Perception, 2006, volume 35, pages 497 – 510

DOI:10.1068/p5491

Structural similarity and spatiotemporal noise effects on learning dynamic novel objects

Quoc C Vuong¶#, Michael J Tarr#

¶ Max Planck Institute for Biological Cybernetics, Spemannstrasse 38, D 72076 Tübingen, Germany; # Department of Cognitive and Linguistic Sciences, Brown University, 190 Thayer Street, Providence, RI 02912, USA; e-mail: quoc.vuong@tuebingen.mpg.de

Received 19 August 2004, in revised form 11 March 2005; published online 15 March 2006

Abstract. The spatiotemporal pattern projected by a moving object is specific to that object, as it depends on both the shape and the dynamics of the object. Previous research has shown that observers learn to make use of this spatiotemporal signature to recognize dynamic faces and objects. In two experiments, we assessed the extent to which the structural similarity of the objects and the presence of spatiotemporal noise affect how these signatures are learned and subsequently used in recognition. Observers first learned to identify novel, structurally distinctive or structurally similar objects that rotated with a particular motion. At test, each learned object moved with its studied motion or with a non-studied motion. In the non-studied motion condition we manipulated either dynamic information alone (experiment 1) or both static and dynamic information (experiment 2). Across both experiments we found that changing the learned motion of an object impaired recognition performance when 3-D shape was similar or when the visual input was noisy during learning. These results are consistent with the hypothesis that observers use learned spatiotemporal signatures and that such information becomes progressively more important as shape information becomes less reliable.

1 Introduction

We live in a dynamic environment. The interplay between our movements relative to other objects and illumination sources produces a continuously changing projection on our retinas. How does our visual system make sense of this visual cacophony to recognize objects? The conventional answer is that the visual system maps dynamic information onto structures that do not vary over time (Marr 1982). For example, popular theories hypothesize that objects are represented as parts and their relations (Biederman 1987; Marr and Nishihara 1978), or as views comprised of visible features (Bülthoff and Edelman 1992; Poggio and Edelman 1990; Tarr and Pinker 1989).

Recently, however, several studies have underscored the need to understand how the visual system directly uses dynamic information for recognition (eg Hill and Pollick 2000; Knappmeyer et al 2003; Lander and Bruce 2000; Liu and Cooper 2003; Newell et al 2004; Stone 1998, 1999; Thornton and Kourtzi 2002; Vuong and Tarr 2004). Many of these studies are motivated by the observation that how visible features change over time is specific to the objects being viewed, as this change depends on both their physical structure (shape and surface appearance) and their movements. Thus, although it is well established that the visual system recovers and refines spatial structures from dynamic visual input (eg Ullman 1984), it is plausible that the visual system also uses the dynamic pattern produced by the movement of that object. Stone (1998) argued that this object-specific dynamic pattern constitutes a *spatiotemporal signature* of the object being viewed, and can therefore provide information that can be used for recognition, in addition to any available shape information. Indeed, studies have shown that dynamic patterns can be used to recognize movements (eg Johansson 1973); to discriminate between male and female actors (eg Mather and Murdoch 1994); to interpret facial expressions (eg Bruce and Valentine 1988); and to recognize individuals (eg Hill and Pollick 2000; Knappmeyer et al 2003; Thornton and Kourtzi 2002), novel

objects (eg Liu and Cooper 2003; Stone 1998, 1999; Vuong and Tarr 2004), or categories of novel objects (Newell et al 2004).

Given the strong evidence that observers use object-specific dynamic patterns for recognition purposes, our goal in the present study was to investigate the conditions under which observers learn to use these spatiotemporal signatures. This is an important issue because the extent to which static and dynamic information is ultimately used in object recognition may depend on how a stimulus class is learned (Wallis and Bülthoff 1999). As highlighted above, investigators have used a wide variety of stimuli (eg faces, human actions, novel 3-D shapes) and an equally wide variety of recognition tasks (eg old/new discrimination, identification, categorization) to study the role of motion in object recognition. Across these different studies, they have consistently found that recognition performance is often affected by subtle changes to the dynamics of the objects. For example, Stone (1998) introduced a *rotation-reversal* manipulation that preserved static cues to object identity (ie 3-D shape and 2-D image features) but disrupted dynamic cues (ie the temporal ordering of views). He reported that this manipulation impaired observers' ability to recognize 'amoebas' rotating rigidly in depth in a complex manner (both in accuracy and response times). Stone replicated his results with point-light displays of the same stimuli (Stone 1999). Liu and Cooper (2003) subsequently reported similar costs for rotation reversal on accuracy in an old/new discrimination task, and on response-time priming in a symmetry judgment task. In their experiments, they used structurally distinctive novel objects rotating about the vertical axis.

In many of these studies, learning the object dynamics is an important component of the study (either during the course of the experiment or from observers' pre-experimental knowledge). Thus beyond demonstrating an important role of motion in object recognition, these previous studies also suggest that learning may shape the visual information that is ultimately used in recognition. However, investigators have not teased apart the factors that may affect the extent to which spatiotemporal signatures are picked-up and used. Here we examined two factors suggested by the literature on object recognition (eg Tarr and Bülthoff 1995): the structural similarity between objects, and the availability of shape and motion information. Both of these factors may make learning the objects more difficult and therefore influence the extent to which their dynamics are used in the recognition process. That is, motion information may be more likely to be used when objects are difficult to learn, as may be the case when the tested objects are highly similar to each other (eg Hayward and Williams 2000) or when objects are visually degraded (eg Lander and Bruce 2000).

