A speed-dependent inversion effect in dynamic object matching

Benjamin Balas

Laboratories of Cognitive Neuroscience, Children's Hospital Boston, Boston, MA, USA

Pawan Sinha

Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA

The representations employed by the visual system for dynamic object recognition remain relatively unclear, due in part to the lack of sufficient data constraining the nature of the underlying encoding processes. In the current study, we examined the limits of invariant recognition for unfamiliar moving objects in the context of a same/different matching task. In Experiments 1 and 2, Observers were asked to evaluate whether pairs of moving objects differed in identity subject to a spatial manipulation (inversion) and a spatiotemporal manipulation (speed change between sample and target). We find evidence of a speed-dependent inversion effect, suggesting distinct modes of processing for fast-moving and slow-moving objects. Furthermore, we observe a deleterious effect of speed change between sample and test stimuli, indicating that the speed of appearance change is encoded by the visual system for recognition. In a third experiment, we also observed a speed-dependency in the extent to which the direction of motion is encoded by the visual system for recognition. These results are discussed in the context of previous proposals regarding dynamic object representations, and in terms of an emerging model of dynamic object perception.

Keywords: object recognition, motion-3D, temporal vision

Citation: Balas, B., & Sinha, P. (2009). A speed-dependent inversion effect in dynamic object matching. *Journal of Vision, 9*(2):16, 1–13, http://journalofvision.org/9/2/16/, doi:10.1167/9.2.16.

Introduction

A growing literature indicates that complex moving objects are "more than the sum of their views" (Vuong & Schultz, 2008). That is, object motion makes a distinct contribution to object recognition, providing useful information beyond the large set of static images contained in a spatiotemporal volume of object appearances recorded over time.

The notion that object motion provides independent information for identification is in accord with decades of work examining the rich percepts evoked from point-light stimuli depicting human locomotion (Johansson, 1973). The gender, mood, and even identity of point-light walkers are readily obtainable from the dynamic stimulus (Cutting, 1987; Kozlowski & Cutting, 1977), despite the fact that individual frames are generally uninformative. The importance of dynamic features in this setting is made particularly clear when spatial factors diagnostic of category are put in conflict with dynamic features (such as a feminine or masculine gait), in which case the dynamic features typically govern the resulting percept (Thornton, Vuong, & Bülthoff, 2003). Experiments with other forms of biological motion (such as facial dynamics) similarly indicate an important role for dynamic information (Knappmeyer, Thornton, & Bülthoff, 2003). Besides the study of biological motion, results from recent years have shown that object motion has a far-reaching effect on the learning and recognition of a wide range of objects. Both object-centered and observer-centered motion of an object can affect recognition for distinct, clearly visible stimuli (Newell, Wallraven, & Huber, 2004), and novel, non-rigid objects appear to be encoded in terms of specific motion sequences (Chuang, Vuong, Thornton, & Bülthoff, 2006). Furthermore, efforts to determine how invariant recognition can be 'broken' have yielded important insights into how robust recognition (Cox, Meier, Oertelt, & DiCarlo, 2005; Wallis & Bülthoff, 2001) and discrimination (Balas & Sinha, 2008) are both learned from temporally extended visual inputs.

We can therefore be confident in saying that object motion "helps" object recognition and is readily used by observers in a variety of tasks. What we lack however, is a detailed understanding of how object motion is represented for the purposes of recognizing a moving object. As we have already said, it is clear that object motion is valuable beyond being a vehicle for a large number of static images; most studies have concentrated on demonstrating this to be the case. In several cases, critical spatiotemporal features for recognition have been identified in domains including speaking faces (Nusseck, Cunningham, Wallraven, & Bülthoff, 2008) and biological motion (Casile & Giese, 2005; Thurman & Grossman, 2008), but these results are fairly task-specific, identifying diagnostic regions in very particular stimuli rather than

doi: 10.1167/9.2.16 Received July 30, 2008; published February 19, 2009 ISSN 1534-7362

examining domain-general properties of moving object recognition. In particular, we argue that there is as yet little data to constrain the nature of the representation supporting the recognition of moving objects. Given the richness of a spatiotemporal input, there is a vast multitude of possible schemes for encoding spatiotemporal appearance. Are flow fields calculated between adjacent "frames" within a sequence? Is the entire ordered volume of 2-D images observed over time maintained as a single perceptual event? Are smaller "chunks" taken out of this large 3-D volume (Dollar, Rabaud, Cottrell, & Belongie, 2005; Ullman & Bart, 2004), or are spatiotemporal "edges" computed to render a temporally extended "primal sketch?" At the moment, we have little insight into which of these strategies is the one employed by the human visual system because we know very little about the basic properties of moving object recognition. Establishing boundaries within the space of possible representational strategies is thus a crucial first step towards the ultimate goal of determining how dynamic object appearance is encoded for recognition.

How shall we establish such boundaries on the representation of moving objects? Exploring the limits of invariant recognition provides exactly the kind of constraints we require to narrow in on specific mechanisms that may be relevant to recognition in a fully dynamic setting. Determining whether recognition is invariant to a particular stimulus transformation is an extremely valuable method for testing the limits of the encoding strategies used by the human visual system. To take static face recognition as an example, our understanding of how faces are represented and recognized was greatly enhanced by many studies detailing the limits of invariant recognition subject to a wide range of manipulations: Spatial filtering studies have determined critical bands of frequencies for efficient recognition (Costen, Parker, & Craw, 1994; Näsänen, 1999). The classic "inversion effect" suggested an orientation-specific encoding strategy employed by the visual system for faces (and possibly other objects processed at an "expert" level) (Diamond & Carey, 1986; Yin, 1969). Studies of color and luminance negation revealed that shape and surface properties are not encoded with invariance to luminance-edge polarity (Galper, 1970), and suggested high invariance to drastic changes in hue (Kemp, Pike, White, & Musselman, 1996). Further study with degraded images revealed that color does appear to be relevant for very blurry images, indicating an important interaction between two stimulus manipulations that face recognition appears to be highly robust to when applied singly (Yip & Sinha, 2002). Taken together, these findings constitute an extensive profile of empirical constraints that must be satisfied by a comprehensive theory of static face representation and recognition. Determining a similarly rich set of empirical constraints for moving object recognition would undoubtedly be a similarly important advance.

