Perception & Psychophysics
2006, 68 (7), 1082-1093

A stereo advantage in generalizing over changes in viewpoint on object recognition tasks

DAVID J. BENNETT
Brown University, Providence, Rhode Island

and

QUOC C. VUONG
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

In four experiments, we examined whether generalization to unfamiliar views was better under stereo viewing or under nonstereo viewing across different tasks and stimuli. In the first three experiments, we used a sequential matching task in which observers matched the identities of shaded tube-like objects. Across Experiments 1–3, we manipulated the presentation method of the nonstereo stimuli (having observers wear an eye patch vs. showing observers the same screen image) and the magnitude of the viewpoint change (30º vs. 38º). In Experiment 4, observers identified "easy" and "hard" rotating wire-frame objects at the individual level under stereo and nonstereo viewing conditions. We found a stereo advantage for generalizing to unfamiliar views in all the experiments. However, in these experiments, performance remained view dependent even under stereo viewing. These results strongly argue against strictly 2-D image-based models of object recognition, at least for the stimuli and recognition tasks used, and suggest that observers used representations that contained view-specific local depth information.

We easily recognize many familiar and unfamiliar objects that vary in shape, color, texture, movements, and so on. Although any or all of these properties can be used for recognition, it is largely assumed that recognition is based predominantly on matching shapes that are recovered from the visual input to shapes that are encoded in short- and long-term visual memory. This assumption has several motivations. First, shape can be derived from different sources of visual information, such as motion or stereo information (Bülthoff, 1991; Marr, 1982). Second, because of multiple inputs to the shape representation, shape is robust to changes to or degradation of the visual input. Finally, in most circumstances, shape can be used to reliably identify objects (see, e.g., Hayward, 1998).

Despite the importance of shape for object recognition, how 3-D shape is represented for recognition remains elusive (Bülthoff, Edelman, & Tarr, 1995). In this regard, one outstanding issue is the extent to which the object representation encodes object-centered 3-D depth and structure (see, e.g., Marr & Nishihara, 1978) as opposed to viewer-centered 2-D views (e.g., Poggio & Edelman, 1990). Another issue is the possibility that the object representation encodes some intermediate shape representation, such as view-invariant qualitative parts (e.g., Biederman, 1987) or view-specific local depth of visible surface patches, such as Marr's (1982) 2.5-D sketch (see also Edelman & Bülthoff, 1992; Williams & Tarr, 1999).

Building on previous work (Edelman & Bülthoff, 1992; Farah, Rochlin, & Klein, 1994; Humphrey & Khan, 1992), in the present study we examined the role of stereo information in object recognition, since this is a strong source of information about 3-D depth and structure, alone or in combination with other depth cues (see, e.g., Bülthoff, 1991; Bülthoff & Mallot, 1988; Landy, Maloney, Johnston, & Young, 1995). Specifically, we examined whether the addition of stereo information facilitates the recognition of objects when they are presented at an unfamiliar viewpoint or at a familiar viewpoint. We did not test novel objects with distinctive part structure, which is often found in real world objects (Biederman, 1987). Rather, we varied the recognition task and stimuli in other important ways over four experiments in an effort to explore at least some of the conditions under which the visual system may encode depth and 3-D structure information. Our secondary aim was to compare our results with those of previous studies in which similar novel objects with no distinctive part structure were used.

Edelman and Bülthoff (1992) initially found that subjects were more accurate at recognizing novel objects under stereo than under nonstereo viewing. Their stimuli were computer-generated wire forms constructed by join-

Part of this research was supported by a postdoctoral research grant to D.J.B. from the Institute for Research in Cognitive Science, University of Pennsylvania, and a Natural Sciences and Engineering Research Council of Canada postgraduate scholarship to Q.C.V. Thanks to Mike Tarr and especially Zili Liu for helpful discussion. Thanks to Nick Foster for programming assistance (regarding aspects of the general spline surface interpolation software used in Experiments 1–3). Thanks to Bill Warren, Mike Tarr, and Fulvio Domini for lab support. The order of authors was determined alphabetically. Address correspondence to D. J. Bennett, Department of Cognitive and Linguistic Sciences, P.O. Box 1978, Brown University, Providence, RI 02912 (e-mail: david_bennett@brown.edu).

ing thin straight tubes together end to end. This stereo advantage was found across a range of viewpoint changes up to 120º from trained viewpoints. A similar advantage, however, was also found for the trained viewpoints (i.e., a 0º change in viewpoint), suggesting that stereo information did not improve view generalization, but did improve overall recognition performance. These findings could have been due to two aspects of their design. First, during training, objects were always shown in stereo, whereas during testing the learned targets were shown in both stereo and nonstereo presentations in randomly intermixed trials. As a result, the overall stereo advantage may have been due to the viewing condition mismatch between the training and testing trials (see also Sinha & Poggio, 1994). Second, with the objects used, subjects already generalized well under nonstereo viewing (e.g., miss rates of around 20% in their Experiment 4). A stereo advantage in view generalization may be evident only when subjects find it difficult to generalize to unfamiliar views under nonstereo viewing. What is clear in Edelman and Bülthoff's data, as they pointed out in their conclusion, is that 3-D depth specified by stereo information is encoded in a view-sensitive fashion. Our present data lend further support to this claim.

Other investigators have found indirect evidence for a stereo advantage in view generalization across a range of novel objects. Farah et al. (1994) compared subjects' ability to generalize to unfamiliar views of thin, smooth, potato-chip-like surfaces and their wire form outlines presented as either real objects seen from a close distance or as video tape recordings. These investigators found that subjects performed better at generalizing to unfamiliar views when presented with physical objects, presumably because they had access to stereo information about 3-D shape. That said, Farah et al. did not provide any statistical justification for this conclusion, and their experiments do not readily admit a stereo-versus-nonstereo comparison (which was not the primary aim of their study); for example, different initial orientations and different rotations were used across the stereo-versus-nonstereo experiments.

Humphrey and Khan (1992) also found evidence suggestive of a stereo advantage in view generalization by using novel objects that had distinctive parts and part structures (these stimuli were also presented as real, physical objects). They found that subjects were more accurate under stereo than under nonstereo viewing when the view changed, but that these subjects performed equally well under stereo and nonstereo viewing when the view did not change. However, the authors raised the possibility that the stereo advantage in their experiment may have resulted from a speed–accuracy trade-off combined with a ceiling effect when the initial and test stimuli were shown from the same view. As they suggested, the slower stereo response times (RTs) observed in their study may well have resulted from the shutter apparatus they used: Refocusing from the shutter to the stimuli may have taken longer under stereo viewing. That said, a speed–accuracy trade-off cannot be definitively ruled out.

Recently, Burke (2005) reported that stereo information reduced both RTs and error rates for large viewpoint differences (between 40º and 80º) in a same–different matching task similar to that used in our Experiments 1–3 (although Burke does not infer that 3-D information is encoded in the object representations). His stimuli consisted of stereo photographs of four bent paper clips. A prism stereoscope was used to present these stimuli. It is not clear, however, that the photographed and thin paper clips were clearly seen as 3-D objects under nonstereo viewing. Therefore, it is of interest to see whether or not there is a stereo advantage for stimuli in which monocular cues to depth are clearly available (e.g., shading and motion), as is the case under most everyday viewing. It is also of interest to see whether or not there is a stereo advantage in an identification task that requires long-term object representations.

So far, the existing evidence suggests a stereo advantage in view generalization, but it is not definitive, at least for some important kinds of stimuli and tasks. It is important to address this issue, because the pattern of view generalization in the presence or absence of stereo cues may reveal the degree to which 3-D depth information is encoded in the object representation. Purely image-based theories of object recognition have a strong history (see, e.g., Poggio & Edelman, 1990; Rock & DiVita, 1987). Such accounts have not been definitively ruled out, although the results from the behavioral studies reviewed so far suggest that some depth information is encoded in the object representations used in object recognition tasks (Burke, 2005; Edelman & Bülthoff, 1992; Farah et al., 1994; Humphrey & Khan, 1992). In addition, a computational study by Liu, Knill, and Kersten (1995) complements these behavioral studies, showing that human subjects performed better than an ideal observer model that used strictly 2-D information.

