Psychonomic Bulletin & Review
1995, 2 (1), 55–82

Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects

MICHAEL J. TARR
Yale University, New Haven, Connecticut

Successful object recognition is essential for finding food, identifying kin, and avoiding danger, as well as many other adaptive behaviors. To accomplish this feat, the visual system must reconstruct 3-D interpretations from 2-D "snapshots" falling on the retina. Theories of recognition address this process by focusing on the question of how object representations are encoded with respect to viewpoint. Although empirical evidence has been equivocal on this question, a growing body of surprising results, including those obtained in the experiments presented in this case study, indicates that recognition is often viewpoint dependent. Such findings reveal a prominent role for viewpoint-dependent mechanisms and provide support for the *multiple-views* approach, in which objects are encoded as a set of view-specific representations that are matched to percepts using normalization procedures.

> Just as you could not fully reconstruct a house from photos all taken from a single vantage point, "snapshots" at many angles must be combined to reconstruct a Burgess organism. Conway Morris told me that he managed to reconstruct the curious *Wiwaxia*—an animal with no modern relatives, and therefore no known prototype to use as a model—by passing countless hours "rotating the damned thing in my mind" from the position of one drawing to the different angle of another, until every specimen could be moved without contradiction from one stance to the next. Then he finally knew that nothing major was missing or out of place.
>
> —Stephen J. Gould, *Wonderful Life* (1989)

The human ability to recognize objects is remarkable—under all but the most degraded conditions, we succeed in understanding the perceptual world around us. This performance is even more astounding when one considers that we encounter 3-D objects from an infinite number of viewpoints producing potentially unique 2-D projections on the retina. The problem, as expressed by Gould (1989) with reference to the paleontologist's reconstruction of 3-D organisms, is that one must find "guides to the three dimensional reanimation of squashed and distorted fossils"—no less than fossils, 2-D images must be "reanimated."

To achieve *shape constancy* (object perception regardless of position, scale, or viewpoint), retinal images must not only be organized into coherent percepts, but must be compared with long-term representations of objects previously seen. The format of these representations may take many forms. For instance, objects might be encoded as spatial representations, descriptive lists of features, or Fourier decompositions of component sine-wave gratings. A growing body of research supports the concept of spatial representations, although the precise mechanisms used to match these to percepts are not yet well understood. Current theories of object recognition have varied widely, arguing for recognition by the location of small sets of unique features (Corballis, 1988), for recognition by the alignment of 2-D input shapes with 3-D models (Ullman, 1989), for recognition by normalizing 2-D input shapes to viewpoint-specific 2-D views (Bülthoff & Edelman, 1992), or for recognition by comparing recovered qualitative descriptions of 3-D parts and their spatial relations (Biederman, 1987; Hummel & Biederman, 1992). Recognition may also rely on texture, shading, color, or motion. It is almost certain that many of these possibilities coexist as mechanisms in visual perception; however, because the latter attributes are often considered precursors to the recovery of shape

This research was conducted in partial fulfillment of the requirements for the PhD in the Department of Brain and Cognitive Sciences at the Massachusetts Institute of Technology. Financial support at MIT was provided by the James R. Killian Fellowship sponsored by the James and Lynelle Holden Fund, a Fellowship from the Whitaker Health Sciences Fund, and an NSF Graduate Fellowship. In addition, parts of this research were funded under NSF Grant BNS 8518774 to Steven Pinker, Professor of Brain and Cognitive Sciences at MIT, and a grant from the Sloan Foundation to the MIT Center for Cognitive Science. Final revisions were completed during a visit to the Max-Planck-Institut für biologische Kybernetik in Tübingen. Jigna Desai and Carmita Signes ran many of the subjects and were stellar research assistants. Many thanks to Kyle Cave, Jacob Feldman, Paul Bloom, Larry Maloney, William Hayward, Pierre Jolicoeur, James Pomerantz, two anonymous reviewers, and in particular to my thesis committee members, Irving Biederman, Ellen Hildreth, and David Irwin for their time, insights, and support. Michael Kubovy provided useful comments on an earlier version. Special thanks to Steve Pinker and to Laurie Heller. Requests for reprints may be sent to M. J. Tarr, Department of Psychology, Yale University, P.O. Box 208205, New Haven, CT 06520-8205 (e-mail: tarr@cs.yale.edu).

 Copyright 1995 Psychonomic Society, Inc.

(Horn & Brooks, 1989), theories of recognition have focused on shape-based representations.

Families of Recognition Theories

Theories of object recognition may be characterized in terms of four issues: (1) the coordinate system or frame of reference (viewpoint dependent or viewpoint invariant); (2) the nature of component features (e.g., surfaces, segmented contours, or parts); (3) the encoding of relations between features (metrically specific or qualitative); and (4) the number of spatial dimensions (2-D, 2-D plus depth, or 3-D). These issues are theoretically independent, but in practice, they tend to cluster into one of several families. Consequently, tests of a single issue—for example, the frame of reference used in the representation—are often taken as generally distinguishing between families. Indeed, approaches to shape constancy may be divided roughly into *complete viewpoint-invariant theories*, *restricted viewpoint-invariant theories*, and *viewpoint-dependent theories*.

Complete viewpoint-invariant theories. This family includes both recognition by orientation-free unique features (Corballis, 1988; Jolicoeur, 1990a) and object-centered theories in which an object is described within the same coordinate system regardless of its orientation, size, or location (Marr & Nishihara, 1978). The essential point in either approach is that object representations and percepts use a common description, derived by recovering a view-independent coordinate system or identifying view-invariant attributes. In either case, an object may be recognized without the need for normalization between familiar and unfamiliar viewpoints.[1]

Restricted viewpoint-invariant theories. This family includes theories in which object representations are composed of qualitatively defined configurations of features (Koenderink, 1990) or structural descriptions (Biederman, 1987; Cooper, Schacter, Ballesteros, & Moore, 1992). The essential point is that percepts arising from a range of adjacent viewpoints will correspond to a single object representation as long as the same viewpoint-invariant configuration of features or parts is available. However, each distinct qualitative configuration will necessitate a separate object description. Thus, if the qualitative configuration is familiar, an object may be recognized over a restricted range of unfamiliar viewpoints without the need for normalization.

