Vision Research 50 (2010) 202–210

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Facilitation by view combination and coherent motion in dynamic object recognition

Alinda Friedman [a,*], Quoc C. Vuong [b], Marcia Spetch [a]

[a] Department of Psychology, University of Alberta, Edmonton, Alberta, Canada T6G 2E9
[b] Institute of Neuroscience, School of Psychology, Newcastle University, Newcastle, England, NE2 4HH, UK

ARTICLE INFO

Article history:
Received 18 August 2009
Received in revised form 30 October 2009

Keywords:
Object recognition
Motion
Viewpoint interpolation
View combination
Similarity in object recognition

ABSTRACT

We compared the effect of motion cues on people's ability to: (1) recognize dynamic objects by combining information from more than one view and (2) perform more efficiently on views that followed the global direction of the trained views. Participants learned to discriminate two objects that were either structurally similar or distinct and that were rotating in depth in either a coherent or scrambled motion sequence. The Training views revealed 60° of the object, with a center 30° segment missing. For similar stimuli only, there was a facilitative effect of motion: Performance in the coherent condition was better on views following the training views than on equidistant preceding views. Importantly, the viewpoint between the two training viewpoints was responded to more efficiently than either the Pre- or Post-Training viewpoints for both the coherent and scrambled condition. The results indicate that view combination and processing coherent motion cues may occur through different processes.

1. Introduction

To interact in a dynamic environment, humans and other active animals must be able to encode and recognize objects under conditions in which both the objects and the viewing conditions may be changing. For example, shadows and light change with time of day and the relative positions of observers and objects may also change over time (e.g., a moving observer or a moving object). The net result of these changes is that the same three-dimensional (3D) object projects different shapes and surface details onto the two-dimensional (2D) retinal array. Thus, the input to the visual recognition system contains both static information (e.g., the 2D shape at a particular moment) and dynamic information (e.g., how that 2D shape changes over time). The role of static cues in object recognition has been investigated extensively (e.g., Biederman, 1987; Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992; Friedman, Spetch, & Ferrey, 2005; Peissig, Wasserman, Young, & Biederman, 2002; Spetch & Friedman, 2003; Spetch, Friedman, & Reid, 2001; Tarr, Bülthoff, Zabinski, & Blanz, 1997), but it is only recently that the role of motion cues has begun to be understood (e.g., Kourtzi & Nakayama, 2002; Kourtzi & Shiffrar, 1999, 2001; Vuong & Tarr, 2004, 2006; Wallis & Bülthoff, 2001), especially in terms of comparisons across species (Cook & Katz, 1999; Cook & Roberts, 2007; Cook, Shaw, & Blaisdell, 2001; Friedman, Vuong, & Spetch, 2009; Loidolt, Aust, Steurer, Troje, & Huber, 2006; Spetch, Friedman, & Vuong, 2006; Spetch et al., 2001).

A *view combination* framework has been able to account for object recognition across species under a variety of circumstances (Bülthoff & Edelman, 1992; Edelman, 1999; Spetch & Friedman, 2003; Friedman et al., 2005, 2009; Spetch et al., 2001; Ullman, 1998). Briefly, view combination is a kind of generalization (cf. Shepard, 1987) in which a novel input view activates all of the stored representations to which it is similar. Therefore, view combination is construed as a process that relies on the existence of object representations in long-term visual memory. A new representation is constructed from the activation of multiple stored representations and if the constructed representation is sufficiently similar (over a threshold) to the novel input image, the novel image is recognized. The behavioral signature of view combination is twofold: First, a novel view that is between two training views (an *Interpolated* view) can be responded to more efficiently than an *Extrapolated* novel view outside of that range by an equivalent distance (e.g., viewpoint 4 in Fig. 1 vs. viewpoints 2 and/or 6 or 1 and/or 7), and second, the Interpolated view can often be responded to about as efficiently as the Training views. Thus far, we have found evidence for view combination with static images (Friedman et al., 2005; Spetch et al., 2001) and more recently, with dynamic objects (Friedman et al., 2009; see also Bülthoff & Edelman, 1992). In Friedman et al. (2009), which compared humans and pigeons, we found that both species showed view combination with both scrambled and coherent motion, but it was stronger for coherent than scrambled motion in humans. We also found an

[*] Corresponding author. Address: Department of Psychology, University of Alberta, Edmonton, Alberta, Canada T6G 2E9. Fax: +1 780 492 1768.
E-mail address: alinda@ualberta.ca (A. Friedman).

0042-6989/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2009.11.010

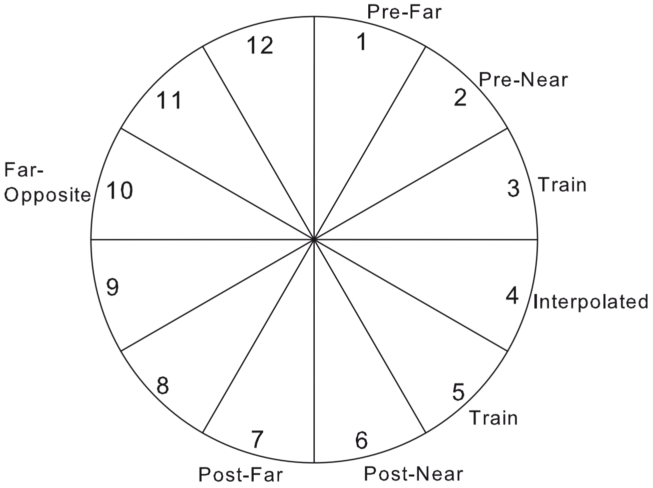

Fig. 1. Schematic diagram of the top view of the different viewpoint conditions for half of the participants. The images were rendered as if the object were in the center of a sphere and the camera was placed at the horizontal great circle of the sphere.

effect of motion; the views that followed the training views were responded to more quickly than those which preceded the training views by an equal amount. Thus, we concluded that view combination models might have to be modified to take account of effects of coherent motion.

In the present study we explored the recognition of dynamic objects further by investigating two potentially different ways of facilitating recognition, one involving view combination and the other involving motion per se. First, we examined whether the perception of smooth coherent motion would enhance view combination effects relative to when the motion was not coherent but the same amount of structure was learned (e.g., Friedman et al., 2009; Kourtzi & Shiffrar, 1999). Second, we examined whether view combination mechanisms observed with dynamic objects are separable from motion-specific mechanisms (Freyd, 1987; Friedman et al., 2009; Stone, 1999; Vuong & Tarr, 2004). To address these goals, we extended our previous study on view combination with dynamic objects in humans (Friedman et al., 2009, Experiment 1a) and we changed the training procedure in order to enhance humans' perception of smooth coherent motion. The procedural changes we made were similar to those shown to affect how pigeons used coherent motion even though the motion itself was not discriminative of object identity (Friedman et al., 2009; Experiments 1b and 2). The finding with pigeons is notable because pigeons are highly sensitive to coherent motion cues when they are discriminative (e.g., Spetch et al., 2006). The changes we made to enhance pigeons' perception of motion during training included, for example, "sweeping" across two different 30° segments of the objects in their correct sequence instead of presenting the segments in a random order.