In their study, using forms similar to those used in Edelman and Bülthoff (1992), Liu et al. (1995) compared ideal observer performance to human performance on a form comparison task to infer the information that humans rely on to carry out the task. They defined ideal observers that used strictly 2-D view information (e.g., *x*- and *y*-coordinates of features), strictly 3-D information (e.g., *x*-, *y*-, and *z*-coordinates of features), or intermediate depth information to perform the comparison task. Their results and analyses ruled out what they called a *2-D–2-D* template matching scheme as a model for the performance of their human subjects with symmetric stimuli, even if it is assumed that new templates learned during testing were stored as the experiment proceeded. On the assumption that subjects do not form new stored templates during testing, Liu et al. also ruled out the 2-D–2-D scheme as a model of subject performance with nonsymmetrical stimuli. With these stimuli, however, subjects performed considerably worse than a corresponding 2-D–2-D learning ideal observer. This raised the possibility that the human subjects still operated with the basic 2-D–2-D scheme but learned new 2-D templates as the experiment proceeded (see, e.g., Jolicœur, 1985; Tarr & Pinker, 1989). The net result is that, at least for their nonsymmetrical stimuli, a strictly 2-D image-based scheme remains an open possibility as a model of subject performance. Thus, from a computational perspective, it is also not conclusive whether human observers use 3-D depth information or not.

To summarize, the evidence to date suggests that subjects rely on some 3-D depth information rather than strictly 2-D views for recognizing various kinds of 3-D objects (Burke, 2005; Edelman & Bülthoff, 1992; Farah et al., 1994; Humphrey & Khan, 1992; Liu et al., 1995). The main goal in the present study is to more conclusively determine the extent to which 3-D depth information is encoded in the visual memory of shapes. To that end, following previous studies, we tested for a stereo advantage in view generalization. In contrast with previous studies, we used a range of different stimuli and different tasks. We also provided a range of monocular cues to 3-D depth, including shading, occlusions, and motion (Bülthoff, 1991). In Experiments 1–3, we used a same–different sequential matching task that tapped short-term memory. For these experiments, we used shaded, closed tube-like objects. In Experiment 4, we used an identification task that tapped long-term memory representations. The stimuli in this experiment were wire-frame objects that rotated in depth. Both of these tasks have been used in many previous studies, and they reflect everyday aspects of visual object recognition. If subjects show a stereo advantage for view generalization across these experiments but performance is still view dependent even under stereo viewing, this would provide direct evidence for object representations that were view dependent but contained view-specific depth information (Edelman & Bülthoff, 1992)—at least for the range of stimuli and tasks used in the present experiments.

EXPERIMENT 1

Example stimuli are shown in Figure 1. The stimuli were randomly deformed tori with the tube diameter held close to constant. With these stimuli, it was not possible to do the same–different task by "counting humps" or looking for local distinguishing features. That is, the stimuli were designed to push subjects toward a global encoding of form. The motivation for using these stimuli was to make it more difficult for subjects to generalize over changes in viewpoint. In light of the evidence surveyed above, such an outcome would increase the likelihood of yielding a stereo advantage in view generalization, since subjects could not consistently do the task by comparing abstract descriptions (e.g., numbers of humps) or local 2-D image features. However, even in nonstereo viewing, there was substantial depth and 3-D structure information available in the form of interposition, shading, and attenuation of illumination with (simulated) distance. In this experiment, subjects wore an eye patch over one eye for the nonstereo viewing condition.

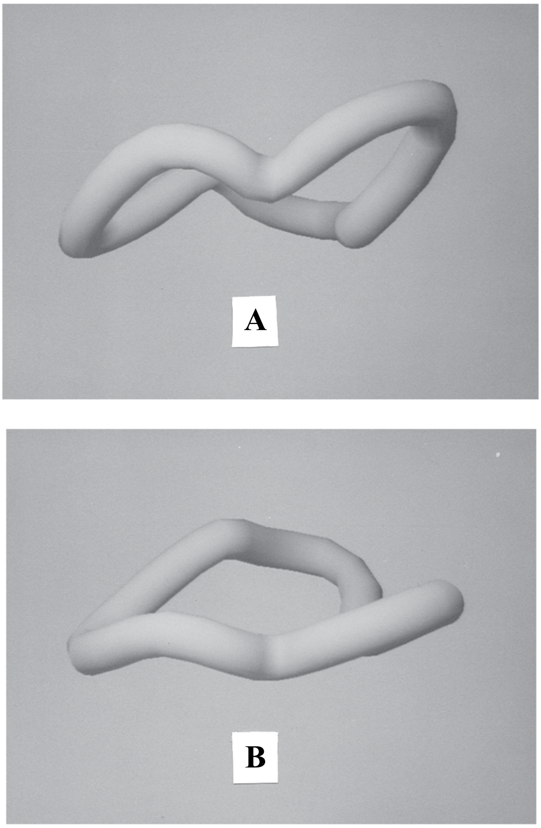

Figure 1. Example stimuli used in a same trial in Experiment 1. (A) Initial presentation. (B) Second presentation, consisting of the form in the initial presentation rotated around a horizontal axis by 38º (forms were shown in simulated stereo for the stereo condition).

Method

Subjects. Nineteen subjects completed the experiment, but 1 did not meet a preset criterion of 1/3 correct responses in both conditions. Most of the subjects were Brown University undergraduates. All of the subjects gave informed consent and were paid for their participation.

Apparatus. The displays were generated on a Silicon Graphics Onyx2 and displayed on a 19-in. monitor with a resolution of 1,280 × 1,024 pixels. The vertical resolution was halved so that the left and right frames could be interleaved. The screen edges were masked off using black poster board. An adjustable riser, with its top close to the subjects and positioned just below eye-level, masked any remaining reflections from the screen light. In this configuration, the riser itself perceptually disappeared. The subject's arms and the computer mouse were hidden from view beneath a table-like construction that supported the riser. The subject's head was stabilized using a chinrest and a padded headrest, although some slight side-to-side head movement was possible.

Stimuli. The stimuli consisted of deformed tori defined by an interpolated circle consisting of seven rings of points. For each stimulus, the simulated large diameter of the original torus was 11.7 cm, and the simulated small diameter was 1.8 cm (simulated magnitudes refer to the measures ideally determined by vergence angle and angular extent). The rings of points were then deformed up–down and in–out within ranges of 6.5 cm, with the constraint that the deformations in these two directions differed by at least 2.25 cm from the deformations of the immediately adjacent rings. *Imagine* rays were drawn from the center of the undeformed torus to the centers of the seven rings of points. The ray to the first circle of points was initially aligned with the *x*-axis, around which the stimuli rotated. The angle of this ray relative to this axis was randomly varied by approximately 51.4º so that the first ring of points, which defined the undeformed torus, would not always be aligned with the axis around which the stimuli rotated. Before the final surface was interpolated, seven additional intervening rings were inserted and positioned to keep the tube diameter approximately constant.

The distance to the screen was 90 cm, and the simulated distance to the center of the undeformed tori was 115 cm. The average horizontal visual angle of the stimuli was about 12º, whereas the average vertical visual angle was about 7º. Although the simulated distance of the deformed tori placed these tori just behind the computer screen, because of the care taken to perceptually isolate the stimuli, the impression was of shapes floating in space, with no impression of a screen surface. Watt, Akeley, Ernst, and Banks (2005) presented evidence that, under stereo viewing and with care taken to perceptually isolate the stimuli, screen cues played no role in a slant perception task (see also Bennett, in press).

Stereo viewing was simulated using Stereographics liquid crystal goggles. Asymmetric viewing frustums (viewing pyramids defined by eye position and screen dimensions) were defined for each eye, with the dimensions adjusted depending on interpupilary distance, which was measured for each subject by sighting over a clear plastic ruler placed on the bridge of the nose. For the nonstereo condition, an eye patch was placed over the nonsighting eye, as indicated (roughly) by handedness. The eye patch was placed under the stereo goggles, and eye width was set to zero. The program that generated the stimuli and collected responses was written in C and Silicon Graphics' GL graphics programming language.

On each trial, the deformed tori could be rotated about the *x*-axis by equal amounts either "front up" or "front down," as is shown in Figure 2. This rotation ensured that the amount of self-occlusion was, on average, the same across trials. The stimuli were blue and were displayed against a gray background. As has already been noted, self-occlusion, shading, and attenuation of illumination with (simulated) distance provided monocular cues to depth and 3-D structure.