Viewpoint-dependent theories. This family includes recognition by alignment to a single canonical view (Palmer, Rosch, & Chase, 1981; Ullman, 1989) or to *multiple views* (Bülthoff & Edelman, 1992; Tarr & Pinker, 1989). The essential point is that the object is represented in a viewer-centered frame of reference determined by the location of the viewer in relation to the object. In the canonical view case, an object may be recognized by normalizing it to the single most salient viewpoint. In the multiple-views case, an object may be recognized by normalizing it to the nearest familiar viewpoint (dubbed *multiple-views-plus-transformation* by Tarr & Pinker, 1989).

Each of these approaches leads to specific predictions about the effect of viewpoint on recognition performance. Viewpoint-invariant theories predict performance that is equivalent across all viewpoints (with the exception of accidental views) or across a range of adjacent viewpoints (assuming that a qualitative description is stable over some variation in viewpoint; Biederman & Gerhardstein, 1993; Koenderink, 1987). In contrast, viewpoint-dependent theories predict performance that varies with the specific viewpoints that are represented. For instance, given a multiple-views representation, recognition performance may be viewpoint invariant at all *familiar* views and therefore indistinguishable from the pattern predicted by viewpoint-invariant approaches. However, because it is often assumed that viewpoint normalization procedures operate with time and/or accuracy proportional to the degree of normalization, recognition performance at *unfamiliar* views is predicted to be dependent on viewpoint. In the canonical-views approach, performance will degrade with increasing distance from the canonical view, whereas in the multiple-views approach, performance will degrade with increasing distance from the nearest familiar view in the representation.

Studies of Discriminations on Shapes Misoriented in the Picture Plane

A survey indicates that there is little consensus concerning which of these approaches best accounts for how the human visual system accommodates variations in viewpoint. Much of this debate centers on the fact that changes in the viewpoint or picture-plane orientation of objects have yielded both viewpoint-dependent and viewpoint-invariant patterns of performance. For example, one of the best known demonstrations of viewpoint dependence in a perceptual task was provided by R. N. Shepard and his colleagues (Cooper & R. N. Shepard, 1973; R. N. Shepard & J. Metzler, 1971; for a comprehensive review see R. N. Shepard & Cooper, 1982). They found that when subjects discriminated standard from mirror-reversed shapes or compared simultaneously presented objects at different viewpoints, response times increased monotonically with increasing misorientation from the upright or between the objects. These effects were taken as evidence for an incremental transformation process known as "mental rotation," which is used to normalize misoriented shapes under many circumstances. However, these results were *not* used as evidence for viewpoint-dependent recognition mechanisms. Shepard and Cooper's tasks all involved handedness discriminations, not naming or identification, so their findings do not address whether mental rotation is used in recognition. In fact, Shepard and Cooper argued that in order to locate the top of an object, subjects must have identified objects by using viewpoint-invariant mechanisms *prior* to mental rotation. Consequently, they suggested that the viewpoint-dependent process of mental rotation is used only to determine handedness.

Subsequent studies appear to support this argument. Corballis, Zbrodoff, Shetzer, and Butler (1978) had subjects name misoriented letters and digits; they found that the time subjects took to name standard (i.e., not mirror-reversed) versions of characters was largely independent of the orientation of the character. Corballis and Nagourney (1978) found that when subjects classified misoriented characters as letters or digits, there was also only a small effect of orientation on decision time. Similarly, White (1980) found that neither category nor identity judgments, preceded by a correct cue for category or identity, exhibited an effect of orientation for either standard or mirror-reversed characters. However, White did find a linear effect of orientation on handedness judgments of the same characters preceded by correct handedness cues. Simion, Bagnara, Roncato, and Umiltà (1982) had subjects perform same/different judgments on simultaneously presented letters separated by varying amounts of rotation. In several of their experiments, they found reliable effects of orientation on response time, but the effect was too small to be attributed to mental rotation as originally conceived. Finally, Eley (1982) found that letter-like shapes containing a salient diagnostic feature (e.g., a small closed curve in one corner, or an equilateral triangle in the center) were recognized equally quickly at all orientations.

The rotation-for-handedness hypothesis. On the basis of these effects, Corballis et al. (1978; see also Corballis, 1988; Hinton & Parsons, 1981) concluded that under most circumstances object recognition (up to, but not including, handedness information) is accomplished by matching percepts to viewpoint-invariant representations. In contrast, handedness discriminations appear to necessitate the use of mental rotation, because the postulated representations do not encode handedness—they match both handedness versions of an object equally well. Therefore, subjects must use other means to assess handedness. Hinton and Parsons suggest that because handedness is inherently egocentric, subjects determine handedness by examining which parts of an object correspond to their left and right sides at the upright. Thus, if an object is misoriented, it must be normalized to the upright canonical view. A similar suggestion has been made by Biederman and Gerhardstein (1993), who posit that recognition is normally viewpoint invariant unless contaminated by the "need to distinguish between mirror reflections of the same object" (p. 1163). Tarr and Pinker (1989) refer to this as the *rotation-for-handedness* hypothesis and conclude that such a hypothesis relegates the use of normalization procedures in object recognition to the "highly circumscribed role" of assigning handedness.