Previous studies using paradigms that tap relatively short-term representations (e.g., representational momentum (Freyd, 1987); priming; apparent motion (Kourtzi & Shiffrar, 1999, 2001)) have found evidence for both view combination and motion-specific mechanisms. For example, Kourtzi and her colleagues (Kourtzi & Nakayama, 2002; Kourtzi & Shiffrar, 1999, 2001) tested human participants with a short-term visual priming paradigm (Sekuler & Palmer, 1992), in which the primes either induced apparent motion or were static images. The primary task was to decide if two shapes presented simultaneously 500 ms after the primes were the same or different with respect to each other. The primes were therefore irrelevant to the main task; the rationale was that if there was facilitation on the same-different judgment, then the orientation of the target shapes must have been primed and by implication, must have had a representation (however brief). The orientation of the target stimuli was either: (1) identical to either of the primes (i.e., to the "learned" views), (2) at a novel angle that was within the rotation angle that separated the two primes (i.e., an "Interpolated" view), or (3) at a novel angle outside the rotation angle that separated the two primes by an equivalent amount of degrees to the second condition (i.e., an "Extrapolated" view).

When the angular disparity of the two primes was small (60º), Kourtzi and Shiffrar (1999; Experiment 1) found significant priming of the Interpolated view for both static and apparent motion conditions, but surprisingly, significant priming of Extrapolated views was found only in the static condition (see their Fig. 3). By comparison, when the angular disparity was large (120°), they found significant priming of the Interpolated view only in the apparent motion condition, and they did not find any significant priming of Extrapolated views for either the static or apparent motion conditions (see also Kourtzi & Nakayama, 2002; Kourtzi & Shiffrar, 2001, who found similar results for 2D objects rotating in the picture plane or deforming objects). Based on these results, Kourtzi and Shiffrar suggested that the visual system can take advantage of an object's motion path and link views within the path of motion (i.e., Interpolated views) even across relatively large rotational differences, possibly creating short term "virtual" views that could facilitate recognition (see Vetter & Poggio, 1994). Consistent with this interpretation, a recent functional imaging study by Weigelt, Kourtzi, Kohler, Singer, and Muckli (2007) showed as much adaptation of the hemodynamic response in object-selective temporal cortical regions to the Interpolated view as to repeated views but no adaptation to Extrapolated views. This hemodynamic adaptation suggests that the brain had formed a representation of the unseen Interpolated view. This newly formed representation may be why unseen views falling within the object's path are efficiently responded to. Kourtzi and colleagues' priming effects are intriguing because they are consistent with the idea that view combination mechanisms also work in short-term visual memory with objects undergoing coherent apparent motion.

In our long-term memory studies with static objects (Friedman et al., 2005), we found a slightly different pattern of results from what Kourtzi and Shiffrar (1999) found with apparent motion in short-term memory. In particular, view combination occurred for pigeons trained with small but not large angles between views (e.g., 60° vs. 90°) and there was no facilitation for Extrapolated views at either angle. Using a similar long-term memory paradigm with humans, the same pattern of results was observed for static views of real-world scenes (the training angles were either 48° or 96° apart; Friedman & Waller, 2008). Thus, it appears that view combination with static stimuli can occur within a wide range of angles in short-term memory and within a narrower range in long-term memory. The present study is an effort to further differentiate effects of view combination and coherent motion in a long-term memory paradigm with objects undergoing real, rather than apparent, motion (e.g., Friedman et al., 2009).

Strictly speaking, view combination should occur with both static and moving objects that are represented in long-term memory, including dynamic objects in which the motion is not coherent (e.g., objects undergoing scrambled motion). Kourtzi's results imply that the range over which view combination is effective might be extended with dynamic objects undergoing coherent motion, whether the motion is real or apparent. Furthermore, Wallis and Bülthoff (2001) showed that coherent apparent rotations of face views were more likely to be linked than the same face stimuli shown in a scrambled order (thereby disrupting the spatio-temporal integrity of the apparent rotation, but controlling for the total amount of an object's structure that was seen; see also Harman & Humphrey, 1999). Together, these results point to the possibility that view combination may be enhanced by coherent motion. In contrast to this expectation, in our previous work with dynamic objects (Friedman et al., 2009), humans showed an equivalent magnitude of view combination in response times for *both* coherent and scrambled motion conditions. However, there was an accuracy difference for the Interpolated view that favored coherent motion. Thus, the evidence for an enhancement of view combination via coherent motion was mixed.

The training angles in our previous study were relatively close (they spanned 60° on either side of the Interpolated view) but there were other aspects of the design, such as presenting the 30° motion segments out of their global order (see below), that may have made it more difficult to obtain differential effects of motion type (coherent vs. scrambled) on response time with respect to view combination. For example, we used the same segments as are shown in Fig. 1, but trained with segments 2, 3, 5, and 6 (or their counterparts on the other half of the circle). These segments were presented individually on each training trial; consequently, the global sequence was never seen in order. By training with one segment at a time, we may have inadvertently weakened the effects of global coherent motion and strengthened the Interpolated view via motion generalization from the preceding training view (in addition to view combination). We correct those issues in the present study to examine whether robust view combination still occurs for both coherent and scrambled motion conditions in a long-term memory paradigm.

To investigate the separate effects of motion per se, independently of effects of view combination, we took advantage of a phenomenon reported by Freyd (1987), which she called *representational momentum* (see also Freyd & Finke, 1984, 1985). In studies of representational momentum, under some circumstances an observer's memory for the final position of an object is distorted in the direction the apparent motion. For example, Freyd and Finke (1984) presented participants with a static figure presented at three different orientations along a path of rotation in the image plane; each figure was separated from the next by a 250 ms interstimulus interval, so that the conditions existed for there to be apparent motion. Observers found it more difficult to decide that a fourth test stimulus that had been slightly rotated forward in the implied direction of apparent motion was different than the actual third stimulus that they had seen, compared to a test stimulus that was slightly rotated backwards in the opposite direction. This discrimination cost for forward rotation disappeared when the three static frames were presented in a scrambled order that did not lead to the percept of apparent motion. Thus, representational momentum has characteristics similar to both the effects of apparent motion on priming (e.g., Kourtzi & Shiffrar, 1999) and of view combination on recognition (e.g., Friedman et al., 2005), in that some novel views are processed more efficiently than other novel views. However, at least empirically, the two mechanisms are different in short-term visual memory because view combination favors views within the training range, whereas representational momentum favors views outside of that range.