Here we used a paradigm similar to that used by Stone (1998, 1999), and Liu and Cooper (2003). Observers first learned to identify either structurally distinct or structurally similar objects that each rotated with a particular motion. Furthermore, the objects could be learned either in the presence or in the absence of spatiotemporal noise. During this learning phase, we developed a training procedure to ensure that observers were given full opportunity to learn shape and motion information. At test, learned objects either moved with their studied motion or with a non-studied motion. In experiment 1, we reversed the studied rotation direction to produce non-studied motion that showed the same studied views of each object (but in reverse order). By comparison, in experiment 2, we presented learned objects moving along a new rotation trajectory that revealed novel views of those objects. Across these two experiments we compared how well observers generalize to novel views of the learned objects (Bülthoff and Edelman 1992). Thus, in contrast to earlier studies, we tested different sets of objects that varied in their structural similarity (and consequently how easily they are recognized) in the same recognition paradigm; we presented objects in the presence or absence of a spatiotemporal noise that presumably degraded both shape

and motion information; and we used a learning procedure to ensure that observers had the opportunity to learn the particular features useful for this challenging task.

2 Experiment 1

Our goal in experiment 1 was to examine the extent to which the *rotation-reversal* effect reported by Stone (1998) and Liu and Cooper (2003) may be influenced by the difficulty of learning the objects. For example, in Stone's (1998) study, observers had the difficult challenge of learning structurally similar amoebas. By comparison, in Liu and Cooper's study, observers had the equally difficult challenge of learning many objects (thirty-two as compared to four objects in Stone's study) from a single exposure and without knowing that their memory for these objects would be subsequently tested. In this experiment, we varied the difficulty of learning in two ways. First, we used either structurally distinctive objects that were 'easy' to recognize or structurally similar objects that were 'hard' to recognize (see Vuong and Tarr 2004). Second, observers could learn either 'easy' or 'hard' objects in the presence or absence of a *dynamic fog* that degraded both shape (including 3-D structure and 2-D views) and motion information. Our working hypothesis is that there should be a larger rotation-reversal effect when the objects are difficult to learn.

2.1 *Method*

2.1.1 *Participants.* A total of forty observers were recruited from the Brown University community (twenty-nine females, eleven males). They participated either for course credit or payment. All observers gave informed consent and were naive to the purposes of the study.

2.1.2 *Stimuli.* Figure 1 shows the two sets of novel 3-D objects used in experiments 1 and 2. These objects were a subset of those used in our earlier study (Vuong and Tarr 2004), and details of their construction can be found in that paper. Each set consisted of eight objects, half of which served as targets and half as distractors.

The first set of eight stimuli consisted of 'easy' objects with structurally distinctive shapes, based on those originally created by Biederman and Gerhardstein (1993). Objects in this set were composed of parts that could be easily discriminated on the basis of non-accidental properties (eg straight versus curved axis of elongation; see Biederman 1987). In contrast, the second set of eight stimuli consisted of 'hard' amoebas with structurally similar shapes (they lacked distinctive parts or features that could be easily used as identity cues), similar to those used in several previous studies of human object recognition (Bülthoff and Edelman 1992; Stone 1998, 1999). The 3-D coordinates of the vertices of each object and their associated surface normals were imported into custom software that rendered the objects with a matte-gray surface. The objects were illuminated by several light sources. All objects were rendered against a black background.

Two trajectories were used to generate 128-frame (2.8°/frame) animations for both 'easy' and 'hard' objects. For the first trajectory, a virtual camera was arbitrarily rotated about the three axes controlled by a parameter t that varied from 0° to 360°. The virtual camera was rotated in the same fashion for the second trajectory but t varied from 360° to 0°. It is important to point out that this produced a completely different trajectory of the same complexity as the first. When either image sequence is presented in increasing frame order, objects appear to tumble in depth with a coherent rotation direction. The same image sequence played in decreasing order depicts each object rotating in the opposite direction. We arbitrarily animated all objects using the first trajectory for all phases of experiment 1. The animations were played at ~50 ms/frame (roughly three screen refreshes), so that it took ~6500 ms to play one entire 360° rotation.