There are some difficulties with evidence supporting the rotation-for-handedness hypothesis. One problem is that in some of the experimental demonstrations of viewpoint invariance, the stimuli contained diagnostic features that subjects may have exploited in their discriminations (a similar criticism may be applied to the studies presented by Biederman & Gerhardstein, 1993; see Tarr & Bülthoff, in press). In Eley's (1982) study, the presence of diagnostic features was deliberate; in White's (1980) study, a cue for either identity or category information permitted subjects to prepare for the task by selecting appropriate diagnostic features. In contrast, a cue for handedness did not permit subjects to prepare for the handedness judgment, since handedness information alone does not specify any specific feature prior to viewing the particular object. Another difficulty is that, disregarding studies subject to the diagnostic feature critique, with one exception (Tarr & Pinker, 1990), viewpoint-invariant recognition has been demonstrated only for familiar objects (e.g., standard versions of English characters or common objects); when novel objects must be recognized, viewpoint-dependent effects are obtained. For instance, using a set of novel polygonal shapes learned at a single orientation, Shinar and Owen (1973) found that familiar/unfamiliar judgment times on misoriented versions of the shapes were orientation dependent, but that this effect disappeared with practice. Similar findings have been reported by Jolicoeur (1985), who had subjects name line drawings of natural objects. Initially, naming times increased as the drawings were oriented farther from the canonical upright. With practice, however, the effects of orientation diminished, suggesting that as objects become increasingly familiar, subjects become less sensitive to their orientation. One explanation for this is that subjects are developing viewpoint-invariant representations, such as object-centered representations or collections of orientation-free features. An alternative explanation is that objects are represented as multiple views, one for each familiar orientation, at which point recognition at familiar viewpoints could be performed without the need for normalization.

It is the latter alternative that complicates the interpretation of studies done with letters, digits, and familiar common objects. Since such shapes are highly familiar, subjects have had a great deal of prior experience recognizing them from many viewpoints (Jolicoeur, 1985; Koriat & Norman, 1985). Therefore, it is possible that subjects have multiple viewpoint-specific representations for each character or object. This account is compatible with Corballis et al.'s (1978) study: although there was only a small effect of orientation on naming latencies for standard versions of characters, there were large effects of orientation on naming latencies for reversed versions. The latter result is consistent with the multiple-views explanation if subjects rarely see mirror-reversed versions of letters and digits (Koriat & Norman, 1985). Corballis et al. also reported a decrease in the effect of orientation with practice—a finding consistent with the hypothesis that new viewpoint-specific representations are encoded at previously unfamiliar viewpoints.

Studies of Discriminations on Objects Rotated in Depth

Because most objects present different visible surfaces across different viewpoints, there is some expec-

tation that viewpoint-dependent mechanisms will be used to achieve shape constancy across rotations in depth (regardless of the mechanisms used across picture-plane rotations). Indeed, in a well-known study of recognition of familiar common objects rotated around the vertical axis, Palmer et al. (1981) found that naming times became progressively slower as the rotation increased away from each object's "canonical" view (established independently via ratings of the preferred view for each object). Moreover, the finding that some objects have been found to have more than one canonical view provides possible evidence for the existence of multiple views. For instance, human heads appear to have two views from which they are most readily identified: head-on and profile. However, such interpretations must be tempered by the fact that the viewpoint effects obtained in Palmer et al.'s study were somewhat smaller than those usually attributed to normalization processes (Cohen & Kubovy, 1993).

Other studies employing familiar objects have likewise yielded somewhat equivocal results. In a study of the effects of practice on picture naming, Bartram (1974) found that naming times diminished most rapidly if the identical picture of an object was presented consistently, but diminished somewhat less if different viewpoints of the same object were presented, and, finally, diminished the least if different exemplars with the same name were presented. This finding is consistent with an exemplar-based account of how multiple views are acquired: specific views will achieve greater familiarity (and hence higher activation and salience during recognition) if presented repeatedly, but the presentation of new views or new objects will distribute the effects of familiarity across viewpoints or exemplars of a class. However, Bartram also obtained evidence for viewpoint-invariant representations: while practice in naming an object from one or several viewpoints did not transfer to different exemplars with the same name, practice in naming an object from a single viewpoint transferred to naming the same object in different viewpoints. This result suggests that subjects were able to derive a viewpoint-invariant representation during practice or were able to make reliable inferences about the appearance of objects from unseen viewpoints ("virtual views"; see Vetter, Poggio, & Bülthoff, 1994).

Bartram (1976) also investigated recognition performance across changes in viewpoint or exemplar in a sequential matching paradigm. Individual objects were presented sequentially, and subjects responded whether the two pictures were the same or different. For line drawings of the objects, the results were consistent with Bartram's 1974 study: response times were fastest for identical pictures, somewhat slower for the same object appearing in differing viewpoints, and slowest for different same-name exemplars. In contrast, using photographs, Bartram found an interaction with the frequency of object class names: for the low-frequency condition, the response time cost for changing viewpoint was nearly as high as the cost for changing exemplar; for the high-frequency condition, there was almost no difference between the response times for identical viewpoints and changed viewpoints, but there was a large cost for changing exemplar (equivalent to that observed for the low-frequency condition). Although Bartram interpreted these findings as evidence for a viewpoint-invariant "object-code," they may be consistent with multiple-views theory if there is a correlation between name frequency and frequency of observation. If this is the case, then high-frequency object classes may be represented by relatively more familiar views than are low-frequency object classes.

More recently, Lawson (1993) has employed line drawings of familiar objects in a series of experiments that included both a sequential picture matching paradigm and a sequential view naming paradigm. The former was similar to that used by Bartram (1976), and the latter involved subjects' naming a single object following its presentation in six views. Lawson found some support for viewpoint-specific object representations—in particular, demonstrating a benefit for canonical views, a benefit for matching views sharing similar image structure, and a benefit for viewing structured view sequences over random view sequences.

In contrast to the explicit tasks in the preceding studies, Biederman and Gerhardstein (1993) and Srinivas (1993) have employed implicit memory priming tasks (for a general discussion of implicit memory, see Roediger, Weldon, & Challis, 1989). In Biederman and Gerhardstein's study (Experiments 1 and 2), the task was entry-level naming of line drawings of common objects across rotations in depth (the entry level may be defined as the categorical level assigned to objects during initial recognition; see Jolicoeur, Gluck, & Kosslyn, 1984). In the study phase, subjects named a series of objects. In the test phase, subjects named the same exemplars or different exemplars of the same object categories shown during study. In either case, at test exemplars appeared at the same viewpoint or at one of several new viewpoints. Two main effects were observed: *same* exemplars were named faster than *different* exemplars of the same category, and naming times for *same* exemplars were relatively invariant across changes in viewpoint. Biederman and Gerhardstein claim that the failure to find effects of viewpoint in an entry-level naming task suggests that objects are "typically" recognized by using viewpoint-invariant mechanisms (but see Tarr & Bülthoff, in press, who argue that viewpoint invariance may have been obtained because of multiple views for familiar objects).[2]