Freyd and Finke's (1984) finding provides evidence that people are sensitive to an implied direction of motion; presumably they would be similarly sensitive to implied views of objects that were actually undergoing coherent motion. Vuong and Tarr (2004) confirmed this expectation. They found that people responded "same" more quickly to novel views of objects that continued the trajectory of an object rotating in depth than they were to respond to views that preceded that rotation trajectory. They argued that it was possible that motion provides the ability to predict the structure or appearance of upcoming views, thereby biasing novel views in the direction of rotation (see also Stone, 1999). However, it was not clear whether observers generated and then represented in memory the predicted "virtual" view that followed the endpoint of the dynamic object's motion. The priming results from Kourtzi and colleagues suggest that Interpolated views generated by motion may be represented; thus, views implied by an object's direction of motion may also be represented in short-term visual memory.

Kourtzi and her colleagues, Freyd and her colleagues, and Vuong and Tarr all examined the role of motion (real or apparent) in the context of short-term visual memory. Together, their results suggest that motion can broaden the tuning functions of studied views (e.g., Kourtzi & Nakayama, 2002), which may affect view combination mechanisms. Their data also suggest that that motion may distort or otherwise bias views in the direction of motion (e.g., Freyd, 1987; Vuong & Tarr, 2004). In the present study, we used movies of novel 3D objects rotating in depth to examine both view combination and motion predictions in a long-term memory paradigm. As noted above, the behavioral signature that indicates a facilitative role for coherent motion is that performance on the sequence of views that follows the trained views (even though that sequence is outside of the training range) should be better than performance on the sequence that precedes the trained views. We will refer to this pattern of responding to dynamic cues as the motion effect.

In our previous research on dynamic object recognition in humans and pigeons (Spetch et al., 2006), shape and motion cues were redundant on training trials, so both cues were discriminative of identity. With structurally similar objects, pigeons could rely on dynamic cues alone to perform accurately but humans maintained a reliance on shape cues (Spetch et al., 2006). In a second series of studies (Friedman et al., 2009) we examined what happened when motion cues were not discriminative of identity. Humans and pigeons were trained to recognize a dynamic object from two different perspectives. In the first experiments (Friedman et al., 2009, 1a and 1b), a full 1/3 of the objects were shown on the training trials (e.g., viewpoints 2, 3, 5, and 6 in Fig. 1). Both species showed clear evidence for view combination, but only humans showed facilitation for novel views that continued the rotation trajectory. We reasoned that pigeons might require better global motion cues to take advantage of motion when it was not discriminative. That is, the four training viewpoints in Experiments 1a and 1b were presented one at a time in a random order; for Experiment 2, which tested only pigeon subjects, we instead presented a "sweep" of two viewpoints in their proper sequential order for the coherent motion condition (e.g., in Fig. 1, viewpoint 3 followed by viewpoint 5, omitting viewpoint 4) and randomized across the same two viewpoints for the scrambled condition. With this procedure, the pigeons now showed both a view combination effect and a facilitatory effect of motion. Because sweeping across two viewpoints of the objects was effective in eliciting about a 10% advantage in accuracy for coherent over scrambled motion in pigeons, and this presentation method also yielded a significant motion effect, we examined whether the same kind of manipulation would magnify the efficacy of dynamic cues for humans and possibly dissociate view combination and motion effects.

To summarize, in the present study, participants learned to discriminate to a criterion between two similar or distinctive objects that were displayed from two viewpoints. The objects were either moving smoothly through a 90° arc that omitted 30° of its structure in the center of the arc, or the same amount of structure was displayed in a scrambled manner. Participants were then tested with dynamic stimuli at the trained viewpoints as well as at novel viewpoints that were inside of the training range or outside of it by distances that were either directly adjacent to the trained distances by the same amount as the Interpolated view or were further away from the training views (see Fig. 1). In both training conditions (coherent and scrambled), motion was not a discriminative cue to object identity. Nevertheless, we expected to observe effects of view combination here and, as we saw in our previous study, we expected to find these effects in both the coherent and scrambled conditions; after all, view combination was originally meant to explain static object recognition. However, if coherent motion enhances view combination by, for example, broadening the tuning functions of the representations, then the view combination effect should be larger for coherent motion than for scrambled motion. In addition, we expected to find a Pre-Post effect – that is, facilitation for viewpoints following the rotation direction – only in the coherent motion condition. The two effects together may translate into a main effect for motion type (coherent vs. scrambled).

It is important to document motion effects in a long-term memory paradigm, because doing so can help rule out recency effects as a principle cause of any differences observed between the efficiency of processing views that precede the first training view and those that follow the second (and most recent) view. In particular, the present and previous experiments used a discrimination learning paradigm (e.g., Friedman et al., 2009) in which all the training views were trained to the same criterion. Thus, if there is no effect of motion, there should be as much generalization to views that precede the first training view as to those that follow the second training view, and performance should be equal for the two cases (Pre vs. Post). We did not think this was likely in the present case because, in our previous work in which recency effects were controlled (Vuong & Tarr, 2004), there was a still evidence for an effect similar to "representational momentum". Nevertheless, from an empirical point of view, a pure recency explanation of facilitation for views that follow the second training view in the sequence predicts that those views should be responded to more accurately and/or more quickly than the Interpolated viewpoint (which precedes the second training view). The present design allows us to test this prediction in addition to the predictions of view combination and motion per se.

2. Method

2.1. Participants

There were 45 volunteers (16 male, 29 female) from the University of Alberta participant pool who received partial course credit as well as performance-based payment for their participation. They were randomly assigned to one of 8 training conditions formed by the combination of type of motion type (scrambled or coherent), whether the stimulus pair that was learned first was distinctive or similar, and which of two sets of particular views were used as the training and test movies (see below). The data from five female participants were not considered further; in one case performance was at chance throughout testing, and four others had less than 70% correct on the training views during testing. This left five participants per group.

2.2. Stimuli and design

The stimuli were two distinctive and two similar shapes (see Fig. 2). Both exemplars of a given type (similar or distinctive) were presented on each trial as animated movies; one exemplar of each stimulus type was arbitrarily assigned to be the S+ for all participants. The particular distinctive shapes that were used were still relatively similar to each other, both parametrically and psychophysically (i.e., the two objects differed by a 30% morph on three parameters for each of their parts, which yielded approximately 60% correct discriminations for rotating pairs; see Schultz, Chang, & Vuong, 2008). The part-structured stimuli in the present study were different than those used in our previous experiments with human and pigeon subjects (Friedman et al., 2009); in that study there were no viewpoint effects for humans responding to the part-structured stimuli. Here, we wanted to see whether part-structured stimuli that were more difficult to discriminate psychophysically, but were still distinctive, would show viewpoint effects.