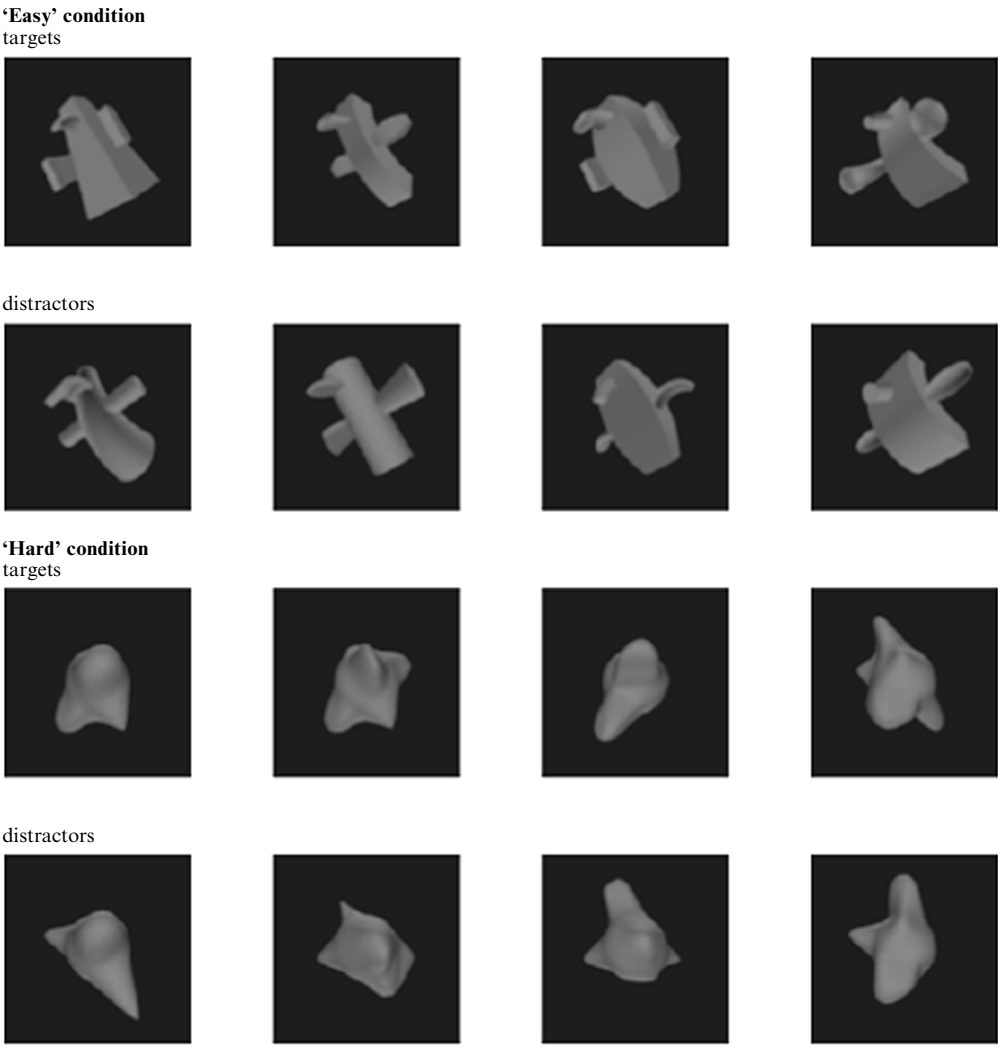

Figure 1. The sets of 'easy', structurally distinct, and 'hard', structurally similar, objects used in experiments 1 and 2.

Finally, in some conditions, we presented objects rotating in a dynamic fog to degrade both spatial and dynamic information. The fog consisted of a pre-computed 3-D fractal noise volume (Perlin 1985). By presenting 2-D slices of this volume on each frame, we were able to smoothly mask random fragments of the rotating object in space and time. On trials when the dynamic fog was presented (for both learning and test phases), we randomly selected a subset of frames and cycled back and forth through these frames on that trial. Thus, the dynamics of the fog was completely independent of the dynamics of the objects. Figure 2 illustrates a time sequence of an object rotating in this fog.

2.1.3 *Design and procedure.* Four main factors were tested in experiment 1 in a mixed design with object type ('easy', 'hard') and learning context (fog, no-fog) as between-participants factors, and test motion (studied, non-studied) and test context (fog, no-fog) as within-participants factors. Ten observers were run in each of the four between-participants conditions.

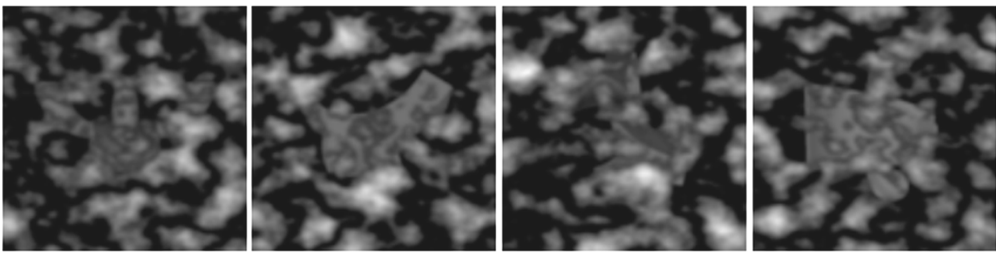

Figure 2. An example sequence of the dynamic fog. Note that the 'easy' object is difficult to see in any particular image. However, when the sequence is animated, the object is easily seen in the dynamic fog. Note also that the dynamics of the fog is independent of the dynamics of the object.

Experiment 1 consisted of two learning phases followed by a test phase. In the first learning phase, observers were shown four objects individually for a full 360° rotation (~6500 ms). To eliminate any effects of seeing a new rotation direction during the test phase, two targets rotated clockwise and the other two rotated counterclockwise (by playing the animation sequence either forwards or backwards). Each rotation direction of the object was randomly determined for each observer at the beginning of the experiment, which established its particular characteristic motion learned by that observer. The starting frame was selected randomly on each trial. Observers were instructed to press the appropriate key for each object after seeing the object make a complete rotation. They were informed that they could not respond until the object was removed from the screen. If observers responded incorrectly, they heard a low 500 Hz tone, and the correct response key was presented on the screen. If they responded correctly, they heard a high 1000 Hz tone. For this phase, observers were instructed to respond as accurately as possible. Each object was presented 30 times for a total of 120 trials. There was a short self-timed break after every 40 trials.

The second learning phase was the same as the first, with the following two exceptions. First, observers were instructed to respond as quickly and as accurately as possible. Thus, in this phase, they did not have to wait for the object to disappear from the screen before responding. Second, if observers responded incorrectly, they heard only the low tone. As in the first learning phase, each object was presented 30 times for a total of 120 trials, with self-timed breaks after every 40 trials.