Srinivas (1993) has also reported several experiments in which she used a naming task. These experiments provide one advantage over Biederman and Gerhardstein's (1993): the stimuli consisted of photographs rather than line drawings of familiar objects. Srinivas used a study phase in which objects were displayed for a relatively long period (2 sec) and then named, followed by a test phase in which objects were displayed for a very brief duration (50 msec in one experiment, and 34 msec and

84 msec in two others) and once again named. In both experiments, Srinivas found an overall decrease in priming with a change in viewpoint from study to test. It was also verified that there was a visual component to these effects by showing that priming for different views was greater than priming for object names. Srinivas interpreted these results as evidence for a multiple-views theory of object recognition in which objects are recognized by normalizing them to view-specific representations. A second effect also provides evidence for view-specific representations. Srinivas reported a somewhat larger decrease in priming when objects were studied in familiar views and tested in unfamiliar views. This result is consistent with the hypothesis that objects are represented at specific familiar views: when either a familiar or an unfamiliar view was presented at study it would have been recognized by a match to a familiar view, so that subsequent testing of that familiar view would result in the largest priming effect. In contrast, subsequent testing of an unfamiliar view would result in somewhat less priming, since recognition would still be mediated by a match to a familiar view.

Several considerations cloud any definitive conclusions from all of the studies just reviewed. First, effects of viewpoint that arise during initial exposure to unfamiliar objects (either because they are completely novel or because they are novel exemplars of familiar object classes) may be transient. Viewpoint-invariant representations may be learned only over repeated exposure. Second, as suggested by Koenderink (1987) and by Biederman and Gerhardstein (1993) viewpoint invariance may be restricted to adjacent viewpoints from which the same qualitative features are visible. According to the latter claim, viewpoint-dependent performance may have been obtained in some of these studies because manipulations of viewpoint failed to control for whether two views (either presented as a sequential match or at study and test) contained the same qualitative configuration. In addition, regardless of whether qualitative changes occur, one must also control for accidental viewpoints (e.g., foreshortened views; see Humphrey & Jolicoeur, 1993) that interfere with normal shape processing.

To control for the possibilities of previously learned multiple views and for diagnostic features, several researchers have employed realistically shaded novel 3-D objects of similar shape. Bülthoff and Edelman (1992, 1993; Edelman & Bülthoff, 1992) used blob-like smooth objects with protruding parts ("amoeboid" objects) and tube-like objects ("paper-clip" objects—similar to those used by Rock & Di Vita, 1987; Rock, Di Vita, & Barbeito, 1981; Rock, Wheeler, & Tudor, 1989) in a same/different single-interval forced choice paradigm. A similar procedure was employed in each experiment: a target object was studied over a range of viewpoints to simulate motion; it was then tested, intermixed with similar distractors displayed at single viewpoints. Response accuracy was generally dependent on the degree of change in viewpoint, but there were differences in the patterns of generalization for unfamiliar views interpolated between studied viewpoints, unfamiliar views extrapolated beyond studied viewpoints, and unfamiliar views generated by rotations around an axis orthogonal to that used during study. Bülthoff and Edelman (1992) suggest that these patterns are inconsistent with viewpoint-invariant theories and viewpoint-dependent theories that rely on linear normalization mechanisms (e.g., alignment models, such as that proposed in Ullman, 1989). Rather, the results appear to support a viewpoint-dependent theory in which objects are recognized by interpolating between familiar 2-D views (Ullman & Basri, 1991). Using similar stimuli, Edelman and Bülthoff (1992) demonstrated that equal exposure to displayed viewpoints still leads to variations in recognition accuracy for these viewpoints. This result indicates that certain views are canonical, regardless of their familiarity. Importantly, their results also established that preferred viewpoints remained following extensive practice with many viewpoints and the inclusion of cues to the 3-D structure (i.e., stereo). Thus, their observation of preferred views cannot be accounted for by initial effects of viewpoint that diminish as viewpoint-invariant representations are acquired. Overall, such results provide further evidence for the multiple-views approach and reinforce the extreme viewpoint dependence of exemplar-specific recognition judgments.

Similar conclusions were reached by Humphrey and Khan (1992) on the basis of three recognition memory experiments (old/new) in which they employed photographs of somewhat more natural novel objects formed out of clay. Humphrey and Khan established that recognition memory for objects was equivalent at several different viewpoints when there was no change in viewpoint between study and test. This finding controls for the possibility that apparently viewpoint-dependent performance is due to variations in shape properties such as foreshortening or canonicity. In two subsequent experiments, recognition memory performance varied systematically with changes in viewpoint from study to test. These results provide still more evidence that object representations are specific to familiar views.

Direct Evidence for View-Specific Representations

As in Bülthoff and Edelman's (1992) and Humphrey and Khan's (1992) studies, Tarr and Pinker (1989) employed stimulus objects that were highly similar to one another, in order to preclude recognition strategies based on unique features. As such, discriminating between the objects was an exemplar-specific rather than categorical task. Each novel 2-D character had a clearly marked bottom and a vertical axis, thereby minimizing possible viewpoint-dependent effects related to finding the top, bottom, or major axis (e.g., McMullen & Jolicoeur, 1992). The experimental procedure was predicated on the different qualitative predictions made by viewpoint-invariant and viewpoint-dependent theories regarding recognition at unfamiliar views. Viewpoint-invariant theories predict roughly equivalent performance across all viewpoints or equivalent performance across all

viewpoints up to a change in visible features or parts; viewpoint-dependent theories predict performance that varies systematically with the distance from the nearest familiar viewpoint. To test these predictions, subjects were taught the objects in a single viewpoint (the upright or near the upright) and then were given extensive practice in naming the objects in several unfamiliar orientations generated by rotations in the picture plane. The subjects were then probed with the now-familiar objects presented in new unfamiliar orientations. Generally, viewpoint-dependent theories predict an initial effect of viewpoint, whereas viewpoint-invariant theories predict no effect of viewpoint. This difference alone is insufficient for testing between theories, however, in that effects of viewpoint may diminish with practice because subjects require time to learn viewpoint-invariant descriptions or because they learn multiple views at familiar practiced orientations. Thus, both types of theories predict equivalent recognition times across all *familiar* viewpoints following practice at naming the objects. In contrast, only multiple-views theories predict that performance will be viewpoint dependent for *unfamiliar* viewpoints of familiar objects—viewpoint-invariant theories predict equivalent recognition times regardless of whether the viewpoint is familiar or not (as long as the qualitative configuration is familiar).