Bitmap images of each stimulus were made at each one degree of viewing angle by moving clockwise around a circle that was centered on the objects. The stimuli were rendered as grayscale objects against a yellow background. When displayed side-by-side on the LCD screen, each object in a pair was displayed in an area that was 450 by 450 pixels (approximately 13.23 by 13.23 cm).

The 360 bitmaps for each object were divided into 12 viewpoints of 30 consecutive bitmaps each. Fig. 1 shows the viewing conditions for four of the eight experimental groups (one group in each motion type × stimulus order condition). For participants in these groups, the Training movies were made from the 60 bitmaps in viewpoints 3 and 5 in the figure. The test stimuli consisted of the two Training viewpoints, as well the two viewpoints that showed the 60 views that preceded viewpoint 3 (Far-Pre and Near-Pre), two viewpoints that followed viewpoint 5 (Near-Post and Far-Post), one viewpoint that showed the views in between the training views (Interpolated), and one viewpoint that was taken from the other side of the figure (Far-Opposite). The remaining four groups were trained with viewpoints 9 and 11, and had corresponding assignments of viewpoints to the other conditions. In particular, the Far-Pre, Near-Pre, Interpolated, Near-Post, Far-Post, and Far-Opposite viewpoints for these groups were 7, 8, 10, 12, 1, and 4, respectively. Thus, the views that were within the Interpolated viewpoint for half the participants were within the Far-Opposite viewpoint for the other half and vice versa; similarly, the views that were in the Far-Post viewpoint for one group were in Far-Pre viewpoint for the other and vice versa. This counterbalancing ensures that particular views could not be responsible for either the view combination or the Pre-Post effects.

Half the participants were assigned to the coherent group, and received all of their training and testing stimuli in the correct consecutive order for each of the two training viewpoints, so that the resulting movie showed smooth motion (the *coherent* group). Phenomenologically, the objects appeared to move smoothly around their vertical axis and then "jump" (via apparent motion) to a new position and continue to move smoothly in the same direction. Although we did not counterbalance for the direction of rotation (right to left or left to right), this variable was counterbalanced in previous studies using similar or identical stimuli as the present study (Spetch et al., 2006; Vuong & Tarr, 2006) and did not affect any of the other variables, which were similar to those used in the present study.

For the remainder of the participants the bitmaps within each viewpoint were first divided into 10 clusters of three bitmaps each; the 10 clusters within each viewpoint were presented in a different randomized sequence on each trial (the *scrambled* group). The resulting stimuli still had some motion but it was choppy. Furthermore, a random selection was made for each training trial in the scrambled condition that determined the order in which the 20 clusters of three viewpoints were seen across the two training segments (either 3 and 5 or 9 and 11); this selection was made independently for the S+ and S− objects. We did this to enhance the unpredictability of the viewing sequence. However, by the end of the training trials, both the coherent and scrambled-motion groups saw the identical amount of structure in each of the stimuli for the same amount of time.

All participants received two blocks of 40 training trials followed by one block of 80 test trials for each object type (distinctive and similar), for a total of 320 trials. On each training trial, the two training viewpoints (e.g., viewpoints 3 and 5) were presented together; across each block of 40 training trials the S+ was randomly selected to be on the right half the time and on the left for the other

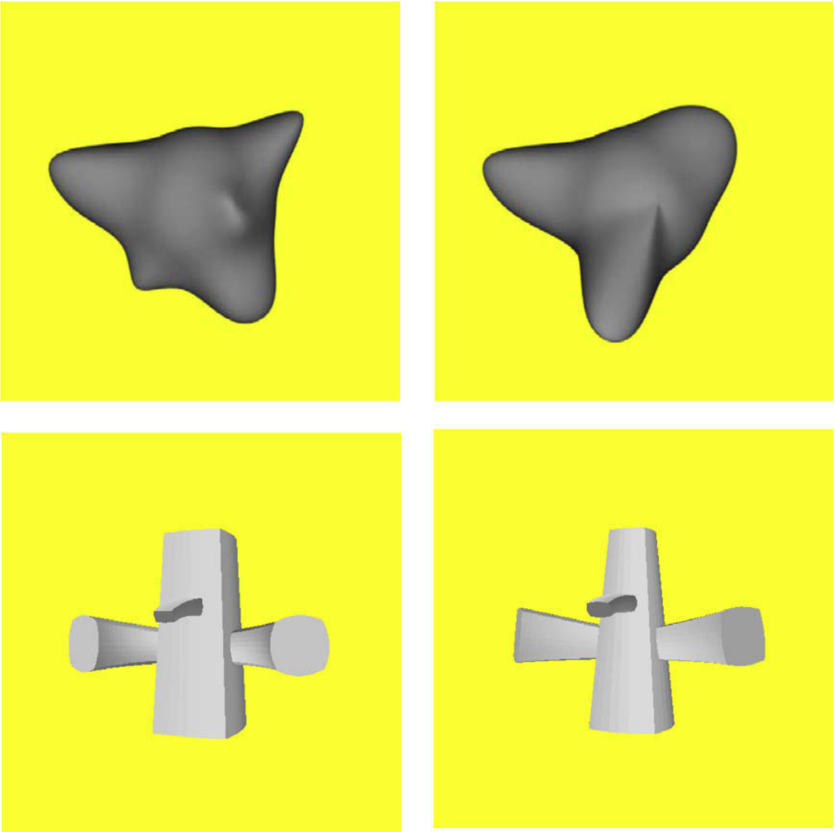

Fig. 2. Stimuli used in the similar (top) and distinctive (bottom) conditions, respectively.

half. For the 80 test trials each viewpoint (e.g., viewpoints 1–7 and 10) was presented twice by itself, randomized in blocks of 16 trials; the S+ in each pair was on the right half the time and on the left the other half. Each viewpoint was thus seen 10 times throughout the test trials.

2.3. Procedure and apparatus

When a participant arrived, he or she was seated in a small room in front of a computer with an NVidia GeForce 7600GS video card. The stimuli were displayed on a 19” Samsung Syncmaster 940BF LCD monitor with a 2 ms gray-to-gray response rate. The functional frame rate was 30 frames/s. The movies were presented as pairs of animations, centered on the screen, which was approximately 60 cm from the participant. There was a button box in front of the monitor with two push button switches that were 8 cm from center to center. Participants responded by pushing the button on the side of the response box that corresponded to the object they thought was the S+.