Design and Procedure. Viewing (stereo vs. nonstereo) was a within-subjects factor blocked by session. Half of the subjects ran in the stereo condition first, whereas the other half ran in the nonstereo session first.

The task was a same–different sequential matching task. Each trial began with the presentation of the first stimulus for 4,000 msec, followed by the presentation of a blank gray field for 1,750 msec. The second stimulus was presented and left in view until the subjects responded. After that, a pattern mask was presented that consisted of a grid of squares of varying lightness (with sides of 4.45º). The subjects were instructed to respond as quickly as possible while maintaining accuracy. Feedback about accuracy was given throughout the experiment as well as during the practice trials. The subjects were also given the screen-presented message "too slow" when their RTs exceeded 4,000 msec. Responses were made by pressing the right mouse button if the two presented stimuli were the same deformed torus and by pressing the left button if they were different deformed tori. The center mouse button was used to begin each trial.

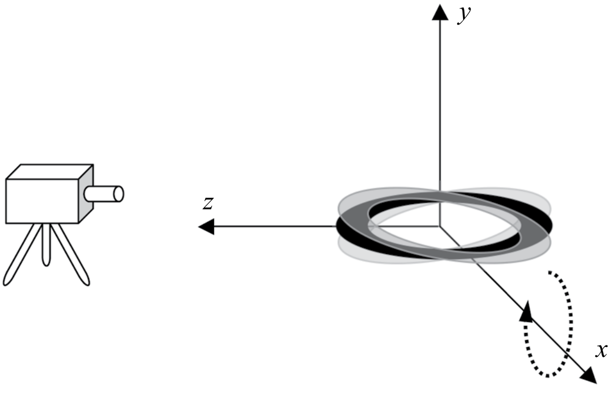

Figure 2. An illustration of the rotation of the tube-like stimulus used in Experiments 1–3. In comparison with no rotation (black ellipse), the object could be rotated about the *x*-axis either front up (dark gray ellipse) or front down (light gray ellipse) by 19º in Experiments 1 and 2, and by 15º in Experiment 3, so that the total differences in orientation were 38º and 30º, respectively, on the different-orientation trials.

In each session there were 144 trials overall, broken into three blocks of 48 trials. Half were *same* trials, in which the first and second stimuli were the same deformed torus, and the remaining trials were *different* trials, in which the two stimuli presented were two different deformed tori. On two thirds of the *same* trials (48/72 trials), the second stimulus presented was rotated about the *x*-axis relative to the orientation of the first stimulus (different-orientation condition). On the remaining third of *same* trials (24/72 trials), the second stimulus was shown at the same orientation as the first (same-orientation condition). The subjects were informed of these percentages. The percentages of different-orientation and same-orientation trials were the same as those of the *different* trials.

Half of the trials began with the first stimulus rotated front up, whereas the other half began with the first stimulus rotated front down (see Figure 2). In Experiment 1, rotations were always 19º up or down so that the subjects were required to generalize over total rotations of 38º on different-orientation trials. The subjects were informed that the stimuli were only rotated about the *x*-axis.

The experimental trials were preceded by 36 practice trials. Immediately before the practice trials, the subjects were shown the various trial conditions. On these example trials, two stimuli were shown in succession and then presented side by side. The subjects were shown how these two stimuli could or could not be rotated to coincide with the same torus. For example, for a *same* different-orientation trial, one stimulus was rotated to correspond to the other. For a *different* different-orientation trial, the subjects were shown that it was not possible to rotate either stimulus to coincide with the other.

Each of the two sessions (one stereo and one nonstereo) took approximately 35–40 min. There were at least 2 days and no more than 2 weeks between sessions.

Results

The results of Experiment 1 are shown in Figures 3 and 4. A viewing (stereo vs. nonstereo) × orientation (same vs. different) ANOVA with percentage correct as the dependent variable yielded main effects of viewing [$F(1,17) = 36.69, p < .001$] and of orientation [$F(1,17) = 134.70, p < .001$]. Importantly, the viewing × orientation interaction was significant [$F(1,17) = 16.31, p = .001$], reflecting the fact that subjects generalized better under stereo viewing (see Figure 3).

For the *same* same-orientation trials, the subjects were close to ceiling for both stereo conditions (stereo, $M = 97.0\%$, $SE = 1.1\%$; nonstereo, $M = 96.7\%$, $SE = 1.5\%$). By comparison, the subjects were more accurate under stereo than under nonstereo viewing for *same* different-orientation trials (stereo, $M = 70.6\%$, $SE = 3.4\%$; nonstereo, $M = 57.3\%$, $SE = 3.4\%$). A post hoc test showed that this difference was significant [$t(17) = 5.15, p < .001$]. Furthermore, although the subjects clearly found the task demanding, performance under nonstereo viewing (*same* different-orientation trials) was greater than chance [$t(17) = 2.16, p < .025$].

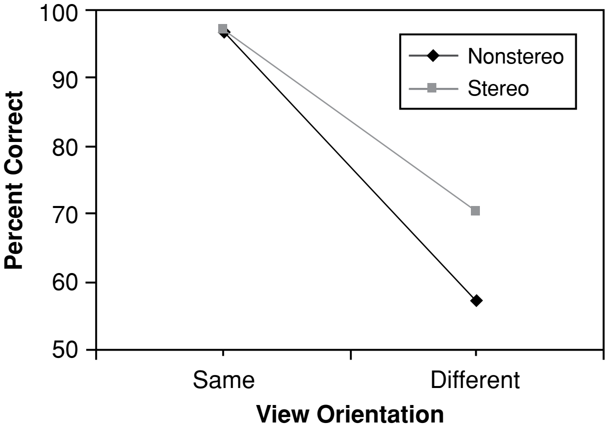

Figure 3. Mean percent correct responses for Experiment 1.

The pattern for RTs was very different from that of the accuracy data (see Figure 4). There was a large effect of orientation [$F(1,17) = 51.00, p < .001$], but there was no effect of viewing [$F(1,17) < 1$]. Furthermore, there was no viewing × orientation interaction [$F(1,17) < 1$]. Indeed, Figure 4 shows that RTs were virtually identical for the two viewing conditions. The key observation is that there is no evidence that a speed–accuracy trade-off accounted for the stereo advantage in generalizing to new viewpoints (cf. Humphrey & Khan, 1992).

Discussion

Consistent with the findings of several previous studies, performance was view dependent under both stereo and nonstereo viewing (see, e.g., Edelman & Bülthoff, 1992; Farah et al., 1994; Humphrey & Khan, 1992; Sinha & Poggio, 1994). Importantly, there was a clear stereo advantage in view generalization. That is, the observers' responses were equally accurate under both stereo and nonstereo viewing when the stimuli were presented at the same orientation, but their accuracy was much poorer in the nonstereo than in the stereo condition when the stimuli were presented at different orientations with only a 38º difference in viewpoint. This stereo advantage is evidence against a purely 2-D image-based model of subject performance. However, the marked view dependency of performance—even under stereo viewing—suggests that the subjects did not build up and use full 3-D models, even though information about 3-D structure was available in both viewing conditions (e.g., shading). That said, however, there were also self-occlusions, and it is possible that they inhibited the subjects from building such models.

The subjects were not tested for stereo anomaly in any of the experiments. However, we do not believe that any stereo defects would have affected the main results. Subjects who are stereoanomalous with brief stimulus presentations are often normal or show reduced stereoanomaly with long stimulus presentations as well as with repeated exposure (Newhouse & Uttal, 1982; Patterson & Fox, 1984). In the present study, we used relatively long stimulus presentation times, and the subjects were repeatedly exposed to similar, stereoscopically presented stimuli. More importantly, any stereo defects would have worked against our hypothesis, since stereoanomalous subjects would not have been expected to show a stereo advantage. That having been said, testing stereoanomalous subjects would be interesting for an examination of whether or not 3-D information from monocular cues (e.g., shading, texture, motion) may facilitate view generalization.