Across all four of Tarr and Pinker's experiments, three results stand out. First, when subjects initially recognized the objects at several orientations in the picture plane, performance was related monotonically to the degree of rotation from the training viewpoint. Second, with further practice at recognizing the objects in these viewpoints, performance became roughly equivalent.[3] Third, following practice, recognition performance at unfamiliar viewpoints varied with the angular difference between the unfamiliar viewpoint and a familiar viewpoint. These findings provided evidence that the novel characters were represented as multiple views and that recognition was accomplished by normalizing the objects in unfamiliar views to the nearest familiar view. In addition, similar effects of viewpoint were obtained when the objects were first recognized and after practice when the then-familiar objects were recognized in unfamiliar viewpoints. In both instances, the effects of viewpoint were comparable to the rates reported for studies in which converging techniques were used to demonstrate the analog nature of mental rotation. The effects were also comparable to the rates obtained in a control experiment run by Tarr and Pinker with the same objects used in handedness discriminations (a task that uncontroversially requires mental rotation). Such equivalence provided further evidence that the normalization mechanisms used for recognition are similar to the processes used in mental rotation.

Tarr and Pinker's (1989) results are inconsistent with the hypothesis that object recognition is accomplished *exclusively* through viewpoint-invariant mechanisms (see, e.g., Biederman, 1987; Corballis, 1988; Marr & Nishihara, 1978). Such theories predict that the diminished effects of viewpoint that come with practice at specific viewpoints should transfer to unfamiliar viewpoints, which they do not. Moreover, because some studies have failed to control for prior familiarity and unique features, these results indicate that some apparently viewpoint-invariant effects may be accounted for by multiple views (but not all; see Murray, Jolicoeur, McMullen, & Ingleton, 1993). These results also challenge the claim that exemplar-specific discriminations rely on viewpoint-invariant contrasts rather than viewpoint-dependent normalization mechanisms (Biederman & Gerhardstein, 1993; Biederman & Shiffrar, 1987). Indeed, in almost every study done with a discrimination between structurally similar objects, large effects of viewpoint have been obtained (e.g., Bülthoff & Edelman, 1992; Humphrey & Khan, 1992; Tarr & Pinker, 1989). Finally, because Tarr and Pinker's tasks *did not* require the assignment of handedness, these results falsify the conjecture that normalization processes are used only when the task requires discriminating between left and right (Biederman & Gerhardstein, 1993; Corballis, 1988). Furthermore, it is unlikely that subjects surreptitiously tried to determine handedness: in one experiment, handedness was made explicitly irrelevant by equating standard and mirror-reversed shapes.

Multiple Views in Three Dimensions

Tarr and Pinker (1989) used a small set of similar 2-D objects, composed of line segments (as opposed to surfaces) and misoriented only in the picture plane. Because of these limitations, their findings may not extend to the recognition of 3-D (and perhaps even 2-D) objects under more ecological conditions, particularly at arbitrary viewpoints in 3-D space. For example, determining the pose of an object or the appropriate axis for normalization may be difficult when the axis does not lie along the line of sight or the major axes of the object. If the human visual system can not determine such information, normalization mechanisms would be of little utility in generic object recognition. Likewise, because many objects maintain a constant orientation with respect to gravity, recognition mechanisms restricted to picture-plane rotations may not generalize to misorientations around other axes (e.g., Hummel & Biederman's, 1992, account for viewpoint-dependent picture-plane effects does not extend to depth rotations). Thus, one alternative account of Tarr and Pinker's (1989) results is that rotations in the picture plane are a special case that the "default" recognition system is incapable of handling. This would be the case, for instance, if viewpoint-invariant mechanisms relied on readily available labels for top and bottom (arboreal species such as monkeys could not rely on such a mechanism because top and bottom are likely to be perturbed quite frequently). Testing this alternative is straightforward: just as mental rotation

has been demonstrated for objects rotated in depth (R. N. Shepard & J. Metzler, 1971), recognition may be examined for novel 3-D objects rotated in depth.

Notably, some findings already indicate that the multiple-views approach generalizes to 3-D. In addition to the results of Bülthoff and Edelman (1992; Edelman & Bülthoff, 1992) and Humphrey and Khan (1992), subjects in Tarr and Pinker's (1989) study who were familiar with the standard versions of 2-D characters appeared to recognize mirror reversals by normalizing them *in depth* around an axis within the frontal plane. This 180º rotation in depth is the shortest path for aligning a mirror reversal of a 2-D shape with its standard (see Parsons, 1987a, 1987b). This finding suggests that recognition mechanisms are capable of using approximately the shortest path rotation, regardless of the axis.[4] Despite this positive evidence, there have been no specific demonstrations of *multiple-views* (as opposed to simply viewpoint-dependent) mechanisms in the recognition of novel 3-D objects rotated in depth. This question was investigated in the experiments presented below.

Multiple Views in the Recognition of 3-D Objects Rotated in Depth

Each of four experiments was designed as a 3-D counterpart to one of Tarr and Pinker's (1989) experiments. The same paradigm was used: subjects were trained on novel objects, here 3-D, and their mirror reversals (enantiomorphs) in a near-upright training viewpoint. During both the initial practice and the probe phases, the objects appeared rotated in depth around the line of sight, the horizontal axis, or the vertical axis.