After signing the consent form, the initial instructions were presented on the computer screen with the experimenter present. The instructions informed the participants that their task was “to learn a discrimination between two stimulus displays of novel objects that are shown as animated movies.” They were also told that they could earn money for accurate and fast responding and that if they scored perfectly the amount they would earn would be $8.00. They were told their total earnings at the end of each training and test block.

During the first training block, participants could not respond until the movies had been shown for three cycles; correct responses earned 1¢ and incorrect responses were penalized 1¢. Participants were asked to look at both objects as much as possible; they were told they would have to guess at first which was the correct object but that they would get feedback on each trial and they should use it to figure out which object was the correct object. The participants received feedback in the form “You earn 1¢” or “You lose 1¢” after each trial.

For each training trial, a beep sounded simultaneously with the onset of a fixation point, which remained on for 750 ms. Then both viewpoints of each object (the S+ and S−) were shown for a total of three cycles per two viewpoints (6 s total), and after the participant responded, the feedback for that trial was displayed for 1 s. This was followed by a 1 s inter-trial interval. Participants could not respond until the end of the three cycles.

For the second block of training trials, the procedure was the same, but the participants were additionally told that each time they responded correctly and “are fast enough” they would earn 3¢; otherwise, if they were correct they would earn 1¢, and if they were incorrect 1¢ would be subtracted from their total. The 3¢ reward was given for responses that were made in under 1 s, but the participants did not know the exact time that was being used as the criterion. During these trials the movies were still repeated for three full cycles but participants could respond any time after the onset of the stimuli.

For the test trials, only one movie segment was shown at a time and no feedback was given. The procedure was otherwise the same as for the second set of learning trials. In addition, participants were warned that some of the animations they would see would be different than what they had previously seen, and that they “should try to decide whether to respond to the right or left side based on which object is the correct one, given your previous feedback.” They were told that the same earning scheme was in place as had been for the previous block of trials but that they would not get feedback.

After finishing the test trials for the first stimulus type, the participants were given a short break, and then they proceeded to the second stimulus type. The procedure was identical.

3. Results

We used $p < .05$ as the criterion for significance throughout the study and report η_p^2 as the measure of effect size. As in our previous study (Friedman et al., 2009), we averaged each participant's correct reaction times (RTs) separately over the distinctive and similar objects and omitted RTs that were more than 3 *SDs* above these means from further consideration. The omitted trials were counted as errors and comprised 1.6% of the total data; they were counted as errors on the assumption that long responses could have been either correct or incorrect by chance. In addition, and irrespective of whether we trimmed the RTs, there were two participants (one each in the coherent and scrambled groups) who had no correct trials for one of the novel test viewpoints; their means for that condition were replaced by the group means.

Figs. 3 and 4 show the RT and accuracy data, respectively. Three aspects of the data are immediately evident. First, there was a large effect of viewpoint on both measures for the similar stimuli, but little or none for the distinctive stimuli. Second, there was very little evidence for an overwhelming benefit of coherent motion over scrambled motion; however, coherent motion was generally facilitative for the Far-Post viewpoint relative to the Far-Pre viewpoint for the similar objects. Third, for the similar stimuli participants were as fast and accurate on the novel Interpolated viewpoint as they were on the Training viewpoints in both motion conditions, but they were slower and less accurate in their responses to the views shown in all the other novel viewpoints. In contrast, participants were fast and accurate on all of the novel views of the distinctive objects.

These observations were confirmed with analyses of variance (ANOVAs) on both RT and accuracy in which motion type was a between-subjects factor and object type (distinctive; similar) and viewpoint (Far-Pre, Near-Pre, Interpolated, Near-Post, Far-Post, and Far-Opposite) were within-subjects factors. Note that all of the tested views were novel.

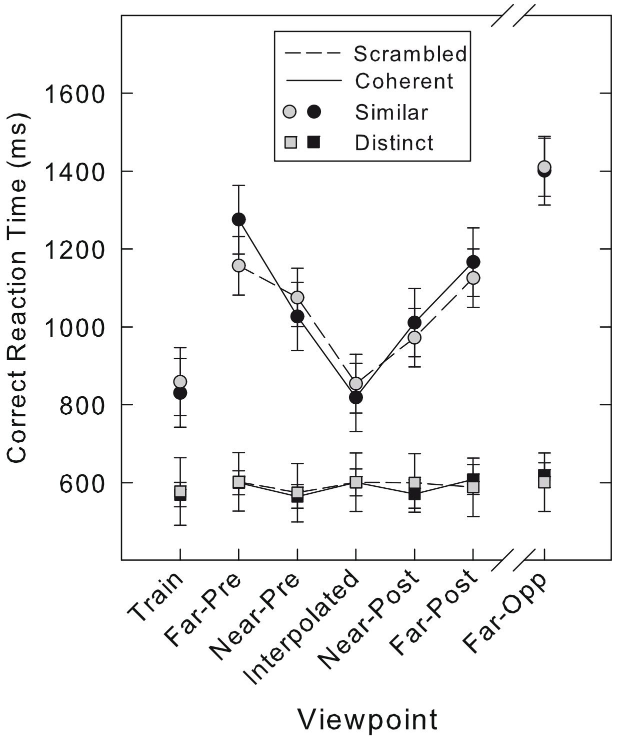

Fig. 3. Correct RTs as a function of object type, viewpoint, and motion type. Error bars in this and all subsequent figures with data are 95% confidence intervals computed from separate ANOVAs on the two motion type conditions, separately for training and testing stimuli (Loftus & Masson, 1994).

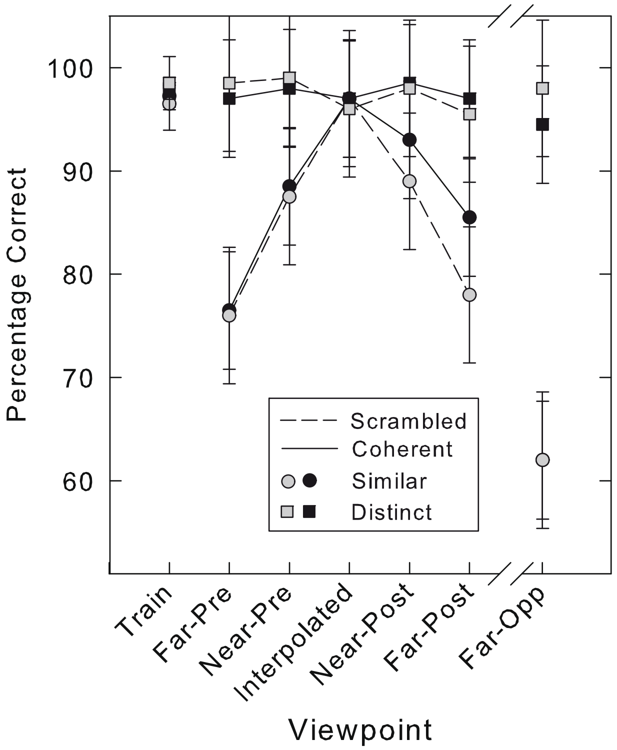

Fig. 4. Percent correct as a function of object type, viewpoint, and motion type.