In the current study, our goal was to begin a line of investigation in this vein on the limits of recognition invariance for isolated dynamic objects. Given how little is known about how the visual system represents moving objects for recognition, this initial examination of the invariance properties of dynamic object recognition is a fundamental contribution towards a coherent theory of perception in natural environments. Beyond the rejection of the "motion = many static views" hypothesis (which is essentially a 'null' hypothesis for the role of object motion) few studies have investigated the invariant properties of dynamic object recognition, with a few notable exceptions. First, possibly the most specific proposal regarding the representation of dynamic appearance is that it may be encoded in terms of so-called "spatiotemporal signatures" that constitute an ordered record of image appearance over time (Stone, 2003). The basis for this hypothesis is the finding that the recognition of rigidly rotating objects is impaired when the direction of rotation at test is different from the direction observed during learning. This is most evident for objects whose appearance is obscured by fog or sparsely represented by dots (Stone, 1998, 1999; Vuong & Tarr, 2004) but also obtains for clearly viewed objects (Liu & Cooper, 2003). Second, Watson, Johnston, Hill, and Troje (2005) found that moving objects exhibited a greater degree of view-invariance than typical static counterparts. This result bolsters the idea that a moving object is more than just a collection of static images, demonstrating a case where the representation of a moving object is more invariant than the representation employed for a static stimulus. These results thus provide us with two important starting points for characterizing dynamic object recognition: one case in which it is invariant (view) and another in which it is not (sequence reversal). These basic findings are important, but are also clearly only a small part of what must be a much larger set of interesting constraints imposed by the representation of dynamic form. Here, we extend what is known about the scope of invariant recognition for moving objects by exploring a subset of these untested constraints.

In Experiments 1 and 2, we examined the effects of two basic transformations of dynamic object appearance, one "spatial" transformation and one "temporal" transformation. In both cases, we used a same/different object-matching task incorporating novel, rigidly rotating objects. Our task obviated the need for category or exemplar training by presenting subjects with a simple image-level judgment. Our spatial manipulation was carried out by varying object orientation (upright or upside-down) at test. This allowed us to examine the extent to which the classic "inversion effect" for faces obtains for moving objects. That is, is the representation of a moving object relatively invariant to the orientation of the individual frames within the sequence, or is there a recognition cost associated with rotating images in the picture plane between sample and test stimuli? Our

temporal manipulation was carried out by varying the speed of the test object. Changing the speed of an object preserves all the image data available to the observer in any single frame (modulo the differences in presentation time), yet systematically alters the dynamic properties of the input. This manipulation allowed us to ask whether or not recognition is invariant to the *speed* or *magnitude* of object motion (a question posed by Stone, 1998 following his observation that the *direction* of object motion is encoded by the visual system). We find that the visual system exhibits an interesting pattern of invariance and impairment to these two manipulations. Our data suggests an intriguing interaction between spatial and spatiotemporal processes. Specifically, both our spatial and temporal transformations affect recognition in a speed-dependent way. The data suggests that moving objects may be encoded differently depending on how quickly appearance changes over time. In Experiment 3, we extend this result by reporting a speed-dependency in the extent to which reversal of an image sequence between study and test affects performance. We discuss all of these results in terms of a cue-combination model of dynamic object perception and suggest avenues for future research.

Experiment 1—Is there an "inversion effect" for matching moving objects?

We began by examining the effect of a spatial manipulation, object inversion, on matching performance. Our primary question was whether or not the ability to match moving objects was at all dependent on the object orientation observed at sample and test. Also, we chose to use two different speeds of object motion so that we could examine how matching performance varied with playback rate. This manipulation also gave us the opportunity to look for evidence of an interaction between object orientation and speed. We presented observers with a same/different task in which they were required to compare upright sample objects to either upright or inverted test objects.

Methods

Subjects

Twelve volunteers from the MIT community participated in the experiment for pay. All subjects reported normal or corrected-to-normal vision, and were between the ages of 18 and 36 years of age. All volunteers were naïve to the design and purpose of the experiment.

Stimuli

Eight "greebles" (Gauthier & Tarr, 1997) were used for this study, each one rendered from a single point-light source using commercially available graphics software (POVray 3D). A 30-frame sequence was created for each greeble, depicting the object rotating about the X and Z axes simultaneously. Each greeble image was 123 × 123 pixels in size, subtending approximately 1.5 degrees of visual angle when presented on screen. Frontal views of all 8 greeble stimuli are presented in Figure 1 below.

Observers were asked to carry out a same/different task using short sequences of the greebles tumbling in space on a uniform black background. The tumbling motion depicted was created by rotating each object about both the X and Z-axes simultaneously, producing a complex sequence of appearances (see Supplementary Movie for examples of "fast" and "slow" motion). On each trial, subjects were asked to fixate on a small white cross presented at the center of the screen while attending to two greeble sequences presented in the near periphery

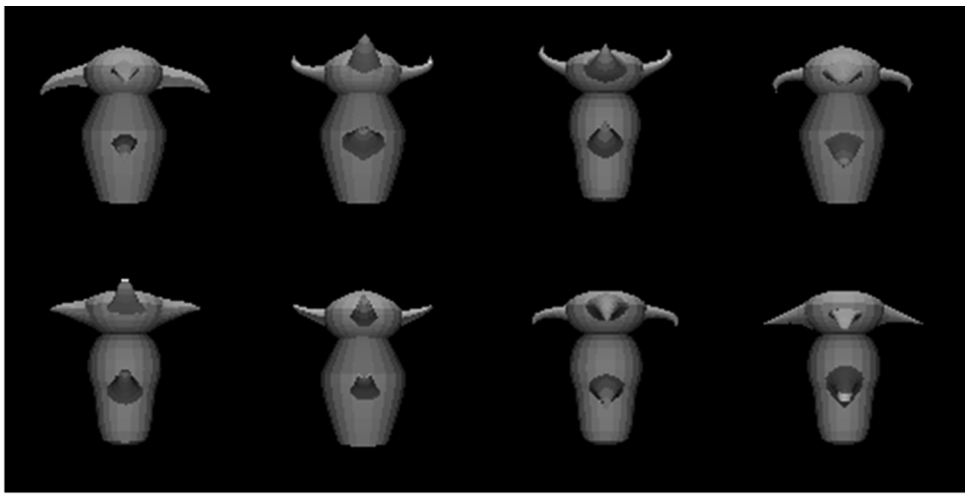

Figure 1. The 8 greebles used in Experiment 1. "Male" and "Female" exemplars of greebles from multiple classes were used.

(approximately 2 degrees out from fixation). First, a greeble was presented to the left of fixation (the "sample") followed by a second sequence to the right of fixation (the "target") that could either depict the same 3D object or a different one. Observers responded "same" by pressing the '1' key on the keyboard and responded "different" by pressing the '0' key. Auditory feedback was provided after each trial in the form of a loud beep that followed incorrect responses. Observers were encouraged to respond as quickly as possible without compromising accuracy. A chin rest was used to fix viewing distance at approximately 50 cm.