There are other possible explanations for the advantage under stereo viewing in Experiment 1 that would stem from using an eye patch to form the nonstereo condition. First, the stereo advantage may be due to the fact that separate estimates (associated with the two images, one for each eye) are available for the same 2-D features under stereo viewing. This seems unlikely, however, given the small differences in the left and right eye images under stereo viewing. Another potential problem with using an eye patch to produce the nonstereo condition is that subjects may have found viewing with one eye unnatural and unfamiliar, and this could have somehow inhibited performance in the nonstereo condition. To address these concerns, the nonstereo condition in Experiment 2 was formed by presenting the same screen image to each eye, meaning that the projected retinal images were essentially identical, given the viewing distance of 90 cm.

EXPERIMENT 2

Experiment 2 was the same as Experiment 1, except that the nonstereo condition was formed by presenting the same image to each eye.

Method

Subjects. Twenty-seven subjects completed the experiment, but 1 did not meet a preset criterion of scoring greater than 1/3 correct in both conditions. Most of the subjects were undergraduates at Brown University. All of the subjects gave informed consent and were paid for their participation.

Stimuli, Design, and Procedure. The stimuli, design, and procedure were the same as in Experiment 1, except that for the nonstereo condition each eye was presented with the same screen image.

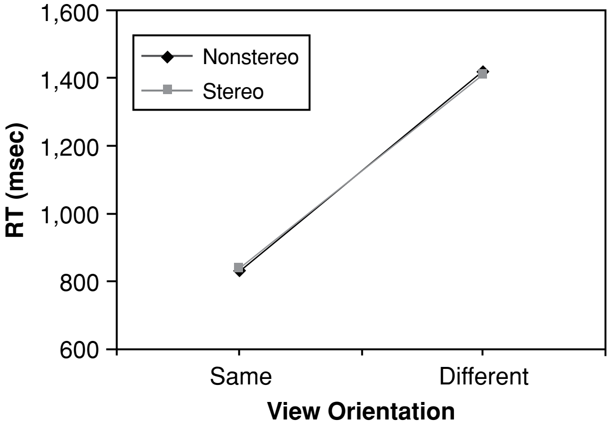

Figure 4. Mean correct response times (RTs) for Experiment 1.

Results

The results are shown in Figures 5 and 6. A viewing (stereo vs. nonstereo) × orientation (same vs. different) ANOVA with percentage correct as the dependent variable yielded main effects of viewing [$F(1,25) = 8.20$, $p < .01$] and orientation [$F(1,25) = 204.54$, $p < .001$]. Furthermore, the viewing × orientation interaction was significant [$F(1,25) = 4.40$, $p < .05$], reflecting the fact that the subjects generalized better under stereo viewing (see Figure 5).

Because the stereo advantage in view generalization seemed to be reduced in Experiment 2 in comparison with Experiment 1 (see Figures 3 and 5), an experiment (nonstereo method) × viewing × orientation ANOVA was conducted with experiment as a between-subjects factor, viewing and orientation as within-subjects factors, and percentage correct as the dependent variable. The three-way interaction, however, was not significant [$F(1,42) = 1.54$, $p = .222$].

As in Experiment 1, for the *same* same-orientation trials, the subjects were equally at ceiling in the stereo and nonstereo conditions (stereo, $M = 97.4\%$, $SE = 0.8\%$; nonstereo, $M = 97.0\%$, $SE = 1.7\%$). In contrast, for the *same* different-orientation trials, the subjects were more accurate under stereo viewing (stereo, $M = 64.3\%$, $SE = 2.7\%$; nonstereo, $M = 56.8\%$, $SE = 3.3\%$). This difference was again significant [$t(25) = 2.72$, $p = .006$]. Finally, performance under nonstereo viewing (*same* different-orientation trials) was greater than chance [$t(25) = 2.06$, $p = .025$].

For RTs (see Figure 6), there were effects of viewing [$F(1,25) = 4.73$, $p = .039$] and orientation [$F(1,25) = 48.89$, $p < .001$]. However, unlike for the accuracy data, there was no viewing × orientation interaction [$F(1,25) = 1.67$, $p = .208$]. It is not clear why there was a significant effect of the viewing condition on RTs (unlike in Experiment 1—or Experiment 3, below). The important observation is that the subjects were faster under stereo viewing; therefore, the stereo advantage in generalizing to new views that was observed in accuracy was not due to a speed–accuracy trade-off coupled with a ceiling effect when the orientation was the same.

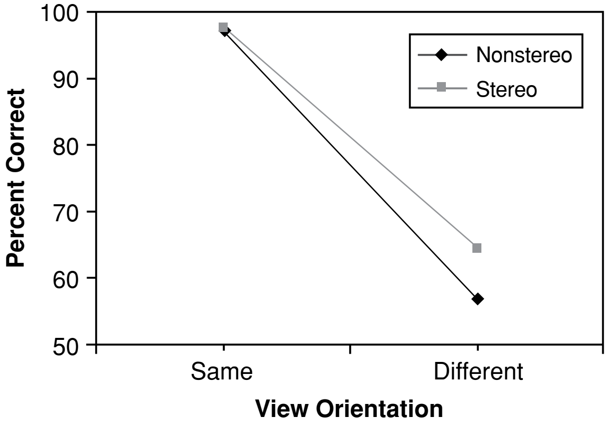

Figure 5. Mean percent correct responses for Experiment 2.

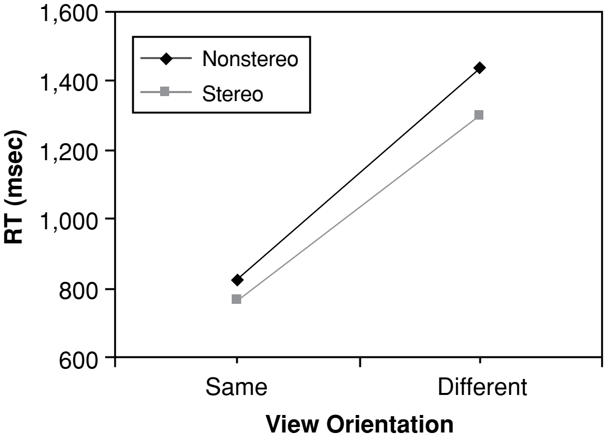

Figure 6. Mean correct response times (RTs) for Experiment 2.

Discussion

In Experiment 2, we replicated the results of Experiment 1 using a different method to form the nonstereo viewing condition. First, we found that recognition performance was markedly view dependent in both viewing conditions. Second, we found a stereo advantage in view generalization when the same image was shown to each eye under nonstereo viewing. Thus, the reported stereo advantage cannot simply be due to the fact that the subjects had access to redundant information about 2-D features under stereo viewing, or to the fact that they had to view the stimuli unnaturally while wearing an eye patch under nonstereo viewing. Taken together, the results of Experiments 1 and 2 provide evidence that depth information is encoded in a view-dependent manner (see, e.g., Edelman & Bülthoff, 1992; Farah et al., 1994; Humphrey & Khan, 1992). These results, therefore, provide evidence against strictly 2-D image-based models or strictly 3-D structural models (see also Liu et al., 1995).

EXPERIMENT 3

Part of the aim in designing the stimuli and choosing the rotation magnitudes for Experiments 1 and 2 was to arrive at a generalization task that was challenging. In fact, performance in the *same* different-orientation trials in the first two experiments was slightly but significantly above chance. Discovering whether or not the stereo advantage in view generalization holds in a task that is, or is expected to be, easier is of interest. Therefore, we attempted to replicate the stereo advantage with smaller depth rotations. This would have presumably made view generalization easier. Thus, Experiment 3 was the same as Experiment 2, except that front-up and front-down rotations of 15º were used, so that subjects were required to generalize over rotations of 30º rather than 38º on different-orientation trials.

Method

Subjects. Twenty-six subjects participated in the experiment. Most of the subjects were undergraduates at Brown University. All of the subjects gave informed consent and were paid for their participation.

Stimuli, Design, and Procedure. The stimuli, design, and procedure were the same as in Experiment 2 (i.e., for the nonstereo condition, each eye was presented with the same screen image), except for the difference in depth rotations.

Results

The results are shown in Figures 7 and 8. A viewing (stereo vs. nonstereo) × orientation (same vs. different) ANOVA with percentage correct as the dependent variable yielded main effects of viewing [$F(1,25) = 12.15$, $p = .002$] and orientation [$F(1,25) = 352.51$, $p < .001$]. Critically, the viewing × orientation interaction was significant [$F(1,25) = 13.27$, $p = .001$], reflecting the fact that the subjects generalized better under stereo viewing (see Figure 7).