In the test phase, observers were presented with the four studied targets intermixed with four unstudied distractors. They were instructed to press the space bar for all distractors, and to continue to respond with the learned letter key associated with each target. During this phase, all objects (both targets and distractors) appeared both in the presence and in the absence of the dynamic fog and rotated in both rotation directions. Thus, on 50% of the test trials, the targets rotated with their studied motion (established during the learning phase), and on the remaining 50% of the trials, they rotated in a non-studied motion (in which the studied motion was reversed). Observers were not informed that the targets would rotate any differently than before. Targets and distractors were shown 10 times in each condition during the test phase, for a total of 320 trials [8 objects (4 targets/4 distractors) $\times$ 2 test contexts $\times$ 2 test motions]. There was a short break after every 40 trials. As in the second learning phase, observers were instructed to respond as quickly and as accurately as possible. No feedback was provided during this phase. The entire experiment took approximately 45 min.

The experiment was run on a Windows PC with a monitor with a 1280×1024 pixel resolution and a 60 Hz refresh rate. The program to present the movies and collect responses was written in C and relied on the OpenGL 1.2 interface to the PC's graphics hardware. Observers sat approximately 50 cm from the monitor. At this viewing distance, each object subtended a maximum visual angle of ~9 deg. The dynamic fog, when

present, filled the entire screen. Responses were collected from the keyboard. The four keys used were 'v', 'b', 'n', and 'm', which were randomly assigned to the four targets for each observer. All observers were instructed to respond with their dominant hand.

2.2 *Results*

Our main focus in the present study is the effect of rotation reversal following learning. Thus, only results of target trials from the test phase were analyzed. For our main analysis, accuracy and correct response times (RTs) for targets presented during the test phase were submitted to a mixed-design analysis of variance (ANOVA) with object type ('easy', 'hard') and learning context (fog, no-fog) as between-participants factors, and test motion (studied, non-studied) and test context (fog, no-fog) as within-participants factors. Response times outside the range of 400 and 6500 ms in this and the subsequent experiment were removed to eliminate anticipatory responses and outliers. This procedure excluded less than 4% of correct trials in both experiments.

For both experiments, we also conducted additional non-parametric tests to determine whether object type and/or learning context influenced the effect of rotation reversal on object recognition. As hypothesized above, both of these factors may affect the difficulty of learning the objects. For these analyses, we first computed a mean percentage of correct responses or RT score for studied and non-studied motion (averaged across test contexts) on a per-observer basis. We then used the Wilcoxon signed rank test to compare the population distribution of these dependent measures for the four different between-participants conditions ('easy' objects learned in fog and no-fog contexts, and 'hard' objects learned in fog and no-fog contexts). A significance level of 0.05 was adopted for all analyses reported.

2.2.1 *Accuracy data.* The mean percentage of correct responses for experiment 1 is plotted in figure 3a as a function of object type and test motion. We plotted this interaction throughout this study because it was the most robust finding. For both experiments, observers performed well above chance levels (20%) in all conditions. Furthermore, there was no indication of any speed–accuracy trade-offs in the data.

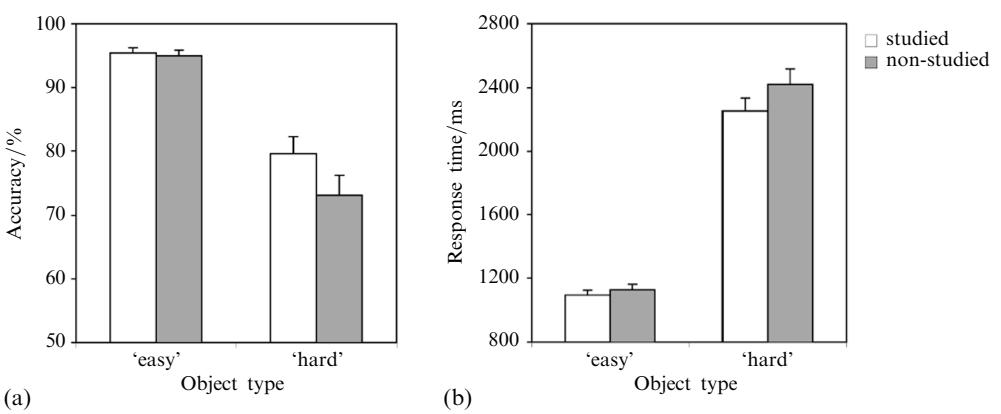

Figure 3. (a) Mean percentage-correct scores in experiment 1 as a function of object type and test motion. Error bars in this and subsequent figures reflect +1 SEM. (b) Mean correct RTs in experiment 1.

In experiment 1, we found main effects of object type ($F_{1,36} = 21.04$, $p < 0.001$) and test motion ($F_{1,36} = 20.31$, $p < 0.001$) on observers' accuracy. As evident in the figure, there was also a significant interaction between object type and test motion ($F_{1,36} = 15.08$, $p < 0.001$), suggesting that the effect of rotation reversal on observers' accuracy was modulated by the structural similarity of the objects. Lastly, there was no significant interaction between learning context and test motion, and no significant

three-way interaction between object type, learning context, and test motion (ps > 0.18). The ANOVA indicates that only object type modulated the effect of rotation reversal on how accurately observers identified targets. This was further confirmed by the non-parametric tests. For 'easy' objects, there was no effect of rotation reversal in the fog learning context ($p = 0.91$) and in the no-fog learning context ($p = 0.38$). For 'hard' objects, there was a rotation-reversal effect in the fog learning context ($p = 0.008$) and marginally in the no-fog learning context ($p = 0.06$).

2.2.2 *Response-time data.* The mean RTs for experiment 1 are plotted in figure 3b. For RTs, all main effects were significant: object type ($F_{1,36} = 132.52$, $p < 0.001$), test motion ($F_{1,36} = 18.89$, $p < 0.001$), learning context ($F_{1,36} = 9.18$, $p < 0.01$), and test context ($F_{1,36} = 111.98$, $p < 0.001$). As for the accuracy data, there was evidence that structural similarity modulated the effect of rotation reversal on RTs, as indicated by the significant interaction between object type and test motion ($F_{1,36} = 8.74$, $p < 0.01$).