Within this basic task, we carried out two manipulations of the greeble sequences. First, on each trial the second greeble could either be presented upside-down or rightside-up, allowing us to examine the existence of an inversion effect for dynamic object matching. Second, the speed of the two sequences displayed on each trial was also varied. Each 30-frame sequence could be played at either 15 Hz or 30 Hz, allowing us to examine the effect of speed on matching performance. Both of these manipulations were carried out in a within-subjects design, with presentation speed blocked and counterbalanced for order across observers, and orientation (upright/inverted test items) randomly interleaved within each block. Regarding our manipulation of object speed, we point out that our design confounds object speed with the presentation time of each frame, meaning that throughout our discussion we must emphasize that either of these factors may underlie our results. Alternate means of manipulating speed could remove this particular confound, but would necessarily introduce other confounding factors. For the present, rather than explore a wide range of speed manipulations (such as frame-dropping, or matching total presentation time over multiple presentations of the frames in a "fast" sequence) we have opted to pursue this manipulation, and noting its particular limitations as we continue.

On each trial, the starting frame of each sequence was randomly selected so that observers could not usefully adopt the strategy of attending to consistent features available in either the first or last image of a sequence. This was done since the first and last images in a sequence are not "masked" by bracketing stimuli, making it possible that with repeated presentation observers would learn to attend only to the beginning and end of each sequence to perform the task. On each trial, the first greeble was selected at random (with replacement) from the full set of 8 objects. This selection routine generally led to distractor objects that differed from the original sample in terms of part shape and part orientation (horns pointing up or down). In a minority of trials (~10%) the only distinction between "different" stimuli was the orientation of parts, but this is not a large enough proportion of such trials to merit special concern. For "same" trials, the first greeble was presented a second time as the test stimulus, with the same starting and ending frames as the sample stimulus. For "different" trials, a second greeble was randomly selected from the set. Observers completed a total of 84 trials within each block for a total of 168 trials in the entire session. Within a block, half of the trials were "same" trials and half were "different." "Same" and "different" trials were further divided into an equal number of upright and inverted trials. Stimuli were presented on a 17-inch Sony monitor. All stimulus display parameters and response collection routines were executed with the Psychophysics toolbox for MATLAB (Brainard, 1997; Pelli, 1997).

Results

For each observer, we calculated d' values in all conditions as our measure of matching performance. Figure 2 contains a bar graph of mean d' values across observers in each experimental condition. All of our statistical tests assume normally distributed data, which is not always the case for d' scores. For the data reported here, we validated that there was no significant violation of normality before continuing.

To assess the effect of our spatial and spatiotemporal manipulations, we carried out a 2 × 2 repeated-measures ANOVA with object speed (slow or fast speeds for both the sample and test objects) and target orientation (test stimulus upright or inverted) as within-subject factors. This analysis revealed a main effect of target orientation ($F(1,11) = 8.825$, $p = 0.013$) but no effect of speed

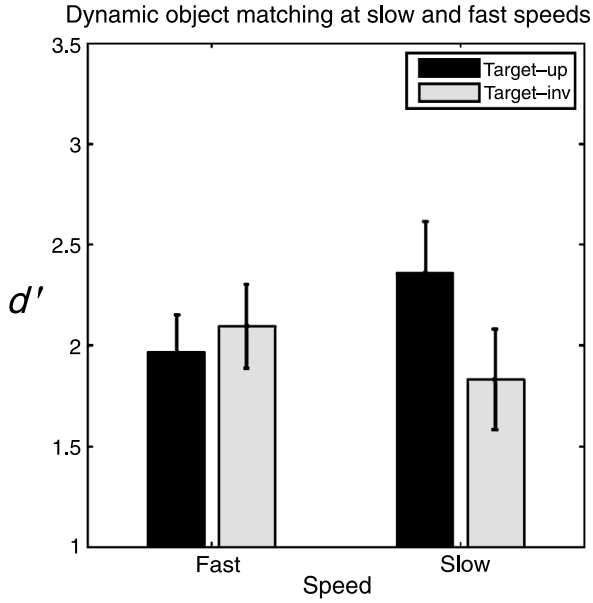

Figure 2. Discriminability in our dynamic object matching task as a function of stimulus orientation and sequence playback speed. In the "slow" condition, observers demonstrate a significant inversion effect that is absent for the "fast" stimuli. Error bars represent ±1 *SEM*.

($F(1,11) = 0.001$, $p = 0.989$). The interaction between speed and orientation was also not significant in this analysis ($F(1,11) = 1.504$, $p = 0.246$).

We continued by carrying out two planned comparisons to examine the inversion effect within each speed condition. We found that for our "fast" trials, there was no apparent cost associated with the inversion of the test stimulus ($t(11) = 0.357$, $p = 0.36$, one-tailed paired-samples test). For our "slow" trials however, there was a significant inversion effect ($t(11) = 2.360$, $p = 0.019$, one-tailed paired-samples test) indicating that the main effect of inversion in our ANOVA was driven primarily by the difference in upright and inverted performance on "slow" trials. It is evident from the figure that the speed-dependency of the inversion effect is not due to ceiling or floor effects for "Fast" objects, but rather reflects a distinct inversion cost for matching slow samples to inverted test objects.

Discussion

Experiment 1 revealed a novel speed-dependent inversion effect for moving object matching. This result places an interesting constraint on the representations supporting performance in our task. The lack of invariance to object inversion for "slow" stimuli indicates that in this setting, appearance is encoded in a highly image-specific fashion, perhaps by simple templates of spatiotemporal appearance sampled during viewing of the complete sequence. The invariance to inversion observed for "fast" objects suggests the use of a different representation of the same set of images, induced by the increase in sequence speed (or perhaps the decrease in presentation time for individual frames). This representation is either robust to the orientation change, or encodes the image in a coarse enough fashion that it is blind to the manipulation. However, the lack of an inversion effect at our fast speed is not simply the result of overall poor performance for quickly moving stimuli. Observers are not at floor performance in this condition, which means that though the representation of spatiotemporal appearance may be crude enough to generalize over picture plane rotation, it is not so crude that the matching task becomes markedly difficult to execute accurately. Indeed, average performance for fast-moving objects is fairly high (d' of $\sim$2), suggesting that observers are still capable of retaining a good deal of useful visual information at this frame rate. Given this, we conjecture that "fast" objects may be represented in a different (not better or worse) way than "slow" objects.