For the *same* same-orientation trials, the subjects were equally at ceiling in both stereo conditions (stereo, $M = 97.6\%$, $SE = 0.9\%$; nonstereo, $M = 96.3\%$, $SE = 1.2\%$). As in Experiments 1 and 2, the subjects were more accurate under stereo viewing for *same* different-orientation trials (stereo, $M = 68.4\%$, $SE = 2.3\%$; nonstereo, $M = 57.9\%$, $SE = 2.7\%$), and this difference was significant [$t(25) = 3.99$, $p < .001$]. Finally, performance under nonstereo viewing for *same* different-orientation trials remained greater than chance [$t(25) = 2.91$, $p < .005$].

As in the previous experiments, the pattern of results for RTs was very different from that for the accuracy data (Figure 8). There was a large effect of orientation [$F(1,25) = 121.00$, $p < .001$] but no effect of viewing [$F(1,25) = 1.62$, $p = .215$]. Furthermore, there was no viewing × orientation interaction [$F(1,25) < 1$]. Again, the key observation is that there is no evidence of a speed–accuracy trade-off.

Discussion

We replicated the two main results of Experiments 1 and 2 using smaller rotations in depth about the *x*-axis. First, performance was view dependent. Second, there was a stereo advantage in view generalization. There was one difference in the data that likely resulted from the smaller depth rotations: Performance under nonstereo viewing on *same* different-orientation trials was clearly greater than 50%, although the task was still challenging under nonstereo viewing. Thus, the stereo advantage in generalizing to new viewpoints was still present when the subjects clearly performed better than chance in the nonstereo condition.

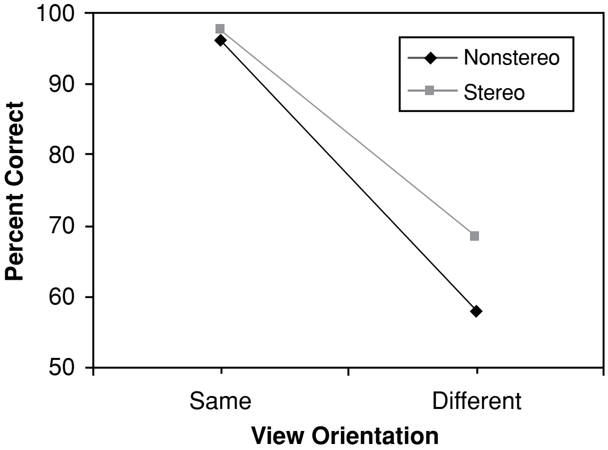

Figure 7. Mean percent correct responses for Experiment 3.

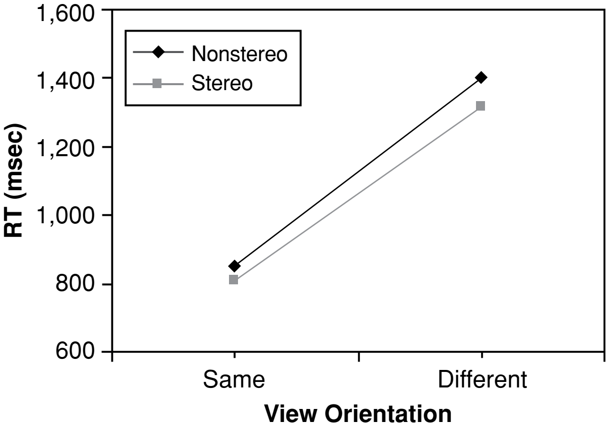

Figure 8. Mean correct response times (RTs) for Experiment 3.

EXPERIMENT 4

Despite the consistency of the results across Experiments 1–3, there are several reasons to question their generality. First, the subjects were presented with only one view of the stimuli before they were tested with new orientations. When multiple 2-D views are available, however, it is possible to relate the views to each other and to the input image so that the subject effectively applies a model of the 3-D structure of the objects to be identified (Liu et al., 1995; Ullman, 1979). It is possible that the stereo advantage will disappear if observers have the opportunity to form multiple views or to derive 3-D structure from multiple views (Ullman, 1979). Second, the stimuli were randomly generated on each trial, limiting the possibility that subjects could learn the 3-D structure of a particular deformed torus over time. In fact, all the previous studies in which the role of stereo information in recognition has been examined have used a small set of target objects seen from many different views (Burke, 2005; Edelman & Bülthoff, 1992; Farah et al., 1994; Humphrey & Khan, 1992; Liu et al., 1995; Sinha & Poggio, 1994). Third, it is possible that the view dependency was due to self-occlusions that may have obscured local features that subjects used to perform the task. It would be of interest to see if the same pattern holds with stimuli that have minimal self-occlusion. Finally, it would be of interest to see whether or not the general pattern of results observed in Experiments 1–3 is also present on a task that taps longer term memory representations of objects.

We addressed these concerns in Experiment 4. In this experiment, we trained subjects to recognize and identify a small number of rotating wire-frame objects. During training, these objects were presented from a particular view. After training, we tested subjects' ability to recognize the learned targets presented from familiar and unfamiliar views. As in Experiments 1–3, training and testing were conducted under either stereo or nonstereo viewing conditions. In addition, we varied the difficulty of recognizing the objects. Using a learning and identification task, this experiment tested whether the view dependency and stereo advantage in view generalization are also present in a task that taps longer term visual representations instead of the short-term memory representations tapped by sequential matching.

Method

Subjects. Twenty-two undergraduate students from Brown University were recruited as subjects. They provided informed consent and were paid for their time.

Stimuli. The stimuli used in Experiment 4 consisted of novel wire-frame objects, examples of which are shown in Figure 9 (see also Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992; Liu et al., 1995). These objects rotated rigidly in depth, and they were viewed in either stereo or nonstereo conditions.

Each wire-frame object consisted of either five or nine white dots arranged in a virtual 3-D space and connected by thin white lines. The *x*-, *y*-, and *z*-coordinates of the dots were constrained to be within a volume of one arbitrary cubic unit. The dots were 3 × 3 pixels in size, and the connecting lines were 1 pixel thick. Both the dots and the lines were antialiased using standard OpenGL 1.2 routines. Thus, there was very little self-occlusion and no shading cues to 3-D shape. The wire frame objects were orthographically projected onto the screen. They subtended between 3.8º and 4.8º of visual angle and were presented centered against a black background. Four five-vertex ("easy") and four nine-vertex ("hard") target wire-frame objects were pregenerated. Distractors were derived from targets by randomly permuting the *z*-coordinates of the target vertices and reconnecting them. During the test phase, randomly generated distractors were presented on each trial.

In the stereo condition, slightly different orthographic projections of the stimulus were presented to each eye. The disparity of each vertex was computed assuming a viewing distance of 60 cm and an interocular eye distance of 6 cm (an average interocular distance). Stereo shutter glasses (NuVision 60GX) were used to present the two different images to each eye. In the nonstereo condition, the same images were presented on the screen and no shutter glasses were used.

Design. There were four between-subjects conditions in Experiment 4. The subjects learned either easy five-vertex or hard nine-vertex targets and viewed the stimuli either with or without stereo information. Five subjects were randomly assigned to easy targets in either the stereo or nonstereo conditions, and 6 were randomly assigned to the hard targets in each of these stereo conditions.

Each subject completed two learning phases and one test phase. During both learning phases, only the target wire-frame objects were presented. Each target was rotated by 0º, 60º, 120º, or 180º about the *z*-axis relative to an arbitrary reference orientation, as is shown in Figure 10. This initial rotation in the image plane defined the *learned orientation* for that target. The assignment of the learned orientation to the target was the same for all the subjects. Following the initial rotation about the *z*-axis, all targets were further rotated in increments of 3.6º per frame to produce a smooth rotation in depth. For the easy targets, this incremental rotation was about the vertical *y*-axis. This produced a simple rotation in depth. By comparison, the hard targets were arbitrarily rotated about both the *x*- and *y*-axis to produce a complex tumbling motion (see, e.g., Stone, 1999). The depth rotation produced a sequence of 100 frames that could be played in ascending order for clockwise rotations in depth, or in descending order for counterclockwise rotations. Two targets always rotated in depth clockwise, and the remaining two targets always rotated in depth counterclockwise.