While multiple-views theories predict that effects of viewpoint will be related to the nearest familiar viewpoint, viewpoint-invariant theories, such as those of Corballis (1988) and Marr and Nishihara (1978), predict no effect of viewpoint for unfamiliar viewpoints. However, a restricted viewpoint-invariant model, such as that proposed by Hummel and Biederman (1992), may predict some effect of rotation in depth. Specifically, because top and bottom must be assigned prior to recognition, their model predicts an effect of viewpoint for frontal plane misorientations. A more recent version of this proposal, as set forth by Biederman and Gerhardstein (1993), only predicts viewpoint invariance for viewpoint changes that do not alter the configurations of features defining the structural description (resulting in a step-like function, but not performance varying monotonically with variations in viewpoint).

EXPERIMENT 1

Experiment 1 provides a baseline measure of the effects of normalization obtained with the stimuli used throughout this study and examines whether subjects learn version- and viewpoint-specific representations in handedness discrimination tasks. Such tasks provide the robust effects of viewpoint on response times that are commonly attributed to normalization processes (R. N. Shepard & Cooper, 1982). Therefore, the pattern of response times and putative rates of rotation from this experiment are useful for comparison with the results of the subsequent experiments, particularly as we assess the role of normalization processes. This experiment also provides a baseline for the rate at which effects of viewpoint diminish with practice. Although practice effects can be attributed to handedness-specific object representations without regard for whether they are viewpoint-dependent or invariant, this experiment also tests their viewpoint specificity. If viewpoint-invariant representations are used, handedness discriminations at unfamiliar viewpoints should take no longer than discriminations at familiar viewpoints—at a minimum, there should be no *systematic* effect of viewpoint on discriminations at unfamiliar viewpoints. Alternatively, if viewpoint-dependent representations are used, handedness discriminations at unfamiliar viewpoints should take increasingly longer with increasing angular distance from familiar viewpoints. However, as suggested by the rotation-for-handedness hypothesis (Biederman & Cooper, 1991; Biederman & Gerhardstein, 1993; Corballis, 1988; Corballis et al., 1978; Hinton & Parsons, 1981), viewpoint-dependent performance for handedness discriminations does not provide evidence for the same mechanisms in recognition. Rather, handedness may be determined by using specialized handedness-specific and, most likely, viewpoint-dependent representations that develop only in response to the need to perform this particular task.

Method

Subjects. Twelve students from the Boston area participated for pay.

Materials. The stimuli consisted of seven left/right and front/back asymmetrical objects described in the Appendix and illustrated in Figure 1 in the training viewpoint. This viewpoint was generated by rotating each object 10º from the upright around each axis.[5] Both the standard and the reversed versions were used. The stimuli were drawn in perspective at the training viewpoint and at 11 viewpoints around each axis (+30º steps, with rotations around the other two axes fixed at 10º each) on an IBM PC EGA monitor. The CRT was approximately 38 cm from a chinrest and the objects were drawn in a 13-cm-diameter (19.4º of visual angle) circle. The surfaces of the objects were of uniform color and the edges of the faces of each cube were drawn with hidden lines removed. To guard against the idiosyncratic effects of a particular object, the objects were grouped into three sets of three named objects each: Set A was composed of objects 1, 2, and 3; Set B, of objects 4, 5, and 6; and Set C, of objects 2, 5, and 7. The three target objects were assigned the names "Kip," "Kef," and "Kor." One third of the subjects, who were not aware of the groupings, received training with each set.

Procedure. The subjects were shown both versions of the target objects on a CRT. The subjects named and built each version of each object five times. The objects were built out of toy blocks that connected at right angles to a prebuilt main axis with a "foot" common to all of the objects that was fixed to a base viewed at a viewpoint of (10º, 10º, 10º). By actually building the objects, subjects were given a great deal of feedback about the 3-D structure of each object—a manipulation that would be expected to bias

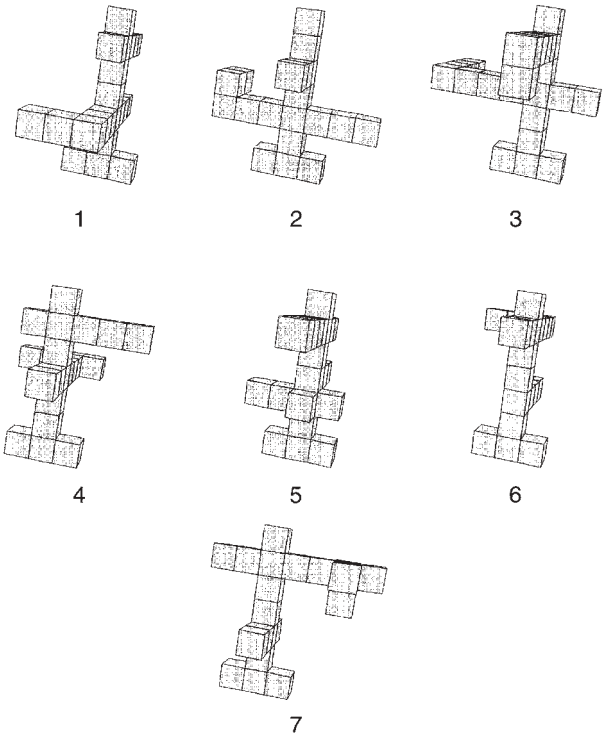

Figure 1. Standard versions of asymmetrical objects in their near-upright training viewpoint. In each of these objects, the bottom of the object is marked by a "foot" of three cubes that terminates the main axis, marking it as well.

them toward learning viewpoint-invariant representations. The subjects then built each object from memory according to the version and name specified by the experimenter. Feedback was given, and the subjects continued to build the objects until they built all three objects in both versions correctly twice.

Throughout the rest of the experiment, the objects were shown one at a time on a CRT. The subjects saw a fixation point, a "+," and then an object. They were instructed to decide as quickly as possible, while minimizing errors, whether the object was the standard or the mirror reversal of one of the trained objects, with the responses corresponding respectively to the right and left keys on a response board. Left-handed subjects had the option of reversing the response board. When subjects made an error, they heard a beep as feedback.