Is it possible that our data can be completely explained by differences in presentation time (and by extension sequence length)? To examine whether it is the speed of the object that matters or the total presentation time seen during a trial, we ran an additional group of 11 subjects on a control task in which objects moved at the "fast" speed already described, but turned through two full revolutions, thus equating total presentation time between this control condition and our original "slow" condition. If it is indeed the case that speed determines the magnitude of the inversion effect, we would expect to find that there is still no inversion effect in this condition. If however, it is total presentation time (and/or sequence length) that governs behavior, an inversion effect should be evident in this manipulation. What we find is that there is no difference between upright and inverted d' scores when "fast" presentation time is matched to our original "slow" condition (Mean upright = 1.7, Mean inverted = 1.74, $t(10) = -0.28$, $p = 0.788$, paired-samples t-test, two-tailed). This suggests that total presentation time is not the relevant factor in this task, but rather that speed (or we may also say local presentation time) determines the magnitude of the inversion effect. We shall therefore continue in Experiments 2 and 3 to define object speed in terms of presentation time/sequence length. Examining the nature of speed-dependent processing in this fashion is important due to the ecological relevance of speed as so defined, since fast-moving objects do give rise to shorter spatiotemporal inputs under natural viewing conditions. The results from this control condition further validate this choice by demonstrating that matching presentation time does not appreciably change performance.

Given the data from these three conditions, what can we begin to say about the nature of the relationship between object representation at fast and slow speeds? We speculate that a trade-off between certainty for individual appearances and certainty for relational features between distinct images may be a crucial determinant of how spatiotemporal appearance is encoded. Specifically, we conjecture that moving objects may be encoded in terms of two distinct types of visual information that are combined in a weighted fashion to yield a final approximation of dynamic appearance in terms of multiple cues.

When objects are moving slowly, observers have relatively good information about individual appearances due to the longer presentation time per image. The result is increased certainty for particular static appearances contained within the sequence. This increased certainty for particular images may encourage substantial reliance on image-based features, "snapshots" from the sequence, that are not robust to inversion. By contrast, a quickly moving object does not lend itself well to the extraction of image features derived from individual frames due to the comparatively brief duration of any individual image. However, this loss of image-based information may be compensated for by increased certainty in some form of relational encoding between images. Examples of such relational features could be the optic flow field between adjacent frames, or the local curvature of the line joining appearances together through a high-dimensional image space (Murase & Nayar, 1993). Despite not containing much information about individual frames within the sequence, the values that describe the local differences

between frames in a sequence could certainly be used as additional, complementary sources of information for recognition. Our data suggest that these representations of spatiotemporal appearance are probably highly invariant to inversion, leading to robust invariance at high speeds when they are employed and poor invariance at slow speeds when they are not. An important aspect of this model is that neither cue is ever categorically "thrown out," but each one varies in terms of how reliable a source of information it is. To clarify this point, in Figure 3 we offer a schematic view of how "fast" and "slow" encoding may differ.

Obviously our data do not allow us to conclude that flow fields or image-space trajectories are actually the mechanisms underlying our data. There are many alternative ways in which "fast" processing might differ from "slow processing." For example, the distinction between "holistic" and "part-based" processing is a key aspect of many theories of object recognition and expertise. Could it be the case that "fast" processing encourages "part-based" processing, which is generally more robust to inversion than "holistic" processing? This is an intriguing proposal in its own right and our data does not permit us to reject it. We also note that it is possible that speed may differentially modulate the extent to which particular features can be extracted from a spatiotemporal sequence, such that the discrimination of two "slow" objects may take advantage of a super-set of image differences relative to the discrimination of two "fast" objects." This hypothesis predicts consistent advantages for "slow-upright" processing (a point we shall revisit in Experiment 2) and also suggests that the inversion effect for slow-moving objects could be erased if only very gross differences between stimuli exist. For the moment, we must entertain these alternatives as important potential mechanisms that could produce our results. However, we also point out that the hypothesis that speed may modulate uncertainty for image-based and sequence-based representation codes provides a useful heuristic framework for describing the recognition of moving objects in terms of easily obtained measures of performance that acknowledge the dynamic nature of the input rather than relying on image features present in a static display. In particular, this hypothesis lends itself well to independent measurements of uncertainty for the recognition of individual images and flow-fields as a function of presentation rate that could be used in an attempt to predict behavior quantitatively in an object matching task. Still, we wish to emphasize that regardless of what computations actually guide behavior in this scenario, the result that image inversion disrupts performance in a speed-dependent fashion is an important first step towards our stated goal of identifying and characterizing constraints on the representation of dynamic object representation.

We continue in Experiment 2 by examining the effect of speed differences between sample and test objects on matching performance. Manipulating the speed of a moving object provides the opportunity to determine the limits of invariant recognition along a basic temporal dimension, complementing our use of a spatial manipulation in Experiment 1. Furthermore, if it is indeed the case that "fast" and "slow" objects are represented in distinct ways, observers' ability to match objects moving at different speeds may provide deeper insight into the similarities and differences between these two processing modes. Put another way, if you retain different information about spatiotemporal appearance as a function of object speed, then a speed difference between objects

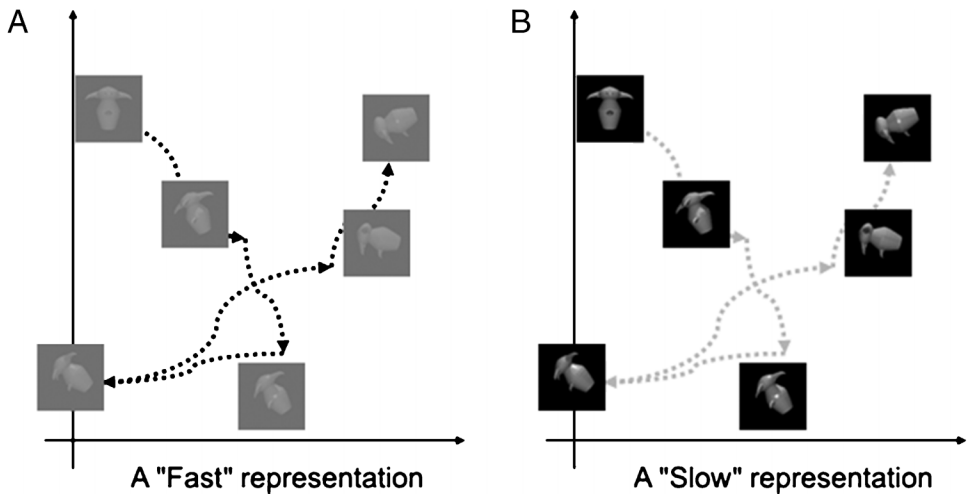

Figure 3. Schematic depictions of candidate representations that might support dynamic object matching in Experiment 1. A "fast" representation (A) may primarily contain information about the sequence of images in relation to one another (de-emphasizing individual images), while a "slow" representation (B) may be more image-based (de-emphasizing sequence) and thus more susceptible to inversion effects.

should incur a significant cost. Measuring visual information that doesn't "match" should make matching more difficult.