During the testing phase, both targets and distractors were presented. Prior to any rotations in depth, the targets were rotated about the *z*-axis either by their learned orientation established during

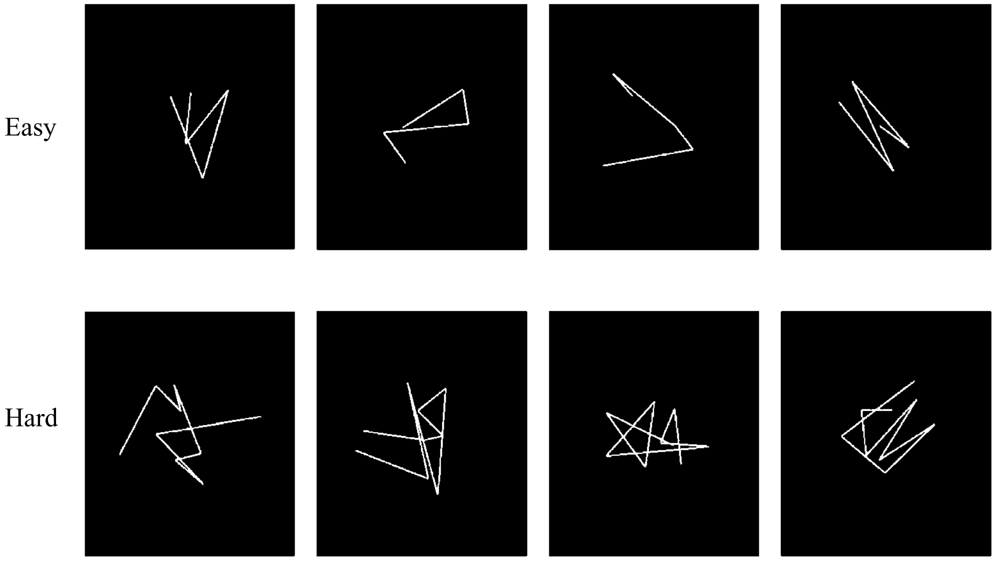

Figure 9. Single frames from the four "easy" and "hard" targets used in Experiment 4. The line widths have been exaggerated in the figure to make them more clearly visible. The actual stimuli consisted of small circles at the bends with 1-pixel-wide lines connecting the circles. The easy targets had five bends, and the hard targets had nine bends.

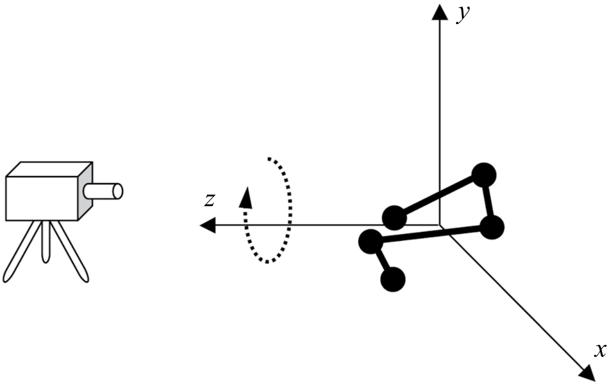

Figure 10. An illustration of the virtual space used in Experiment 4. Each wire-frame object was initially placed into this virtual space in an arbitrary pose. The trained viewpoint of each object is then defined with respect to the amount of rotation about the viewing *z*-axis. That is, each target was rotated by a fixed amount about the *z*-axis (dotted arrow). For easy objects, targets were subsequently rotated about the *y*-axis to produce the final image sequence. For hard objects, targets were rotated about both the *x*- and *y*-axes.

learning (same-orientation condition) or by a different orientation (different-orientation condition). The different orientation was randomly sampled between 0º and 360º, excluding the learned orientation associated with that target. Finally, the targets were rotated in depth both clockwise and counterclockwise during the test phase. Thus, on half the trials the targets rotated in the same direction as in the learning phase, and on the remaining trials they rotated in a different direction. To summarize, object type (easy vs. hard) and viewing (stereo vs. nonstereo) were between-subjects factors, and orientation (same vs. different) and rotation direction (same vs. different) were within-subjects factors.

Procedure. As was mentioned above, Experiment 4 consisted of two learning phases followed by a test phase (see Vuong & Tarr, 2006). In the first learning phase, the subjects were shown four target objects. For easy objects, each target was presented for 10 frames (approximately 700 msec). For hard objects, each target was presented for a full 360º rotation (100 frames, approximately 7,000 msec). The starting frame was randomly selected for each object on each trial. The subjects were instructed to press the appropriate key for each object after the object disappeared from the screen. They were informed that they could not respond until this event occurred, and they were encouraged to pay attention to the stimulus for the entire presentation. If the subjects 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, 1,000-Hz tone. For this phase, the subjects were instructed to respond as accurately as possible. Each object was presented 40 times, for a total of 160 trials. There was a short self-timed break after every 40 trials.

In the second learning phase, the subjects were informed that they could respond at any time after the stimulus was presented, and they were further instructed to respond as quickly and as accurately as possible. The only feedback they received in this phase was the low tone that signaled an incorrect response. For the easy stimuli, each target was presented 80 times, for a total of 320 trials, and for the hard stimuli each target was presented 40 times, for a total of 160 trials. Again, there were self-timed breaks after every 40 trials.

Finally, in the test phase, the subjects were presented with targets intermixed with randomly generated distractors. Recall that distractors were derived from one of the four targets by randomly permuting the *z*-coordinate of the vertices on a trial-by-trial basis. The subjects were instructed to press the space bar for any distractors and to continue to respond with the learned letter key associated with each target. As in the learning phases, targets were presented in the same orientation or in a different orientation (via a rotation of the object about the *z*-axis) and were rotated in the same direction or in a different direction in depth. Each target was presented 10 times, for a total of 160 trials (4 targets × 2 orientations × 2 rotation directions). There were also 160 distractor trials. Thus, there were 320 trials in this phase. The subjects were instructed to respond as quickly and as accurately as possible. No feedback was provided during this phase. As before, there was a short break after every 40 trials. The entire experiment took approximately 45 min to 1 h.

The stereo condition was conducted on a PC computer with a screen resolution of 1,024 × 768 pixels and a refresh rate of 120 Hz (60 Hz for each eye). The nonstereo condition was conducted on a different PC computer with the same screen resolution and a 60-Hz refresh rate.[1] Both conditions were conducted in a dimly lit room. Self-written code using the C language was used to control stimulus presentation and to collect responses from the keyboard. The subjects sat approximately 60 cm from the monitor. The four keys used were "v," "b," "n," and "m," which were randomly assigned to the four targets for each subject. All the subjects were instructed to respond with their dominant hand.

Results

An initial set of analyses did not reveal any effects of rotation direction on either RTs or accuracy (but see Stone, 1999; Vuong & Tarr, 2006). We believe that this may be due partially to the lack of self-occlusion, but future studies in this direction are needed. For the goal of the present study, we focused on the effects of the stereo condition on recognition performance (as in Experiments 1–3). Thus, for the data reported below, we collapse RTs and accuracy across rotation direction.

For all the analyses reported, we eliminated correct RTs longer than 5,000 msec or shorter than 300 msec to control for outliers and anticipatory responses.

Learning phase. We first analyzed the RT and accuracy data from the second learning phase, during which the subjects made speeded responses. Both dependent variables were submitted to an ANOVA with object type (easy vs. hard) and viewing (stereo vs. nonstereo) as between-subjects factors. For accuracy, there were no significant main effects or interactions (all *p*s $> .2$). Importantly, there were no differences in learning for subjects in the two stereo conditions (for easy objects, $M_{\text{stereo}} = 93.4\%$, $SE = 1.0\%$, $M_{\text{nonstereo}} = 92.4\%$, $SE = 1.4\%$; for hard objects, $M_{\text{stereo}} = 95.1\%$, $SE = 1.9\%$, $M_{\text{nonstereo}} = 95.6\%$, $SE = 1.4\%$). For RTs, there was only a significant effect of object type [$F(1,18) = 29.47$, $p < .001$]. Not surprisingly, the subjects responded more quickly to easy targets ($M = 800$ msec, $SE = 25$) than to hard targets ($M = 1{,}388$ msec, $SE = 96$).