There were robust main effects of object type for both measures, $F(1, 38) = 64.61$, $MSE = 489{,}441.88$, $\eta_p^2 = .63$ for RT, and $F(1, 38) = 70.79$, $MSE = 360.50$, $\eta_p^2 = .65$ for accuracy. For RTs, the distinctive objects were responded to faster than the similar objects at all of the novel views; for accuracy, the exception was the Interpolated views, which were highly accurate for both object types. There were also main effects of viewpoint type for both measures, $F(5, 190) = 24.04$, $MSE = 34{,}204.06$, $\eta_p^2 = .39$ for RT, and $F(5, 190) = 14.20$, $MSE = 231.81$, $\eta_p^2 = .27$ for accuracy. However, both main effects were modified by the interaction between object type and viewpoint, $F(5, 190) = 20.79$, $MSE = 34{,}720.26$, $\eta_p^2 = .35$ for RT, and $F(5, 190) = 14.74$, $MSE = 198.85$, $\eta_p^2 = .28$. We discuss this interaction further below. No other effects were significant in either omnibus analysis (all Fs < 1.00). It should be noted that exactly the same results were obtained for both measures when we conducted ANOVAs that excluded the Far-Opposite viewpoint, for which performance was particularly poor for the similar stimuli. We did these ANOVAs to ensure that the Far-Opposite viewpoint was not driving the interactions. Importantly, in the ANOVAs that did not use the Far-Opposite viewpoint, the quadratic component of the interaction between object type and viewpoint was significant for both measures, $F(1, 38) = 32.04$, $MSE = 41{,}625.05$, $\eta_p^2 = .457$ for RT, and $F(1, 38) = 25.38$, $MSE = 171.84$, $\eta_p^2 = .400$ for accuracy. However, the quadratic component of the triple interaction was not significant for either measure, which is evidence for view combination effects of the same magnitude occurring with similar stimuli in both motion conditions.

3.1. Effects of view combination

Further examination of the Viewpoint × Object Type interaction showed that for the similar stimuli there was very clear evidence that view combination took place in both motion conditions. For RTs to similar stimuli in the coherent condition, performance to the Interpolated view (819 ms) was significantly faster than performance to both the Near-Pre (1027 ms) and Near-Post views

(1011 ms), $t(19) = 2.40$, $SD_{diff} = 387$ ms and $t(19) = 2.15$, $SD_{diff} = 400$ ms, respectively, and there was no difference in RT between the Interpolated and Training viewpoints, $t < 1.00$. Similarly, for similar stimuli in the scrambled condition, there was no difference in RT between the Training and Interpolated views, $t < 1.00$, but performance was significantly faster to the Interpolated view (854 ms) than to both the Near-Pre (1075 ms) and Near-Post viewpoints (972 ms) views, $t(19) = 3.18$, $SD_{diff} = 310.46$, and $t(19) = 2.26$, $SD_{diff} = 233.07$, respectively. The finding that participants were able to interpolate with similar stimuli when the motion was scrambled replicates our previous finding (Friedman et al., 2009) and suggests that the interpolation mechanism, in itself, is insensitive to the difference between coherent and scrambled motion cues when stimuli are structurally similar.

For the distinctive stimuli in the coherent motion condition, responses to the Interpolated view (601 ms) were slightly slower than responses to the Training (569), Near-Pre (566), and Near-Post (571 ms) views; $t(19) = 3.62$, $SD_{diff} = 39$ ms for Training vs. Interpolated; $t(19) = 4.06$, $SD_{diff} = 40$ ms for Near-Pre vs. Interpolated, and $t(19) = 2.62$, $SD_{diff} = 51$ ms for Near-Post vs. Interpolated, respectively. It is likely that performance for the distinctive objects is nearly at floor; it is thus not clear what to make of any RT differences in this condition. No differences among these conditions were significant in the scrambled motion condition.

The accuracy data mirrored the pattern observed in RT data for the effect of view combination with similar stimuli, although some of the effects only approached significance. There were no differences in accuracy between the Training and Interpolated viewpoints for either the coherent or scrambled condition, $t(19) < 1.00$ for both. For similar stimuli in the coherent condition, participants were more accurate on the Interpolated viewpoint (97.0%) than on either the Near-Pre (88.5%) or Near-Post viewpoints (93.0%), $t(19) = 2.60$, $SD_{diff} = 14.6\%$ and $t(19) = 1.80$, $SD_{diff} = 10.0\%$, $p = .09$, respectively. For the scrambled condition, participants were also more accurate on the Interpolated viewpoint (97.0%) than on either the Near-Pre (87.5%) or Near-Post (89.0%) viewpoints, $t(19) = 2.03$, $SD_{diff} = 20.8\%$, $p < .06$, and $t(19) = 2.37$, $SD_{diff} = 15.1\%$, respectively. None of the effects were significant for the distinctive stimuli.

3.2. Effects of motion type

We next examined the differences, if any, between performance on the viewpoints that preceded or followed the Training viewpoints (see Fig. 5). For both RT and accuracy, we conducted ANOVAs in which Object Type, Near/Far, and Pre vs. Post position were within-subjects factors and motion type was the between-subjects factor. The main effect of Near-Far was significant, $F(1, 38) = 35.86$, $MSE = 18{,}441.67$, $\eta^2_p = .49$ for RT, and $F(1, 38) = 31.74$, $MSE = 88.87$, $\eta^2_p = .46$ for accuracy, as was the interaction between Near-Far and Object Type, $F(1, 38) = 22.63$, $MSE = 16{,}705.40$, $\eta^2_p = .37$ for RT, and $F(1, 38) = 17.63$, $MSE = 94.46$, $\eta^2_p = .32$ for accuracy. Most importantly, for RT, but not accuracy, there was a 4-way interaction among the factors, $F(1, 38) = 4.41$, $MSE = 11823.16$, $\eta^2_p = 10$. As can be seen in Fig. 5, accuracy on the distinctive objects was at ceiling for all of the views, which may have mitigated against revealing the 4-way interaction for this measure.