Experiment 2—Does comparing dynamic objects at different speeds impair recognition?

To further probe the nature of dynamic object representations, we continued by asking whether or not a speed change between sample and test stimuli incurred a cost on discriminability. To do so, we extended the design used in Experiment 1 to incorporate two additional "speed change" conditions, a "slow-to-fast" condition and a "fast-to-slow" condition. The inversion manipulation used in Experiment 1 was included in our design as well, allowing us to attempt to replicate our findings from Experiment 1 and also to directly compare the impact of changing inversion between sample and test to that of changing speed.

Methods

Subjects

Twelve observers, none of whom had participated in Experiment 1, volunteered to participate in Experiment 2. All observers were between 18 and 35 years of age and reported normal or corrected-to-normal vision. All observers were naive to the purpose of the experiment.

Stimuli

The same greeble objects employed in Experiment 1 were used here. All rendering settings and image parameters were preserved. Display characteristics and viewing distance was also maintained between Experiments 1 and 2.

Procedure

Overall task design was very similar to Experiment 1 with a few critical differences. First, the "slow-to-fast" and "fast-to-slow" conditions were added into our design, including both upright and inverted test stimuli. In the slow-to-fast condition, sample sequences played at 15 fps had to be compared to test sequences played at 30 fps. The fast-to-slow condition required the opposite comparison of a 30 fps sample sequence to a 15 fps test sequence. Second, instead of adopting a blocked design as in Experiment 1, we opted in Experiment 2 for a fully randomized design. We did so to ensure that the results of Experiment 1 were not somehow related to a change in observers' strategy between blocks rather than a true difference in sensitivity and to minimize observers' ability to predict the speed of the sample sequence. Otherwise, all stimulus display and response collection procedures were identical to those described in Experiment 1.

Results

As before, observers' performance was characterized using d' as the measure of sensitivity to object differences. After verifying that our data did not violate assumptions of normality, we first examined whether or not we were able to replicate the speed-dependent inversion effect observed in Experiment 1. We carried out two planned comparisons of upright to inverted sensitivity in the "fast" and "slow" conditions (in which no speed change is carried out between sample and test sequences). These tests revealed no significant effect of inversion in the "fast" condition ($t(11) = 0.013$, $p = 0.49$, one-tailed paired-differences t-test), but a significant effect in the "slow" condition ($t(11) = 1.864$, $p = 0.044$, one-tailed paired-differences t-test), replicating the effect observed in Experiment 1.

We continued by examining the potentially deleterious effects of introducing a speed change between sample and test stimuli. To do so, we carried out a $2 \times 2 \times 2$ repeated-measures ANOVA with sample sequence speed ("fast" or "slow"), target orientation ("upright" or "inverted"), and relative sample/target speed ("same speeds" or "different speeds") as within-subjects factors. We observed a main effect of speed change ($F(1,11) = 12.91$, MSe = 2.71, $p = 0.004$), favoring comparisons between same-speed samples and targets, but no other significant main effects or interactions. Finally, we also point out that the d' values in this experiment are lower than those reported in Experiment 1. This is likely the result of our use of a fully randomized design in this task, rather than the blocked design employed in Experiment 1. A graph of the data from all conditions is displayed in Figure 4.

Discussion

There are multiple interesting features of the data obtained from Experiment 2. First, the data demonstrates that the speed-dependent inversion effect obtained in Experiment 1 is replicable, bolstering the evidence for speed-dependent processing of spatiotemporal appearance. Second, the data also provides evidence that there is a significant sensitivity cost accompanying a speed change between sample and target stimulus sequences. This latter result places an important constraint on the underlying representation of dynamic object appearance. Moving objects (whether fast or slow) are apparently

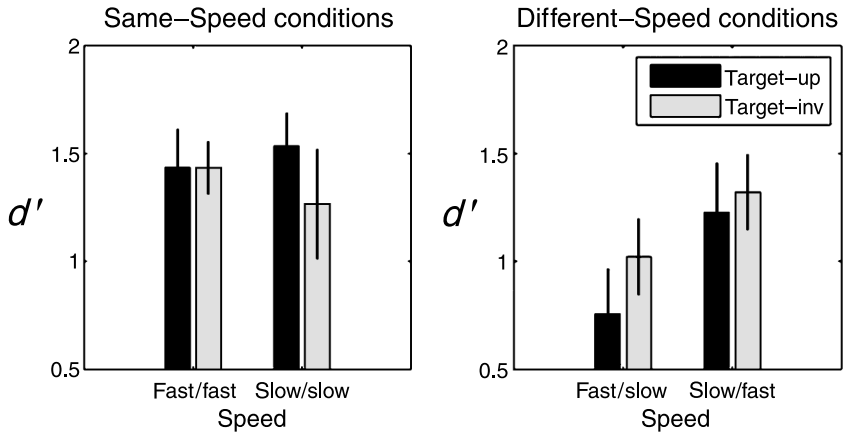

Figure 4. Sensitivity to object differences in our dynamic matching task as a function of relative speed and target orientation. We observe the speed-dependent inversion effect reported in Experiment 1, as well as a significant sensitivity cost when attempting to match objects moving at different speeds. Error bars represent ±1 *SEM*.

encoded at the speed at which they move, rather than solely in terms of the sequence of images contained in the sequence. We also briefly note that matching a fast object to a slow one appears to be more difficult than matching a slow object to a fast one. While incidental to the questions that motivated this experiment, this asymmetry may be an important clue to the representations employed at different speeds. We reiterate however, that regardless of the specific mechanistic interpretation of the results of Experiment 2, the key contribution is to establish limits on the invariant properties of moving object recognition. Here, we have observed that the speed of a moving object (or one can also say the magnitude of object motion) is encoded for recognition. Introducing a speed change between sample and test stimuli thus incurs a significant recognition cost, which directly indicates that the rate of appearance change over time is a relevant perceptual feature and indirectly suggests that the nature of the information retained after viewing a "fast" object may differ (either qualitatively or quantitatively) from the information retained after viewing a "slow" object.