Test phase. On the basis of the analysis of the learning phase data, we believed that the subjects in both stereo conditions learned the targets quite well. All the observers were able to identify the targets very accurately (>90%; chance level was 20%). The key question was, how well did subjects generalize to nonstudied orientations in the presence and absence of stereo information? To that end, we submitted the data to a mixed-design ANOVA with object type and viewing as between-subjects factors, and orientation (same vs. different) as a within-subjects factor.

For the accuracy data, the critical finding is the significant interaction between viewing and orientation [$F(1,18) = 14.99$, $p < .001$]. This interaction is shown in Figure 11 separately for easy and hard targets. Averaged across both easy and hard objects, the subjects were better able to generalize to different orientations in the stereo condition ($M_{\text{different view}} = 72.0\%$, $SE = 3.1\%$) than in the nonstereo condition ($M_{\text{different view}} = 56.5\%$, $SE = 3.2\%$). By comparison, they were equally accurate at identifying both types of targets in the same orientation, irrespective of the viewing condition (stereo, $M_{\text{same view}} = 89.1\%$, $SE = 2.4\%$; nonstereo, $M_{\text{same view}} = 91.4\%$, $SE = 1.9\%$). Overall, there were also main effects of viewing [$F(1,18) = 6.18$, $p = .023$] and orientation [$F(1,18) = 233.64$, $p < .001$].

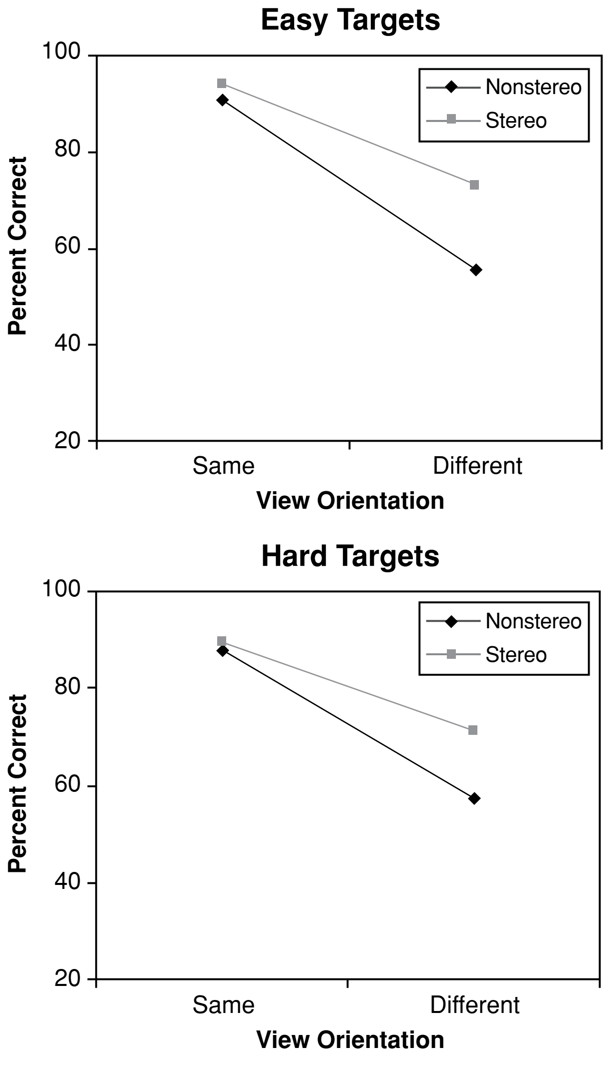

Figure 11. Mean percent correct responses during the test phase of Experiment 4. The data are plotted separately for easy and hard objects. The view orientation is relative to that during the learning phases.

The mean RTs for easy and hard objects are presented in Figure 12. For RTs, a slightly different pattern emerged. In contrast to the accuracy data, there was no significant interaction between viewing and orientation ($F < 1$). There was also no significant effect of viewing [$F(1,18) = 1.51$, $p = .235$]. The effect of orientation [$F(1,18) = 185.55$, $p < .001$] was significant, as were the interaction between object type and orientation [$F(1,18) = 52.58$, $p < .001$] and the three-way interaction between object type, viewing, and orientation [$F(1,18) = 4.99$, $p = .038$]. Finally, the subjects responded more quickly to easy than to hard targets [$F(1,18) = 60$, $p < .001$].

Discussion

All the subjects were able to learn both easy and hard targets relatively quickly during the learning phases. Importantly, they were able to maintain this high level of identification during testing when targets were presented in their learned orientations. However, when targets were presented at nonstudied orientations during the test phase, there was a clear benefit in accuracy under stereo viewing. Moreover, there was no evidence that this stereo advantage found in generalizing to new views was the result of a speed–accuracy trade-off.

There are several points to note about the target objects, which contrast with the stimuli used in Experiments 1–3. First, there was little self-occlusion, so local features (e.g., the vertices of the wire frame) were almost always visible. Second, rigid depth rotation provided a strong basis for the estimation of 3-D shape, even under nonstereo viewing (Ullman, 1979). Third, over the course of training, the subjects could form multiple-view representations that allowed for perfect view generalization (Tarr & Pinker, 1989; Ullman & Basri, 1991).

In principle, the subjects could have derived strictly 2-D view representations or strictly 3-D representations to perform the recognition task, which may have eliminated any stereo advantage in view generalization. For the same reason, one might expect view-invariant performance with these stimuli during the testing phase, at least under stereo viewing. Like the subjects in Experiments 1–3, the subjects in Experiment 4 seemed to use stereo information to build up representations that were more invariant to viewpoint changes, presumably because they incorporated additional or more accurate depth and shape information. However, there was still a robust view dependency under stereo viewing, indicating that the subjects did not make use of available information specifying 3-D structure to build up full, view-independent 3-D models of the stimuli.

GENERAL DISCUSSION

There were two striking and robust findings across the four experiments reported. First, there was a stereo advantage in generalizing to unfamiliar views of novel objects (see, e.g., Farah et al., 1994; Humphrey & Khan, 1992). Second, performance was view dependent (Edelman &

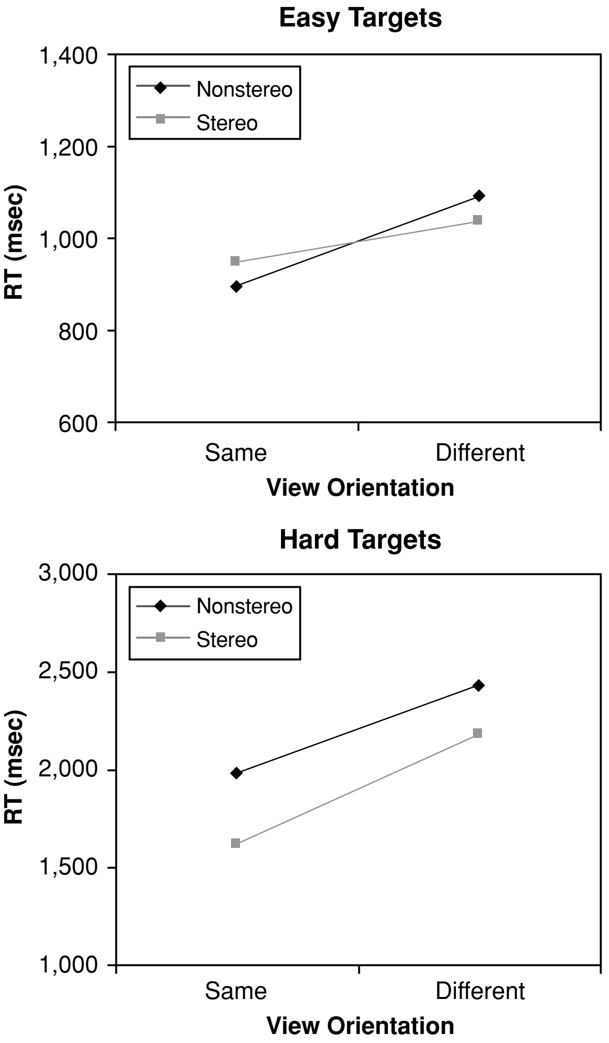

Figure 12. Mean correct response times (RTs) during the test phase of Experiment 4. The data are plotted separately for easy and hard objects. The view orientation is relative to that during the learning phases.