There was also some evidence that the availability of shape and motion information modulated the effect of rotation reversal on RTs, but this did not reach significance in the omnibus ANOVA [the interaction between learning context and test motion was marginally significant ($F_{1,36} = 2.55$, $p = 0.12$) and the three-way interaction between object type, learning context, and test motion was not significant ($F < 1$)]. However, the trend in the RT data is in this direction, as shown in table 1. The non-parametric tests provide further statistical evidence for this trend: for 'easy' objects, rotation reversal slowed observers' responses during the test phase if they initially learned the objects in the fog context ($p = 0.02$) but not when they learned the objects in the no-fog context ($p = 0.92$). By comparison, for 'hard' objects, rotation reversal impaired RTs irrespective of the learning context ($p = 0.04$ in both contexts).

Table 1. Mean percentage of correct responses and RTs (SEM in parentheses) as a function of object type, learning context, and test motion for experiments 1 and 2.

Object type	Learning context	Test motion	Correct responses/%	RT/ms
Experiment 1				
Easy	no-fog	studied	97.0 (0.8)	1047 (43)
		non-studied	96.1 (1.1)	1047 (42)
	fog	studied	93.9 (1.4)	1150 (42)
		non-studied	93.8 (1.5)	1213 (55)
Hard	no-fog	studied	81.5 (4.1)	2020 (97)
		non-studied	76.8 (4.8)	2147 (115)
	fog	studied	77.7 (3.5)	2488 (110)
		non-studied	69.6 (4.0)	2699 (132)
Experiment 2				
Easy	no-fog	studied	95.5 (0.9)	1069 (49)
		non-studied	96.5 (0.7)	1105 (54)
	fog	studied	97.4 (0.6)	1040 (28)
		non-studied	95.8 (0.8)	1112 (34)
Hard	no-fog	studied	81.4 (3.9)	2012 (128)
		non-studied	58.0 (5.6)	2211 (131)
	fog	studied	87.0 (1.5)	2202 (115)
		non-studied	64.5 (3.0)	2620 (139)

2.3 *Discussion*

Our present results replicate Stone's (1998, 1999) original findings with a different learning procedure. For 'hard' objects similar to Stone's stimuli, rotation reversal was clearly detrimental to observers' recognition performance, as measured by both accuracy and

response times. We also found an effect of rotation reversal only on response times for 'easy' objects when these were initially learned in the presence of dynamic noise. Liu and Cooper (2003), on the other hand, found an effect of rotation reversal only on accuracy using stimuli similar to our 'easy' objects. The different response measure (RT versus accuracy) affected by rotation reversal in our study and theirs may reflect the different methods used to create a difficult learning context. We used spatiotemporal noise to degrade the availability of shape and motion cues during learning. This degradation may have impaired how quickly observers used shape and motion cues in general. However, accuracy was not affected because these objects can be accurately discriminated on the basis of their shape. Liu and Cooper, on the other hand, had observers study many objects from a single exposure. This procedure may have impaired how accurately observers could encode the different objects.

Overall, we believe that the present data support the hypothesis that observers learn to use spatiotemporal signatures to recognize objects when recognition is difficult during the learning phase, as when objects have similar shape or when both shape and motion information are degraded. This hypothesis is consistent with the results reported for the recognition of dynamic faces (for a review see O'Toole et al 2002).

3 Experiment 2

Experiment 1 indicates that when the recognition task was made difficult owing to the similarity of the targets or the presence of spatiotemporal noise, rotation reversal was more likely to disrupt the learned motion associated with each target object. Critically, we held constant the shape and surface information available for the recognition task—both the studied motion and its reversal contained the same views of each object. The purpose of experiment 2 was twofold. First, we wanted to contrast this 'pure' effect of rotation reversal with the more ecological situation of showing a completely different motion trajectory (and that consequently showed novel views). Second, we wanted to test whether observers could generalize studied views acquired during learning to novel views at test in order to compensate for changes to the studied motion (eg Bülthoff and Edelman 1992). Here observers learned the same objects as those in experiment 1 in one of two possible motion sequences and were then tested with both sequences. In contrast to the first experiment, these two sequences changed the studied motion and revealed different sets of 128 views of each object.

3.1 *Method*

3.1.1 *Participants.* A new group of forty members of the Brown University community (twenty-seven females, thirteen males) were recruited as observers for either course credit or payment. All observers gave informed consent and were naive to the purpose of the study.

3.1.2 *Design and procedure.* The design of experiment 2 was identical to that of experiment 1. Procedurally, two critical changes were made to the learning and test phases. First, for both learning phases, two targets were animated with one animation sequence in which the parameter that controlled the virtual camera, t, varied from 0° to 360° (the same trajectory as that used in experiment 1), and the remaining two targets were animated with the second animation sequence, in which t varied from 360° to 0°. We reiterate that both trajectories were therefore of equal complexity. As in the previous experiment, the assignment of a given animation sequence to a given target object was randomly determined for each observer. Second, during the test phase, observers were shown each target (and distractor) in both animation sequences. Thus, on 50% of the trials, targets rotated with their studied motion, and on the remaining 50% of the trials, targets rotated with a non-studied motion.

3.2 *Results*

As in experiment 1, the main focus in experiment 2 was whether changing the studied motion affected recognition performance for 'easy' and 'hard' objects. As before, accuracy and RT data obtained from target trials were submitted to the same mixed-design ANOVA as that used in the first experiment. We also conducted non-parametric tests comparing differences between studied and non-studied motion across the different between-participants conditions.