Before continuing, we revisit an issue we raised in the discussion of the results from Experiment 1 that we can now address given the data reported here. Could a possible explanation of the speed-dependency of the inversion effect be that a slow-moving object is encoded via a super-set of features relative to the encoding of a fast-moving object? That is, two slow-moving objects might be perceived to differ in terms of both "gross" and "fine" features. If we imagine that both inversion and increased speed compromise the ability to detect "fine" features but not "gross" features, then we might obtain the data from Experiment 1 solely by this difference in static feature encoding as a function of local presentation time. This is already a difficult explanation to apply to our data since it is not clear that there is a wide range of "gross" and "fine" differences between these stimuli, but the data from Experiment 2 makes it possible to reject on different terms. This proposal directly predicts that what differentiates performance across the fast/slow and upright/inverted manipulations is especially high accuracy for slow-moving, upright objects, with approximately equally low performance in the other conditions. We point out that in Experiment 2, it is actually the case that slow-moving inverted object performance stands out as being particularly poor. We may thus tentatively reject this particular explanation of our result, though it is an important hypothesis to keep in mind for future research. The interaction of task difficulty with various spatiotemporal manipulations has previously been examined with interesting results (Vuong & Tarr, 2006), and may still yield interesting results in this context as well.

In our final experiment, we extend our investigation of invariant recognition by examining the effects of sequence reversal on object recognition across a range of speeds to help further elucidate the differences between the distinct modes of processing we have suggested here. The deleterious effects of sequence reversal on dynamic object recognition have been observed in several contexts (Stone, 1998; Vuong & Tarr, 2004, 2006) making this effect a compelling target for the current study. In particular, given that the reversal manipulation is primarily temporal in nature (since all static appearances are unchanged) we expect that reversal only has deleterious effects at high speeds, where the dynamic "signature" of appearance may be more vital to recognition and matching. This also provides an important control for our results from Experiments 1 and 2, both of which found effects at slow speeds only. By exploring the speed-dependence of the reversal effect, we have the opportunity to further verify that our results are not some artifact of overall "better" or more efficient processing for slow-moving objects and also to extend our paradigm to a new domain.

Experiment 3—Is the effect of sequence reversal speed-dependent?

In this final experiment, we ask whether or not the previously reported effect of sequence reversal on moving object recall (Stone, 1998; Vuong & Tarr, 2004, 2006) and recognition is speed-dependent. Given our characterization of speed as a determiner of the relative uncertainty for static image appearance and sequence-based features, we predict that the effects of sequence reversal should be more prominent at fast speeds than at slow speeds. At slow speeds, we expect that the observer has more certainty regarding individual frames, making a slow sequence robust to reversal (since all frames are identical). At fast speeds, we conjecture that the reliance on a "spatiotemporal signature" is more pronounced since individual frames can no longer be encoded with high fidelity. Thus, reversal should incur a greater cost. We explore this issue by adapting a recall task used by Vuong and Tarr (2006) to examine the nature of the reversal effect.

Methods

Subjects

Nine observers, none of whom had participated in Experiments 1 or 2, volunteered to participate in Experiment 3. All observers were between 18 and 35 years of age and reported normal or corrected-to-normal vision. All observers were naive to the purpose of the experiment.

Stimuli

The greeble stimuli used in Experiments 1 and 2 were also employed here. Again, all rendering settings and image parameters were preserved. Display characteristics and viewing distance were maintained from Experiments 1 and 2.

Procedure

Observers in this experiment carried out a two-stage learning and recall task adapted from Vuong and Tarr (2006). In the first phase of the experiment, observers learned to correctly label four unique greebles with the labels 1, 2, 3 and 4. On each trial, a greeble sequence was presented in the center of the screen (all display parameters being the same as described earlier for Experiments 1 and 2) and observers were asked to guess the correct label following completion of the sequence. Incorrect guesses were signaled with a short beep. Initially, observers could only guess at the correct label for each greeble, but learned to assign the correct labels rapidly. Each greeble was presented 30 times during the course of the learning phase (presentation order randomized) for a grand total of 120 learning trials per subject. Critically, two of the greebles presented during learning moved at our pre-defined "fast" speed and the remaining two moved at our pre-defined "slow" speed.

In the second phase of the experiment, observers were asked to perform an old/new task using the greeble stimuli. On each trial of this test phase, observers were presented with a single greeble sequence at the center of the screen. If they believed that the greeble was one of those presented during the learning phase, observers were asked to respond with the label (1–4) that they believed went with that object. Otherwise, if they believed it to be a "new" greeble not presented during the learning phase, they were asked to press the space bar. No feedback was provided during this phase of the experiment. The four greebles used in the learning phase were intermixed with four new greebles, half of which moved at a "fast" speed and half of which moved at a "slow" speed. The speed of all "old" greebles was preserved between learning and test phases. Finally, all greeble sequences were presented both in "forward" and "reversed" order so that we could measure the reversal effect for fast and slow objects. Each greeble was presented 20 times in the "forward" direction and 20 times in the "reversed" direction for a grand total of 320 trials. Accuracy and Response Time were measured for all subjects. Response time was defined relative to the start of the second stimulus, as subjects were free to respond throughout the entire presentation of this sequence.

Results

Per our predictions regarding the nature of dynamic object representation at different speeds, the critical questions are:

1. Do we observe a "reversal effect" such that test sequences are harder to correctly identify if frame order differs from that learned during training?
2. Does the "reversal effect" depend on speed? Specifically, is it reduced for slow-moving objects?

We examined observers' "hit rates" (The proportion of "old" objects correctly labeled), "misses" (the proportion of "old" objects labeled as "new"), and the response time for correct identifications. RTs were pre-processed to remove responses taking longer than 3 seconds, ensuring a robust estimate of the mean that is not driven by a small number of outlying data points.

Analysis of both "hits" and "misses" revealed no effect of the reversal manipulation on observers' performance. Sequences reversed at test were on average labeled correctly no less frequently than the original sequences, and similarly reversed sequences were not rejected as "new" any more frequently than the originals. However, analysis of observers' average RT did yield an interesting

	Forward	Reversed	Forward-Reversed *p*-value (one-tailed, paired-samples *t*-test)
Fast objects	960 ms (70)	1070 ms (90)	$p = 0.05$
Slow objects	890 ms (60)	940 ms (70)	$p = 0.24$

Table 1. Mean RTs by condition for correctly labeled sequences in Experiment 3. Numbers in each cell denote the group mean, with the standard error in parentheses.

effect. Specifically, "fast" objects did suffer from a reversal effect such that observers were significantly slower at test to correctly label reversed sequences than original sequences (Table 1).