Further examination of the interaction in the RTs showed that most of the difference in performance on Pre-Post viewpoints took place in the coherent motion group when they were responding to similar stimuli. This was especially evident in the viewpoints that were furthest away from the training views. In particular, for similar stimuli undergoing coherent motion, there was a significant 109 ms facilitation in performance for the Far-Post views compared to the Far-Pre views, $t(19) = 2.29$, $SD_{diff} = 213$ ms. In contrast, there was only a 16 ms difference for the Near-Post vs. Near-Pre viewpoints undergoing coherent motion, $t(19) < 1.00$. The differences between the Far-Post and Far-Pre (32 ms) and Near-Post and Near-Pre (103 ms) views of the similar stimuli in the scrambled-motion group were not significant. Furthermore, none of the Pre-Post comparisons were significant for the distinctive stimuli, regardless of motion type.

For the similar stimuli undergoing coherent motion, none of the specific Pre-Post tests were significant for accuracy, although the differences were in the expected direction. The difference in accuracy between the Far-Pre and Far-Post viewpoints was 9.0% and between the Near-Pre and Near-Post viewpoints it was 4.5%. For scrambled motion, the differences in accuracy between Far-Pre and Far-Post viewpoints were small: 2.0% for the Far viewpoint and 1.5% for the Near viewpoint.

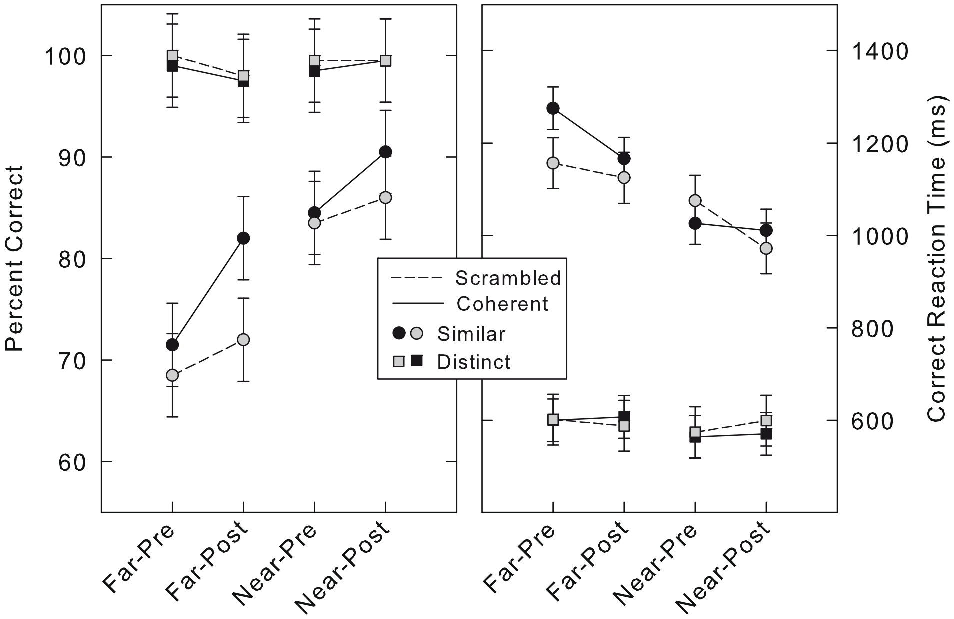

Fig. 5. Correct RTs and percent correct as a function of object type, motion type, and whether the segments were near or far from the training views, and before or after them in the sequence.

4. Discussion

The data from the present experiment can be summarized as follows. First, similar objects that had no part-structure were much more difficult to recognize and were responded to much more slowly than objects with a part-structure at all novel viewpoints except the Interpolated view. Thus, view combination mechanisms can overcome the advantage part-structure often confers to recognition (e.g., Biederman, 1987), insofar as for our stimuli with no part-structure, the Interpolated views were responded to as well as the Trained views. This replicates our previous finding (Friedman et al., 2009). Furthermore, the present results suggest that view combination mechanisms do not necessarily depend on the total amount of the object seen; in our previous study observers saw 1/3 of each object during training whereas here they only saw 1/6 of each object. Importantly, however, the "gap" between views was the same across the two studies; we have shown elsewhere (and the theory predicts) that view combination mechanisms can only be engaged to combine information from multiple views if the learned views are sufficiently close (Friedman et al., 2005).

Second, there were effects of motion per se, insofar as for similar stimuli, views that followed the training views were recognized faster than equally distant preceding views; in contrast to view combination effects, the motion effects occurred only when the training motion was coherent rather than scrambled. It is important that the present motion effects were obtained in a long-term memory paradigm, because as noted earlier, they allow us to rule out recency as the cause of the facilitation of performance to the Far-Post views. Further, a recency effect predicts that the Near-Post views would be responded to more accurately and/or quickly than the Interpolated view, but this was not the case.

The effects of motion were approximately the same order of magnitude in the present experiment as they were previously for humans (Friedman et al., 2009, Experiment 1a), and there was still not an overall effect of motion type (coherent vs. scrambled). In the previous study the difference between the Pre and Post conditions for similar stimuli was 138 ms. In the present experiment, the difference between the Far-Pre vs. Far-Post conditions, which were the identical viewpoints as in the present study, was 109 ms. From one perspective, the similar size of the motion effect across the two studies is surprising because the motion swept across two viewpoints on each trial in the present study, whereas it swept across only a single viewpoint on each trial in the previous study. Pigeons in our previous study showed an effect of coherent motion when the motion swept across two viewpoints but not when it swept across a single viewpoint. We therefore expected that the overall sense of global motion would be enhanced in the present study with humans relative to that seen in the previous study. However, it appears that either the enhanced sense of motion failed to affect the speed with which views that followed the trained views were processed, or, that the sweeping motion across two segments did not enhance the global direction of motion for humans as much as it appeared to have done for pigeons. On the other hand, in the present experiment, much less of the objects' overall shapes were exposed during training (60° vs. 120°). Thus, from this perspective, sweeping across the two training viewpoints was effective in maintaining the same magnitude of Pre-Post effect in RT as was training with twice as much of the objects' structure but with less overall global coherence attributable to the motion.

Third, there were large effects of viewpoint for the similar objects, with performance generally being a graded function of distance from the Training viewpoints. The important exception was stimuli at the Interpolated viewpoint, which were as efficiently processed as stimuli at the Training viewpoints. Further, for similar objects, evidence for view combination of about the same magnitude was obtained in both the coherent and scrambled motion conditions. This replicates the previous findings for humans (and pigeons), but much more strongly. Consistent with previous work (Stone, 1998, 1999; Wallis & Bülthoff, 2001) we found that an explanation for view combination based purely on temporal associations between images is not sufficient to explain the pattern of recognition performance in data (but see Liu, 2007; Wang, Obama, Yamashita et al., 2005, for some counterevidence with humans and monkeys, respectively). For example, by immediately going from viewpoint 3 to viewpoint 5, we placed very disparate views in close temporal proximity. On a purely temporal association account, we would not expect any benefit for Interpolated views if viewpoints 3 and 5 become associated into a single representation. In other words, on a temporal association account the tuning functions that originally represented each of Views 3 and 5 independently would broaden to become one function that would exclude the information in viewpoint 4 (see Fig. 1).