Design. Both the standard and the mirror-reversed versions of the objects were displayed at the training viewpoint and at rotations of 130º around the *x*-, *y*-, or *z*-axis (Figure 2). The subjects ran in 12 "practice" blocks consisting of 6 preliminary trials followed by 72 trials corresponding to the three objects in their standard and reversed versions in the training viewpoint six times and in the other 3 practice viewpoints two times. This was followed by a "surprise" block consisting of 6 preliminary trials followed by 432 trials corresponding to the three objects in their standard and reversed versions in the training viewpoint six times and in 33 other viewpoints, defined by +30º increments starting at 10º around the *x*-, *y*-, or *z*-axis, two times each.

In the surprise block, the 6 preliminary trials were composed of viewpoints used in practice blocks. In all blocks, the order of the trials following the preliminary trials was determined randomly. There was a self-timed break every 40 trials.

The subjects were run in a total of four sessions, each of which was approximately 1 hour long (training plus 2 blocks; 4 blocks; 4 blocks; 2 blocks plus surprise block—the practice blocks prior to the surprise block ensured that any effects in the surprise block were not due to a beginning-of-session effect). Not counting preliminary trials, each subject was run in a total of 26 trials for every object at a particular handedness and practice viewpoint and three times that number at the training viewpoint.

Results

Response time means for each viewpoint were calculated by block, averaging over all objects. Here and in subsequent experiments no clear "outliers" were apparent in the data, so no data were discarded other than incorrect responses and preliminary trials. It is generally accepted that clockwise and counterclockwise rotations of the same magnitude produce approximately equal response times because subjects normalize through the shortest path to the upright (R. N. Shepard & Cooper, 1982). This assumption may be extended to include ro-

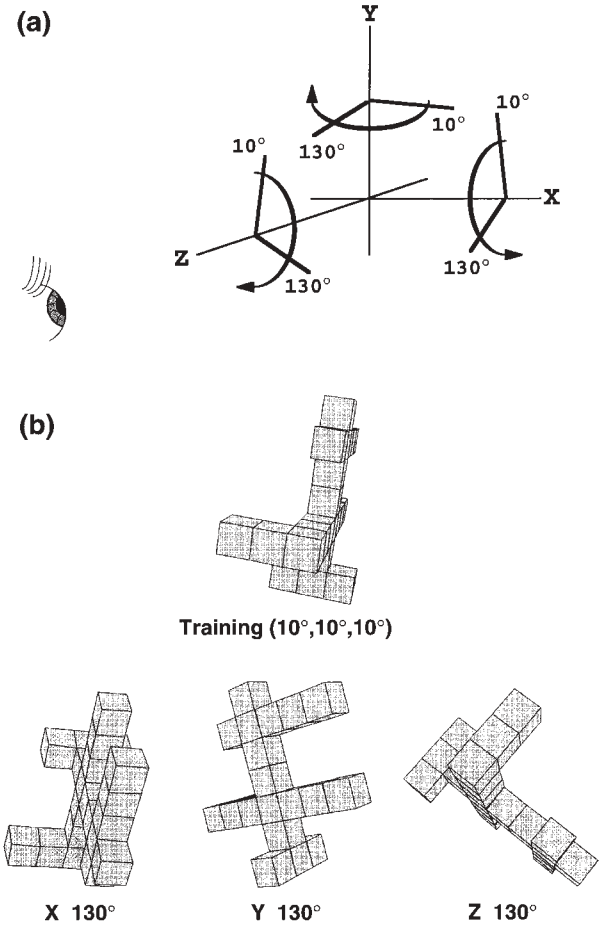

Figure 2. (a) Angular layout of practice viewpoints in all experiments—with the exception of the one-version condition of Experiment 3, in which the practice viewpoints were the training viewpoint and 40º, 70º, 100º, and 190º around each axis. The direction of rotation around each axis is indicated with an arrow. A rotation around one axis is accompanied by constant rotations of 10º around the other two axes. (b) Standard version of Object 1 in the four practice viewpoints.

tations of equivalent magnitude *around the same axis of rotation*, whether or not the rotation is in the picture plane (see Parsons, 1987c; misorientations were generated by a change in viewpoint around only one axis, meaning that the shortest path was always a rotation around the same axis used to generate the misorientation). This assumption is supported by the finding that mean response times for equidistant viewpoints around a common axis fall near a single straight line. Thus, the effect of viewpoint may be characterized by plotting the response time means against the shortest angular distance from a given viewpoint and calculating the putative rate of rotation as measured by the slope of the best-fitting line determined by the method of least squares.

Three analyses of variance (ANOVAs), one for each axis of rotation, for data collapsed over Blocks 1–12 with version and viewpoint as factors revealed a reliable effect of viewpoint ($p < .01$) for each axis [x, $F(1,11) = 28.5$; y, $F(1,11) = 77.7$; z, $F(1,11) = 46.7$]; a reliable effect of version for the z-axis [$F(1,11) = 15.1$, $p < .01$]; and a version × viewpoint interaction for the z-axis [$F(1,11) = 5.2$, $p < .05$]. Standard and reversed versions were collapsed in all subsequent analyses. Mean response times for the first and last practice blocks are shown in Figure 3, and putative rates of rotation (slopes in degrees/second) are listed in Table 1. Over the course of the 12 practice blocks, overall response times decreased (Figure 3) and rates of rotation, while remaining viewpoint dependent, became faster (Figure 4; error rates ranged from about 9%–34% in Block 1 to about 1%–4% in Block 12).[6] These patterns were found to be statistically reliable in two-way ANOVAs for each axis. There was a reliable effect of block [$p < .001$; x, $F(11,121) = 24.4$; y, $F(11,121) = 26.4$; z, $F(11,121) = 21.7$], a reliable effect of viewpoint [$p < .001$; x, $F(1,11) = 34.0$; y, $F(1,11) = 71.8$; z, $F(1,11) = 47.2$], and a reliable block × viewpoint interaction [$p < .001$; x, $F(11,121) = 8.7$; y, $F(11,121) = 5.1$; z, $F(11,121) = 6.4$]. These interactions confirm that the effect of viewpoint diminished with practice.