3.2.1 *Accuracy data.* The accuracy data for experiment 2 are shown in figure 4a. As in experiment 1, there were significant effects of object type ($F_{1,36} = 51.25$, $p < 0.001$), motion type ($F_{1,36} = 53.65$, $p < 0.001$), and their interaction ($F_{1,36} = 50.47$, $p < 0.001$). Although the omnibus ANOVA revealed a similar pattern of results in the accuracy data across the two experiments (see the figures 3a and 4a), the non-parametric tests revealed a slightly different pattern. Here for 'easy' objects, there was an effect of changing the studied motion in the fog learning context ($p = 0.03$), which was not found in the first experiment in which the studied motion was simply reversed. However, there was no effect of changing the studied motion in the no-fog learning context ($p = 0.92$) for 'easy' objects; and there was an effect of changing the studied motion in both the fog and no-fog learning contexts ($p = 0.002$ for both conditions) for 'hard' objects.

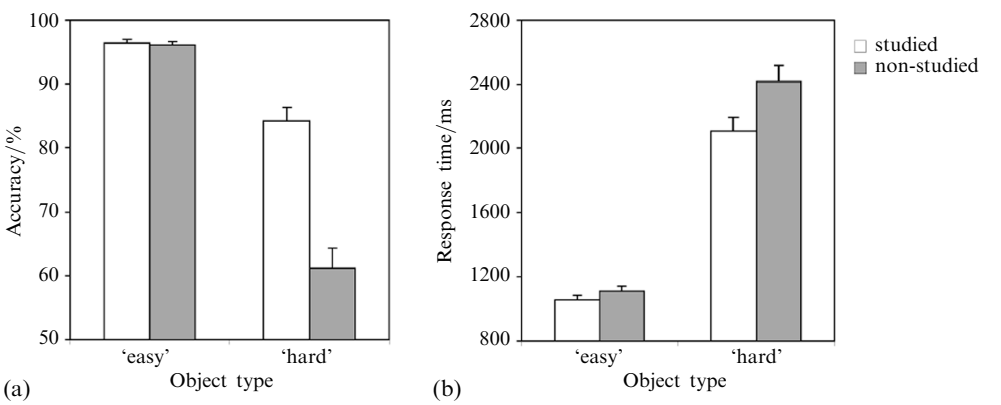

Figure 4. (a) Mean percentage-correct scores in experiment 2. (b) Mean correct RTs in experiment 2.

3.2.2 *Response-time data.* The mean RTs for experiment 2 are shown in figure 4b. Again, for RTs, object type ($F_{1,36} = 93.99$, $p < 0.001$), test motion ($F_{1,36} = 42.40$, $p < 0.001$), and test context ($F_{1,36} = 101.65$, $p < 0.001$), were significant, as was the interaction between object type and test motion ($F_{1,36} = 21.05$, $p < 0.001$). There were also significant interactions between learning context and motion type ($F_{1,36} = 5.26$, $p < 0.05$), and between text context and motion type ($F_{1,36} = 4.68$, $p < 0.05$). Lastly, the three-way interaction between object-type, learning context, and test motion was marginally significant ($F_{1,36} = 2.69$, $p = 0.11$) as was the interaction of all four factors ($F_{1,36} = 3.07$, $p = 0.09$). In contrast to experiment 1, the omnibus ANOVA indicated that test context had more effect on how quickly observers could identify studied objects. However, for both 'easy' and 'hard' objects, changing the studied motion so that observers saw novel views of the studied objects slowed their responses in both learning contexts. The results of the non-parametric tests were consistent with this interpretation. For 'easy' objects, changing the studied motion slowed observers' responses in both learning contexts (fog: $p = 0.01$; no-fog: $p = 0.02$). For 'hard' objects, changing the studied motion impaired RTs in the fog learning context ($p = 0.002$). However, in the no-fog learning context, changing the studied motion did not significantly impair

response times ($p = 0.11$). We cannot account for this non-significance, but the trend in the RTs is again in the right direction, as shown in table 1.

3.3 *Discussion*

In experiment 2, we changed both the studied motion and the views that were seen. Despite this critical difference, the results of experiment 2 replicate those of experiment 1. For 'hard' objects, there was an effect for changing the studied motion at test and no interactions with learning context for both accuracy and response times. For 'easy' objects, we found an overall effect of changing the studied motion, and no interaction with whether observers learned these objects in spatiotemporal noise or not. As in experiment 1, changing the studied motion affected response times but not accuracy (see table 1). Again we believe that learning 'easy' structurally distinct objects in spatiotemporal noise influenced how observers processed dynamic information, rather than whether they encoded dynamic information or not. We have shown elsewhere that observers are sensitive to the direction of rotation even for these stimuli, suggesting that the dynamics of the objects are, in fact, encoded in the object representation (Vuong and Tarr 2004).

Many researchers have suggested that seeing many different views of an object facilitates the development of object representations that are invariant to image variations arising from changes in viewpoint or object pose, particularly if differing views are linked in a temporally ordered sequence that gives rise to apparent motion (eg Kourtzi and Shiffrar 1999; Lawson et al 1994; Wallis and Bülthoff 2001; but see Harman and Humphrey 1999). However, it is not known precisely what form this invariant representation may take. Here we found that observers were unable to use the novel views to compensate for changes to the motion direction: their performance was impaired by the novel motion, and they generally made more errors and responded more slowly in this experiment compared to experiment 1 (see table 1). Thus, across both experiments, our results suggest that observers concurrently encode spatial and dynamic information in the form of spatiotemporal signatures, much in the same manner that observers encode multiple views of static objects experienced from different viewpoints (Tarr and Pinker 1989).

