCNAMARA, T. P., DIWADKAR, V. A., BLEVINS, W. A., & VALIQUETTE, C. M. (2006). Representations of apparent rotation. *Visual Cognition*, **13**, 273-307.

NEWCOMBE, N. S., & HUTTENLOCHER, J. (2000). *Making space: The development of spatial representation and reasoning*. Cambridge, MA: MIT Press.

O'KEEFE, J., & NADEL, L. (1978). *The hippocampus as a cognitive map*. Oxford: Oxford University Press, Clarendon Press.

SHELTON, A. L., & MCNAMARA, T. P. (1997). Multiple views of spatial memory. *Psychonomic Bulletin & Review*, **4**, 102-106.

SHEMYAKIN, F. N. (1962). Orientation in space. In B. G. Anan'yev et al. (Eds.), *Psychological science in the USSR* (Vol. 1, Pt. 1, pp. 186-255). Washington, DC: U. S. Office of Technical Reports.

SIEGEL, A. W., & WHITE, S. H. (1975). The development of spatial representations of large-scale environments. In H. Reese (Ed.), *Advances in child development and behavior* (Vol. 10, pp. 10-55). New York: Academic Press.

SPETCH, M. L., & FRIEDMAN, A. (2003). Recognizing rotated views of objects: Interpolation versus generalization by humans and pigeons. *Psychonomic Bulletin & Review*, **10**, 135-140.

SPETCH, M. L., FRIEDMAN, A., & REID, S. L. (2001). The effect of distinctive parts on recognition of depth-rotated objects by pigeons (*Columba livia*) and humans. *Journal of Experimental Psychology: General*, **130**, 238-255.

TARR, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review*, **2**, 55-82.

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

TOLMAN, E. (1948). Cognitive maps in rats and men. *Psychological Review*, **55**, 189-208.

ULLMAN, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, **32**, 193-254.

ULLMAN, S. (1996). *High-level vision: Object recognition and visual cognition*. Cambridge, MA: MIT Press.

(Manuscript received March 8, 2007;
revision accepted for publication September 27, 2007.)