Thus far, the main empirical difference between view combination and motion effects is that view combination facilitates recognition of views that are within the range of the training views for static objects and those undergoing either coherent or scrambled motion. Although we did not explicitly test this hypothesis here, there are clearly angles between training views that are sufficiently large that they would mitigate against facilitation by view combination because the tuning functions for the objects would not overlap (c.f., Friedman et al., 2005). In contrast, motion effects like those observed here, as well as representational momentum, facilitate recognition of views outside of the range of the training views in the direction of the global coherent motion (and not for scrambled motion). Thus, it may be that as long as coherent motion is observed, the effective angle(s) for achieving facilitation by motion may be relatively large (c.f., Kourtzi & Shiffrar, 1999). These are important empirical differences that imply that the process(es) underlying each type of effect (facilitation of recognition by view combination and motion) are different, in both short- and long-term memory. View combination is usually implemented as a form of stimulus generalization (e.g., Edelman & Bülthoff, 1992). Some form of generalization may also underlie motion effects for novel views that follow the path of coherent rotation. However, even if generalization underlies both effects, it clearly manifests itself differently for view combination and motion-specific effects like representational momentum. For example, predicting a "next" view or biasing a particular set of views based on a coherently moving object may depend upon non-structural information like velocity, prior knowledge or expectations about the physical world and the properties of the object (e.g., Reed & Vinson, 1996), and so on. Elaborating what these differences are between view combination and motion effects requires future research with both short- and long-term memory paradigms.

Acknowledgments

This research was supported by Grants from the Natural Sciences and Engineering Research Council of Canada to A.F. and M.S. We are grateful to Bernd Kohler for programming the experiment, and Lauren Jerke for collecting the data.

References

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

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.

Cook, R. G., & Katz, J. S. (1999). Dynamic object perception by pigeons. *Journal of Experimental Psychology: Animal Behavior Processes, 25*, 194–210.

Cook, R. G., & Roberts, S. (2007). The role of video coherence on object-based motion discriminations by pigeons. *Journal of Experimental Psychology: Animal Behavior Processes, 33*, 287–298.

Cook, R. G., Shaw, R., & Blaisdell, A. P. (2001). Dynamic object perception by pigeons: Discrimination of action in video presentations. *Animal Cognition, 4*, 137–146.

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.

Freyd, J. J. (1987). Dynamic mental representations. *Psychological Review, 94*, 427–438.

Freyd, J. J., & Finke, R. A. (1984). Representational momentum. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 10*, 126–132.

Freyd, J. J., & Finke, R. A. (1985). A velocity effect for representational momentum. *Bulletin of the Psychonomic Society, 23*, 443–446.

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., Vuong, Q. C., & Spetch, M. L. (2009). View combination in moving objects: The role of motion in discriminating between novel views of similar and distinctive objects by humans and pigeons. *Vision Research, 49*, 594–607.

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

Harman, K. L., & Humphrey, G. K. (1999). Encoding "regular" and "random" sequences of views of novel three-dimensional objects. *Perception, 1999*, 601–615.

Kourtzi, Z., & Nakayama, K. (2002). Distinct mechanisms for the representation of moving and static objects. *Visual Cognition, 8*, 248–264.

Kourtzi, Z., & Shiffrar, M. (1999). The visual representation of three-dimensional, rotating objects. *Acta Psychologica, 102*, 265–292.

Kourtzi, Z., & Shiffrar, M. (2001). Visual representation of malleable and rigid objects that deform as they rotate. *Journal of Experimental Psychology: Human Perception & Performance, 27*, 335–355.

Liu, T. (2007). Learning sequence of views of three-dimensional objects: The effect of temporal coherence on object memory. *Perception, 36*, 1320–1333.

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

Loidolt, M., Aust, U., Steurer, M., Troje, N., & Huber, L. (2006). Limits of dynamic object perception in pigeons: Dynamic stimulus presentation does not enhance perception and discrimination of complex shape. *Learning & Behavior, 34*, 71–85.

Peissig, J. J., Wasserman, E. A., Young, M. E., & Biederman, I. (2002). Learning an object from multiple views enhances its recognition in an orthogonal rotational axis in pigeons. *Vision Research, 42*, 2051–2062.

Reed, C. L., & Vinson, L. G. (1996). Conceptual effects on representational momentum. *Journal of Experimental Psychology: Human Perception & Performance, 22*, 839–850.

Schultz, J., Chang, L., & Vuong, Q. C. (2008). A dynamic object-processing network: Metric shape discrimination of dynamic objets by activation of ocipito-temporal, parietal and frontal cortex. *Cerebral Cortex, 18*, 1302–1313.

Sekuler, A. B., & Palmer, S. E. (1992). Perception of partly occluded objects: A microgenetic analysis. *Journal of Experimental Psychology: General, 121*, 95–111.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317–1323.

Spetch, M. L., & Friedman, A. (2003). Recognizing rotated views of objects: Interpolation versus generalization by humans and pigeons. *Psychological 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.

Spetch, M., Friedman, A., & Vuong, Q. C. (2006). Dynamic object recognition in pigeons and humans. *Learning & Behavior, 34*, 215–228.

Stone, J. V. (1998). Object recognition using spatiotemporal signatures. *Vision Research, 38*, 947–951.

Stone, J. V. (1999). Object recognition: View-specificity and motion-specificity. *Vision Research, 39*, 4032–4044.

Tarr, M. J., Bülthoff, H. H., Zabinski, M., & Blanz, V. (1997). To what extent do unique parts influence recognition across changes in viewpoints? *Psychological Science, 8*, 282–289.

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

Vetter, T., & Poggio, T. (1994). Symmetric 3D objects are an easy case for 2D object recognition. *Spatial Vision, 8*, 443–453.

Vuong, Q. C., & Tarr, M. J. (2004). Rotation direction affects object recognition. *Vision Research, 44*, 1717–1730.

Vuong, Q. C., & Tarr, M. J. (2006). Structural similarity and spatiotemporal noise effects on learning dynamic novel objects. *Perception, 35*, 497–510.

Wallis, G., & Bülthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences, 98*, 4800–4804.

Wang, G., Obama, S., Yamashita, W., Sugihara, T., & Tanaka, K. (2005). Prior experience of rotation is not required for recognizing objects seen from different angles. *Nature Neuroscience, 8*, 1768–1775.

Weigelt, S., Kourtzi, Z., Kohler, A., Singer, W., & Muckli, L. (2007). The cortical representation of objects rotating in depth. *The Journal of Neuroscience, 27*, 3864–3874.