4 General discussion

The two experiments reported here were motivated by the question: How does object motion affect observers' ability to recognize objects? The results presented here converge with previous studies demonstrating that observers use spatiotemporal signatures to recognize dynamic stimuli (eg Hill and Pollick 2000; Knappmeyer et al 2003; Newell et al 2004). As we, and others, have shown, changing the studied motion can severely impair observers' recognition performance, even if 3-D shape and 2-D views are fully preserved (Liu and Cooper 2003; Stone 1998, 1999).

Our contribution to this growing literature is that we tested factors that affect how learned spatiotemporal signatures are ultimately used for recognition purposes; namely, we examined the structural similarity of the target objects and the availability of shape information during learning. Both of these factors have been found to affect the recognition of objects presented as static images (eg Biederman and Gerhardstein 1993; Hayward and Williams 2000), and thus we predicted that they would also mediate the recognition of moving objects. The present results are consistent with this prediction. In experiment 1, we found interactions between structural similarity and availability of shape information in modulating the rotation-reversal effect (Liu and Cooper 2003; Stone 1998, 1999). For 'easy', structurally distinctive, objects, observers' response times were affected by rotation reversal only if they had studied the objects in a noisy context. By comparison, with 'hard', structurally similar, objects, accuracy and response times

were affected by a rotation reversal in both learning contexts, probably because these stimuli were already difficult to recognize (see also Vuong and Tarr 2004).

In experiment 2, the studied motion of each object was changed so that observers saw novel views of the targets (but the same 3-D structure). In this case, recognition performance decreased when novel views of studied objects were introduced, for both 'easy' and 'hard' objects and in both learning contexts (as measured by an increase in response times). However, a direct comparison across the two experiments may be difficult because the new trajectory used in experiment 2 (but not in experiment 1) may have made it more difficult to recognize objects even though both trajectories were constructed to have the same complexity.

We focused on the learning component in the present study because in previous studies it had not been systematically investigated how the difficulty of learning objects may influence the visual information observers use to recognize those objects. Indeed, as far as we can tell, the test stimuli or learning context used in previous studies generally made it difficult to learn the stimuli. The stimuli formed a homogeneous class (eg faces, amoebas, arm movements); they were degraded in some manner (eg shown as point-light displays); or observers had to learn many items from limited exposures. Our strategy to address this issue was to test qualitatively different types of stimuli and learning contexts using a difficult individual-level identification task.

It is important to point out that we used a procedure to provide observers with every opportunity to learn both static (3-D shape and 2-D views) and dynamic (spatiotemporal signatures) cues to recognize the target objects. In our learning procedure, observers were initially required to see the entire rotation sequence (first learning phase). Following that, they were encouraged to respond as quickly and as accurately as possible (second learning phase). They were provided with explicit feedback throughout both learning phases. Other researchers have used different learning procedures. For example, Knappmeyer et al (2003) found that observers learned very subtle characteristic facial movements of specific individuals incidentally. In their study, observers were merely exposed to animated faces and asked to answer questions about each individual (eg "which person is more friendly?"). Similarly, Liu and Cooper (2003) had observers incidentally learn their set of objects by having them decide whether the object could be used for support or as a tool. Other researchers have used famous or well-known individuals so that observers would be familiar with the idiosyncratic movements of those individuals from normal experience (eg Cutting and Kozlowski 1977; Lander and Bruce 2000). Indeed, a possible avenue for future research is to test different learning procedures; for example, whether observers learn objects incidentally or explicitly.

We also acknowledge that there is one potential confound in the present study that provides an alternative account of our data. During the test phase, it is possible that observers were surprised when learned objects moved in a different manner, and this could have caused them to make more errors or respond more slowly. To address this issue, we divided the accuracy and RT data in experiments 1 and 2 into two blocks and looked at the results only on the second block. That is, we only looked at the last five (out of 10) presentations of each target object and in each test condition. We assumed that any surprise effects should have disappeared by the second half of the test phase. The results are presented in table 2. A comparison of tables 1 and 2 shows a similar pattern of results for both experiments. Thus, even after seeing learned objects moving with both studied and non-studied motions, and in the absence and presence of the dynamic fog, observers were still sensitive to the dynamics of the objects acquired during the learning phase.

Lastly, our data suggest that structural similarity and availability of stimulus information (shape and motion cues) had different effects on how observers ultimately used spatiotemporal information, although both factors generally made the recognition task

Table 2. Mean percentage of correct responses and RTs (SEM in parentheses) of experiments 1 and 2 as a function of object type, learning context, and test motion. The means were computed from the last five trials of each studied object presented in each condition during the test phase.

Object type	Learning context	Test motion	Correct responses/%	RT/ms
Experiment 1				
Easy	no-fog	studied	98.3 (0.7)	995 (41)
		non-studied	97.8 (0.7)	1007 (41)
	fog	studied	95.2 (1.6)	1100 (40)
		non-studied	95.0 (1.7)	1114 (41)
Hard	no-fog	studied	81.9 (4.8)	2027 (122)
		non-studied	77.2 (5.3)	2163 (143)
	fog	studied	80.6 (3.9)	2438 (101)
		non-studied	70.4 (4.2)	2485 (134)
Experiment 2				
Easy	no-fog	studied	96.7 (0.8)	1027 (49)
		non-studied	97.5 (0.7)	1025 (51)
	fog	studied	97.7 (0.7)	972 (32)
		non-studied	96.2 (1.0)	1013 (33)
Hard	no-fog	studied	77.6 (4.3)	2038 (137)
		non-studied	58.5 (5.8)	2161 (147)
	fog	studied	86.8 (1.6)	2112 (118)
		non-studied	65.8 (3.2)	2461 (144)

more difficult during learning. However, as is often the case for static objects, structural similarity seemed to be the critical factor in our study (Tarr and Bülthoff 1995). The availability of shape and motion information, on the other hand, had a weak effect that was evident only in comparing differences between studied and non-studied motion (mostly) RT distributions.