Discussion

Our final experiment provides further evidence supporting our conjecture that speed may critically modulate observers' certainty regarding frame-based and sequence-based sources of information regarding object appearance. As we predicted, the reversal effect was more prominent for fast objects than for slow objects, suggesting that sequence-based aspects of appearance ("spatio-temporal signatures") are more prominently relied upon at fast speeds than at slow speeds. At slow speeds, the representation of object appearance may be more heavily weighted towards frame-based information, which is robust to the reversal manipulation.

Our results from Experiment 3 do stand in contrast to earlier studies of the reversal effect in that we only observe significant effects in the response time data obtained from correct classifications. This may be due to our use of clearly visible objects throughout learning and test phases. Previous work describing the reversal effect has been carried out using stimuli that were somehow degraded, either by using only point-lights to describe their form (Stone, 1998), introducing spatio-temporal noise (Vuong & Tarr, 2004), or choosing particular items that are very difficult to discriminate from one another (Vuong & Tarr, 2006). Further investigation of this issue may benefit from a more complete analysis of how image quality and speed interact in producing a robust reversal effect. We note however, that these previous investigations of sequence reversal typically also reported RT differences similar in nature to those that we report here, validating our data relative to the existing literature. At present, our result provides initial support for our basic formulation of object speed as a determiner of uncertainty for distinct aspects of dynamic appearance and lays the foundation for continued investigation of this topic. It also serves as a useful counterpart to the results already reported in Experiments 1 and 2 insofar as we have identified a speed-dependent effect that manifests itself at fast speeds rather than slow speeds. Thus, it is not simply the case that some artifactual aspect of our design favors one speed over another.

General discussion

The experimental results presented here are obviously only a first step towards a comprehensive theory of dynamic object perception and recognition, but they provide important constraints to guide further elaboration of such a theory. In particular, the possibility that moving objects are processed in a speed-dependent fashion is a novel proposal to the best of our knowledge, and raises several interesting questions.

We have thus far attempted to be conservative as to whether or not we believe the distinction between "fast" and "slow" processing is truly categorical in nature or represents two essentially arbitrary points on a continuous spectrum. While we have suggested that these processes may be distinct (involving separate measurements of static appearance and what we have called "relational" properties of images over time) we cannot firmly state at present whether the difference between "fast" and "slow" processing is qualitative or quantitative in nature. That is, does object speed determine which of two independent mechanisms is employed for stimulus encoding, or are we really seeing different aspects of the same general-purpose mechanism? This is a very difficult question to answer within the domain of moving object recognition, complicated in part by an ongoing debate as to whether the visual system employs multiple systems for speed perception in the domain of low-level motion perception. While there is some evidence that the visual system can be well-described in terms of multiple modes of processing "tuned" to different speed ranges (Edwards, Badcock, & Smith, 1998; Khuu & Badcock, 2002), the competing view that a unitary mechanism may subserve motion perception at all speeds remains an open possibility (van Boxtel, van Ee, & Erkelens, 2006). Even if there were more agreement regarding the visual system's organization relative to speed for low-level motion processing, it is also always difficult to extrapolate from studies of low-level perception to high-level visual processes applied to complex objects. We suggest that what is necessary is an ongoing effort to characterize the role of speed and/or acceleration in moving object recognition. Observing behavior across a wider range of matching tasks would provide further insight into the influence of object speed on recognition, as would more detailed parametric investigation of the influence of speed on matching abilities. Manipulating object speed offers a distinct advantage over traditional "scrambling" manipulations of object sequences

in which frames within a coherent sequence are randomly re-arranged on a global scale (Harman & Humphrey, 1999) or within smaller temporal neighborhoods (Vuong & Schultz, 2008). Sequence scrambling remains one of the primary manipulations applied to spatiotemporal visual inputs despite the fact that it introduces very large artifacts into the stimulus. Just as pixel-scrambling destroys a large amount of visual structure that we know is "important" to the human visual system, so too does temporal scrambling likely compromise a wide range of temporally extended structure that is equally important. By contrast, speed manipulations maintain spatiotemporal smoothness to a much higher degree and afford the experimenter more control over the variables under consideration.

Also, as we alluded to earlier, our proposal that the representation of a moving object may be defined by the uncertainty brought on by reduced presentation time as speed increases could easily be explored further. This proposal is broadly consistent with existing results regarding object localization abilities in phenomena like the flash-lag effect (Nijhawan, 1994) and so-called "representational momentum" studies (Freyd & Finke, 1984). In both of these phenomena, observers make systematic (and predictive) errors in estimating object position that are subject to the speed of object motion (Freyd & Finke, 1985). Just as our own results from Experiment 2 indicate that the speed of a complex 3-D object is encoded by the visual system for recognition, these previous studies indicate that the speed of simpler objects is encoded by the visual system for the purposes of spatial localization. The main difference between the models that underlie these results and our own proposal is that both the flash-lag effect and RM phenomena are usually defined in terms of the spatial location of an object within the visual field (x, y position or the orientation of a simple rectangle or other schematic shape). We suggest that analogous mechanisms can be applied to object appearance by recasting positional uncertainty as estimation of object "position" within a high-dimensional appearance space that is scaled by perceptual similarity in a potentially non-linear way. While this proposal certainly introduces some complex experimental issues (most prominently the need to measure how discriminability between static images varies across appearance space), the basic intuition is the same. In both cases, speed limits certainty within the relevant space, be it the physical space of the visual field or the more abstract space of object appearance. Beyond providing important constraints on invariant recognition, our finding of speed-dependency in object recognition thus represents an intriguing bridge between existing results obtained from simple geometric "objects" and complex 3-D forms.

There are also multiple additional invariant properties that remain to be investigated. Again, by analogy with classic investigations of face and object recognition, we have yet to examine how changes in illumination, viewing angle, size, or even simple translation affect the recognition of object motion. As a simple example, our own task required observers to carry out a matching task in the periphery, thereby introducing a slight impairment in form perception due to decreasing acuity. Examining the effects of image blur on matching performance in a systematic way would provide further insights into the trade-off between encoding detailed estimates of appearance and robust spatiotemporal properties of the dynamic stimulus. Finally, exploring how representations of moving objects are learned as an observer gains more experience with a particular set of objects is an additional important issue we have yet to address. How are invariant properties of recognition acquired? All of these questions represent key extensions of this line of research that would provide further empirical data to constrain the set of potential mechanisms that could underlie generic moving object perception.