Bülthoff, 1992; Farah et al., 1994; Humphrey & Khan, 1992; Sinha & Poggio, 1994). We obtained these results in a sequential matching task with shaded, closed, and randomly deformed tori (Experiments 1–3) and in a learning and identification task with rotating wire-frame objects (Experiment 4). The stereo advantage in view generalization, coupled with an overall view dependency in experiments in which different tasks and stimuli were used, argue against models of subject performance that posit 3-D structural representations (e.g., Marr & Nishihara, 1978) or models that posit 2-D image-based representations (e.g., the simplified model proposed by Poggio & Edelman, 1990).

Previous studies found evidence suggestive of a stereo advantage in view generalization (Farah et al., 1994; Humphrey & Khan, 1992). However, these earlier studies were not conclusive. For instance, Farah et al. did not directly compare stereo and nonstereo viewing under controlled conditions. Similarly, Humphrey and Khan's stereo advantage may be due in part to a speed–accuracy trade-off, as the authors also noted. In the present study, we directly compared stereo and nonstereo viewing under controlled conditions and across a range of tasks and stimuli. Here, we provide clear evidence that, across a range of stimuli (i.e., shaded, deformed tori and rigidly rotating wire-frame objects) and recognition tasks (i.e., matching and identification), stereo information facilitates view generalization (see also Burke, 2005).

Similar to the results of Experiments 1–4, several previous studies have also consistently found view dependency in recognition performance under both nonstereo viewing (see, e.g., Rock & DiVita, 1987; Tarr, 1995; Tarr, Williams, Hayward, & Gauthier, 1998) and stereo viewing (see, e.g., Edelman & Bülthoff, 1992; Sinha & Poggio, 1994). The stereo advantage that we observed in our study, in combination with the exhibited view dependency, permits several conclusions to be drawn regarding the nature of the underlying object representation. First, the stereo advantage suggests that 3-D depth information is encoded in the object representation because the stereo advantage was found mostly in the accuracy data with no evidence of any speed–accuracy trade-offs. Second, our study as a whole suggests that stereo information enhances the 3-D depth percept or representation to which monocular cues also contribute. This suggestion is made on the basis of our finding a stereo advantage when other cues to 3-D depth were available, including motion information, which is also a strong depth cue (Ullman, 1979).

Third, and finally, the results of all our experiments, but particularly Experiment 4, suggest that subjects did not learn full 3-D representations, even if this had been possible during the learning phase (Liu et al., 1995; Sinha & Poggio, 1994). In combination with previous studies (Burke, 2005; Edelman & Bülthoff, 1992; Farah et al., 1994; Humphrey & Khan, 1992), the present results suggest that—at least for a range of tasks and stimuli—the human visual system encodes 3-D shape information and that this additional information is specified from a particular view.

These conclusions must, however, be tempered for at least two reasons. First, the stimuli used in Experiments 1–4 did not have a clearly defined part structure that could facilitate viewpoint generalization (Biederman, 1987). Thus, it is possible that there would have been no stereo advantage in generalization with objects constructed out of geons or geon-like parts because these parts are specified by image information (Biederman, 1987; but see Humphrey & Khan, 1992). It is also possible that stereo information would not have yielded an advantage under brief presentations, which may hinder the building up of more fully integrated representations containing depth and 3-D structure information (and possibly also hinder their transformation in the matching process). However, Uttal, Davis, and Welke (1994) showed that compelling stereo depth can be recovered from very brief stimulus presentations (<1 msec). In any case, our results show that there is a stereo advantage in generalizing to new orientations with a range of stimuli and tasks that are similar in important respects to those used in a number of previous object recognition studies.

Marr and Nishihara (1978) first proposed that, in the process of deriving a description of an object, the visual system derives an intermediate representation that incorporates depth information at each image point—what they called the *2.5-D sketch* (see also Marr, 1982). This representation is, in Marr's view, subsequently used to reconstruct a full 3-D representation of objects in the world, but is not part of the object representation per se. Recently, however, several investigators have challenged this last reconstruction stage in object recognition. Rather, on the basis of the pattern of view dependency often observed in recognition performance, these investigators suggest that this intermediate representation is sufficient for recognizing objects (see, e.g., Edelman & Bülthoff, 1992; Williams & Tarr, 1999). Using a very different possible–impossible object decision task, Williams and Tarr reached a similar conclusion. Our results provide direct evidence for this proposal and consequently constrain any computational theory of object recognition (see also Bülthoff et al., 1995).

To conclude, although the data suggest that stereo information leads to a more robust representation that allows better generalization to new views, this representation does not embody a full 3-D, object-centered specification of object structure. Rather, the pattern of results in the present study suggests that—at least with the stimuli and tasks used—subjects encode representations that embody view-dependent depth and 3-D structural information. The representations that seem to be involved are reminiscent of Marr's (1982) 2.5-D sketch.

REFERENCES

Bennett, D. J. (in press). Does an estimate of environmental size precede size scaling on a form comparison task? *Perception*.

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

Bülthoff, H. H. (1991). Shape from X: Psychophysics and computation. In M. S. Landy & J. A. Movshon (Eds.), *Computational models of visual processing* (pp. 305-330). Cambridge, MA: MIT Press.

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

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

Bülthoff, H. H., & Mallot, H. A. (1988). Integration of depth modules: Stereo and shading. *Journal of the Optical Society of America A*, **5**, 1749-1758.

Burke, D. (2005). Combining disparate views of objects: Viewpoint costs are reduced by stereopsis. *Visual Cognition*, **12**, 705-719.

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

Farah, M. J., Rochlin, R., & Klein, K. L. (1994). Orientation invariance and geometric primitives in shape recognition. *Cognitive Science*, **18**, 325-344.

Hayward, W. G. (1998). Effects of outline shape in object recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **24**, 427-440.

Humphrey, G. K., & Khan, S. C. (1992). Recognising novel views of three-dimensional objects. *Canadian Journal of Psychology*, **46**, 170-190.

Jolicœur, P. (1985). The time to name disoriented natural objects. *Memory & Cognition*, **13**, 289-303.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research*, **35**, 389-412.

Liu, Z., Knill, D. C., & Kersten, D. (1995). Object classification for human and ideal observers. *Vision Research*, **35**, 549-568.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: 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: Series B*, **200**, 269-294.

Newhouse, M., & Uttal, W. R. (1982). Distribution of stereoanomalies in the general population. *Bulletin of the Psychonomic Society*, **20**, 48-50

Patterson, R., & Fox, R. (1984). The effect of testing method on stereoanomaly. *Vision Research*, **24**, 403-408.

Poggio, T., & Edelman, S. [Y.] (1990). A network that learns to recognize three-dimensional objects. *Nature*, **343**, 263-266.

Rock, I., & DiVita, J. (1987). A case of viewer-centered object perception. *Cognitive Psychology*, **19**, 280-293.

Sinha, P., & Poggio, T. (1994). *View-based strategies for 3D object recognition* (AI Memo No. 1518, CBCL Paper No. 106). Cambridge, MA: Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

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

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

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

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience*, **1**, 275-277.

Ullman, S. (1979). *The interpretation of visual motion*. Cambridge, MA: MIT Press.

Ullman, S., & Basri, R. (1991). Recognition by linear combination of models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **13**, 992-1006.

Uttal, W. R., Davis, N. S., & Welke, C. (1994). Stereoscopic perception with brief exposures. *Perception & Psychophysics*, **56**, 599-604.

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

Watt, S. J., Akeley, K., Ernst, M. O., & Banks, M. S. (2005). Focus cues affect perceived depth. *Journal of Vision*, **5**, 834-862.

Williams, P., & Tarr, M. J. (1999). Orientation-specific possibility priming for novel three-dimensional objects. *Perception & Psychophysics*, **61**, 963-976.

NOTE

[1] The stereo and nonstereo conditions were conducted on separate PC machines because of technical problems. However, we matched important aspects of the experiment (e.g., ambient light, screen resolution, refresh rate, image size, experiment code). Furthermore, there was essentially no difference in recognition performance in the two viewing conditions when the view orientation remained the same (see Figures 11 and 12).

(Manuscript received February 23, 2005;
revision accepted for publication October 17, 2005.)