For the two types of objects we used, the results suggest that spatial and dynamic information may be weighted differently (see Foster and Gilson 2002, for how object parts and views are summed in object recognition). In our experiments, spatial and dynamic information about structurally similar objects may have been equally weighted in the object representation because motion information would help observers discriminate between visually similar objects. By comparison, shape information may have been weighted more than motion information for structurally distinct objects, as these can be accurately and quickly identified on the basis of shape. Consequently, for 'hard' objects, changing the studied motion would impair both accuracy and response times in both the absence and presence of spatiotemporal noise. In contrast, for 'easy' objects, the presence of spatiotemporal noise during learning may affect how quickly shape and motion cues are used. In any case, future studies will be required to further explore this important issue. Our data provide a starting point to further investigate precisely how learning affects the combination of shape and motion cues for object recognition, so that appropriate cue-combination models may be formulated. For example, it would be interesting in the future to systematically vary the structural similarity of targets from 'easy' to 'hard'.

In summary, we used rotation reversal (Liu and Cooper 2003; Stone 1998, 1999) as a means to investigate the information observers use to recognize dynamic objects. Combined with earlier results, our present findings suggest that how an object's movements unfold over time also contributes to the recognition of that object. That is, observers can learn to directly associate specific dynamic information with specific objects (or classes of objects), particularly if this is informative with regard to object identity. Hence, the term 'signature' is appropriate: spatiotemporal signatures capture space-time structures projected onto our retinas by a dynamic world.

Acknowledgment. QCV was supported by a Natural Sciences and Engineering Research Council of Canada scholarship (NSERC).

References

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

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

Bruce V, Valentine T, 1988 "When a nod's as good as a wink: The role of dynamic information in facial recognition", in *Practical Aspects of Memory: Current Research and Issues* volume 1, Eds N N Gruneberg, P E Morris, R N Sykes (Chichester, UK: Wiley and Sons) pp 169 – 174

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 of the USA* **89** 60 – 64

Cutting J, Kozlowski L, 1977 "Recognizing friends by their walk: Gait perception without familiarity cues" *Bulletin of the Psychonomic Society* **9** 353 – 356

Foster D H, Gilson S J, 2002 "Recognizing novel three-dimensional objects by summing signals from parts and views" *Proceedings of the Royal Society of London B* **269** 1939 – 1947

Harman K L, Humphrey G K, 1999 "Encoding 'regular' and 'random' sequences of views of novel three-dimensional objects" *Perception* **28** 601 – 615

Hayward W G, Williams P, 2000 "Viewpoint dependence and object discriminability" *Psychological Science* **11** 7 – 12

Hill H, Pollick F E, 2000 "Exaggerating temporal differences enhances recognition of individuals from point light display" *Psychological Science* **11** 223 – 227

Johansson G, 1973 "Visual perception of biological motion and a model for its analysis" *Perception & Psychophysics* **14** 201 – 211

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 – 1936

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

Lander K, Bruce V, 2000 "Recognizing famous faces: Exploring the benefits of facial motion" *Ecological Psychology* **12** 259 – 272

Lawson R, Humphreys G W, Watson D G, 1994 "Object recognition under sequential viewing conditions: evidence for viewpoint-specific recognition procedures" *Perception* **23** 595 – 614

Liu T, Cooper L A, 2003 "Explicit and implicit memory for rotating objects" *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20** 554 – 562

Marr D, 1982 *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (San Francisco, CA: W H Freeman)

Marr D, Nishihara H K, 1978 "Representation and recognition of the spatial organization of three-dimensional shapes" *Proceedings of the Royal Society of London B* **200** 269 – 294

Mather G, Murdoch L, 1994 "Gender discrimination in biological motion displays based on dynamic cues" *Proceedings of the Royal Society of London B* **258** 273 – 279

Newell F N, Wallraven C, Huber S, 2004 "The role of characteristic motion in object categorization" *Journal of Vision* **4** 118 – 129

O'Toole A J, Roark D A, Abdi H, 2002 "Recognizing moving faces: a psychological and neural synthesis" *Trends in Cognitive Sciences* **6** 261 – 266

Perlin K, 1985 "An image synthesizer" *Computer Graphics* **19** 287 – 296

Poggio T, Edelman S, 1990 "A network that learns to recognize three-dimensional objects" *Nature* **34** 263 – 266

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, 1995 "Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993)" *Journal of Experimental Psychology: Human Perception and Performance* **21** 1494 – 1505

Tarr M J, Pinker S, 1989 "Mental rotation and orientation-dependence in shape recognition" *Cognitive Psychology* **21** 233 – 282

Thornton I M, Kourtzi Z, 2002 "A matching advantage for dynamic human faces" *Perception* **31** 113 – 132

Ullman S, 1984 "Maximizing rigidity: The incremental recovery of 3-D structure from rigid and rubbery motion" *Perception* **13** 255 – 274

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

Wallis G, Bülthoff H H, 1999 "Learning to recognize objects" *Trends in Cognitive Sciences* **3** 22 – 31

Wallis G, Bülthoff H H, 2001 "Effects of temporal association on recognition memory" *Proceedings of the National Academy of Sciences of the USA* **98** 4800 – 4804

ISSN 0301-0066 (print) ISSN 1468-4233 (electronic)

PERCEPTION
VOLUME 35 2006

www.perceptionweb.com