Conclusions

We have observed in three experiments evidence for speed-dependent processing of dynamic objects. Our first experiment demonstrated that invariance to picture-plane inversion varied as a function of object speed, suggesting the existence of a trade-off between detailed image-specific representation of appearance and potentially coarser, orientation-invariant encoding as a function of speed. Our second experiment confirmed this finding, and additionally found that object speed is encoded for recognition, as evidenced by a significant drop in performance when speed changes between sample and test stimuli. These results provide important constraints on the nature of dynamic object recognition, suggesting that particular stimulus dimensions are encoded robustly only in certain regimes of spatiotemporal appearance change. The emerging description of dynamic object perception is therefore governed by the interaction of spatial processing with spatiotemporal processing, which raises a host of interesting questions for further study.

Acknowledgments

BB was supported by a National Defense Science and Engineering Graduate fellowship. Thanks to Dick Held, Ming Meng, and Meg Moulson for helpful comments on an earlier version of this manuscript.

Commercial relationships: none.
Corresponding author: Benjamin Balas.
Email: Benjamin.Balas@childrens.harvard.edu.
Address: Laboratories of Cognitive Neuroscience, Children's Hospital Boston, 1 Autumn St., AU457, Boston, MA 02115, USA.

References

Balas, B., & Sinha, P. (2008). Observing object motion induces increased generalization and sensitivity. *Perception, 37,* 1160–1174. [PubMed]

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436. [PubMed]

Casile, A., & Giese, M. A. (2005). Critical features for the recognition of biological motion. *Journal of Vision, 5*(4):6, 348–360, http://journalofvision.org/5/4/6/, doi:10.1167/5.4.6. [PubMed] [Article]

Chuang, L., Vuong, Q. C., Thornton, I. M., & Bülthoff, H. H. (2006). Recognising novel deforming objects. *Visual Cognition, 14,* 85–88.

Costen, N. P., Parker, D. M., & Craw, I. (1994). Spatial content and spatial quantisation effects in face recognition. *Perception, 23,* 129–146. [PubMed]

Cox, D. D., Meier, P., Oertelt, N., & DiCarlo, J. J. (2005). 'Breaking' position-invariant object recognition. *Nature Neuroscience, 8,* 1145–1147. [PubMed]

Cutting, J. E. (1987). Perception and information. *Annual Review of Psychology, 38,* 61–90. [PubMed]

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General, 115,* 107–117. [PubMed]

Dollar, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance,* 65–72.

Edwards, M., Badcock, D. R., & Smith, A. T. (1998). Independent speed-tuned global-motion systems. *Vision Research, 38,* 1573–1580. [PubMed]

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, 6,* 443–446.

Galper, R. E. (1970). Recognition of faces in photographic negative. *Psychonomic Science, 19,* 207–208.

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research, 37,* 1673–1682. [PubMed]

Harman, K. L., & Humphrey, G. K. (1999). Encoding 'regular' and 'random' sequences of novel three-dimensional objects. *Perception, 28,* 601–615. [PubMed]

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 1,* 201–211.

Kemp, R., Pike, G., White, P., & Musselman, A. (1996). Perception and recognition of normal and negative faces: The role of shape from shading and pigmentation cues. *Perception, 25,* 37–52. [PubMed]

Khuu, S. K., & Badcock, D. R. (2002). Global speed processing: Evidence for local averaging within, but not across two speed ranges. *Vision Research, 42,* 3031–3042. [PubMed]

Knappmeyer, B., Thornton, I. M., & Bülthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research, 43,* 1921–36. [PubMed]

Kozlowski, L. T., & Cutting, J. E. (1977). Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics, 21,* 575–580.

Liu, T., & Cooper, L. A. (2003). Explicit and implicit memory for rotating objects. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29,* 554–562. [PubMed]

Murase, H., & Nayar, S. K. (1993). Learning and recognition of 3D objects from appearance. *IEEE Workshop on Qualitative Vision,* New York.

Näsänen, R. (1999). Spatial frequency bandwidth used in the recognition of facial images. *Vision Research, 39,* 3824–3833. [PubMed]

Newell, F. N., Wallraven, C., & Huber, S. (2004). The role of characteristic motion in object categorization. *Journal of Vision, 4*(2):5, 118–129, http://journalofvision.org/4/2/5/, doi:10.1167/4.2.5. [PubMed] [Article]

Nijhawan, R. (1994). Motion extrapolation in catching. *Nature, 370,* 256–257. [PubMed]

Nusseck, M., Cunningham, D. W., Wallraven, C., & Bülthoff, H. H. (2008). The contribution of different facial regions to the recognition of conversational expressions. *Journal of Vision, 8*(8):1, 1–23, http://journalofvision.org/8/8/1/, doi:10.1167/8.8.1. [PubMed] [Article]

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442. [PubMed]

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

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

Stone, J. V. (2003). Computer vision: What is the object? In *Prospects for AI, Proc. Artificial Intelligence and Simulation of Behavior,* Birmingham, UK (pp. 199–208). Amsterdam: IOS Press.

Thornton, I. M., Vuong, Q. C., & Bülthoff, H. H. (2003). A chimeric point-light walker. *Perception, 32,* 377–383. [PubMed]

Thurman, S. M., & Grossman, E. D. (2008). Temporal "Bubbles" reveal key features for point-light biological motion perception. *Journal of Vision, 8*(3):28, 1–11, http://journalofvision.org/8/3/28/, doi:10.1167/8.3.28. [PubMed] [Article]

Ullman, S., & Bart, E. (2004). Recognition invariance obtained by extended and invariant features. *Neural Networks, 17,* 833–848. [PubMed]

van Boxtel, J. J., van Ee, R., & Erkelens, C. J. (2006). A single system explains human speed perception. *Journal of Cognitive Neuroscience, 18,* 1808–1819. [PubMed]

Vuong, Q. C., & Schultz, J. (2008). Dynamic objects are more than the sum of their views: Behavioural and neural signatures of depth rotation in object recognition [Abstract]. *Journal of Vision, 8*(6):39, 39a, http://journalofvision.org/8/6/39/, doi:10.1167/8.6.39.

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

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

Wallis, G., & Bülthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences of the United States of America, 98,* 4800–4804. [PubMed] [Article]

Watson, T., Johnston, A., Hill, H., & Troje, N. (2005). Motion as a cue for viewpoint-invariance. *Visual Cognition, 12,* 1291–1308.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81,* 141–145.

Yip, A. W., & Sinha, P. (2002). Contribution of color to face recognition. *Perception, 31,* 995–1003. [PubMed]