Table 1
Putative Rates of Rotation (in Degrees/Second) From Experiments 1–4, Broken Down by Axis of Rotation and Object Handedness

	Slope		
Block No.	*x*-axis	*y*-axis	*z*-axis
Experiment 1: Handedness			
Practice 1	61	82	94
Practice 12	250	244	303
13—familiar (10º, 130º)	167	213	217
13—unfamiliar (160º–340º)	74	85	556
Experiment 2: Recognition			
Practice 1	76	167	125
Practice 12	455	588	417
13—familiar (10º, 130º)	294	769	455
13—unfamiliar (160º–340º)	76	119	169
Experiment 3: Both Versions			
Practice 1			
Standard	159	233	167
Reversed	77	141	119
Practice 12	909	1,429	909
13—familiar (10º, 130º)	588	5,000	714
13—unfamiliar (160º–340º)	149	156	208
Experiment 3: One Version			
Practice 1			
Standard	385	333	253
Reversed	449	839	283
Practice 12	1,039	2,946	841
13—familiar (10º . . . 190º)	2,000	2,000	714
13—unfamiliar (220º–340º)	149	154	167
Experiment 4			
Practice 1	79	233	141
Practice 12	588	909	526
13—familiar (10º, 130º)			
Standard	303	1,667	526
Reversed	256	556	385
13—unfamiliar (160º–340º)			
Standard	108	233	123
Reversed	103	244	286

Note—Familiar, unfamiliar: trials in the surprise blocks in which familiar or unfamiliar viewpoints were presented, respectively.

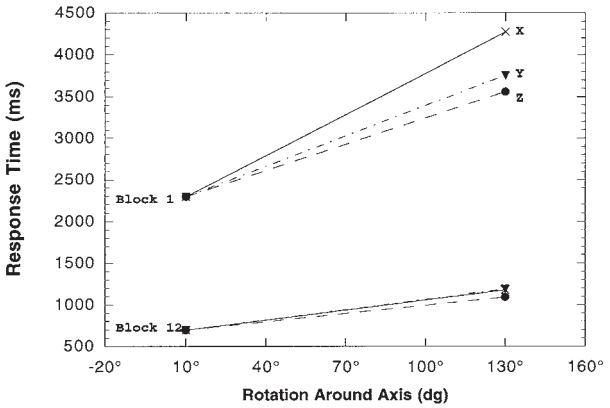

Figure 3. Mean handedness discrimination times collapsed over version for early (Block 1) and late (Block 12) trials in Experiment 1. Each axis of rotation is displayed separately.

Raw slopes for the unfamiliar viewpoints in the surprise block, computed by averaging across viewpoints equidistant from the nearest familiar viewpoint, may underestimate the actual rate of rotation, because the viewpoints between 40º and 100º do not appear to be rotated to the nearest familiar viewpoint (Figure 5). Therefore, here and for all subsequent experiments, Table 1 lists a post hoc estimate of the rate of rotation that was obtained by including only unfamiliar viewpoints from 160º to 340º.

Recognition times and error rates generally increased with the distance from the nearest familiar viewpoint (Figure 5). Slopes for familiar viewpoints were relatively flat, while slopes for unfamiliar viewpoints were comparable to those obtained in Block 1 (Table 1). Here and in all subsequent experiments, no evidence for a

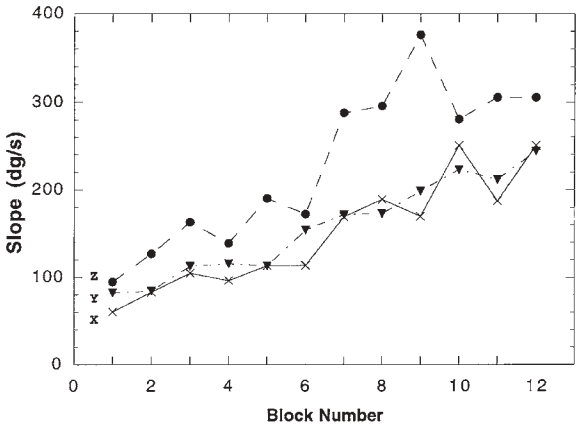

Figure 4. Changes in putative rate of rotation with practice (by blocks of 72 trials) in Experiment 1 (larger values reflect faster rates). Each axis of rotation is displayed separately.

speed–accuracy tradeoff was found in any block. This may be seen in the functions for all three axes of rotation over the range of unfamiliar viewpoints from 160º to 340º. Minima appear near the familiar viewpoints of 10º and 130º. Peaks appear at the unfamiliar viewpoints of 220º for the *x*-axis, 280º for the *y*-axis, and 190º/220º for the *z*-axis.

Discussion

Effects of viewpoint. The results of Experiment 1 are consistent with those of previous studies of handedness discriminations over rotations in depth (Parsons, 1987c; R. N. Shepard & Cooper, 1982). Three predictions were confirmed: (1) response times increased with angular distance from the training viewpoint; (2) patterns of response times were roughly consistent with normalization through the shortest 3-D path to a familiar viewpoint; and (3) the putative rate of rotation varied with the apparent axis of the shortest path rotation. In addition, the results of Experiment 1 are consistent with the findings of Tarr and Pinker (1989): with extensive practice, performance became nearly equivalent at all familiar viewpoints; practice effects at familiar viewpoints *did not* transfer to unfamiliar viewpoints; and judgments at unfamiliar viewpoints appeared to be made by normalization to the nearest familiar viewpoint.

These results provide a baseline measure of normalization for a novel set of 3-D objects similar to those used in previous studies (Jolicoeur, Regehr, Smith, & Smith, 1985; Parsons, 1987c; R. N. Shepard & J. Metzler, 1971). The rates obtained here were of roughly the same order of magnitude as those obtained in other studies (Tables 1 and 2). However, some discrepancies should be pointed out. First, the relative ordering of speed of rotation around axes (slowest to fastest: *x*, *y*, *z*) is the inverse of the ordering obtained by Parsons (slowest to fastest: *z*, *y*, *x*—Parson's estimate of picture-plane rotation rate is comparable to the rate obtained by Jolicoeur et al., 1985). Second, the absolute magnitude of rates is slower than that obtained in S. Shepard and D. Metzler's (1988) one-stimulus condition, in which subjects judged the handedness of a singly presented object rotated around the vertical axis.

There are several possible explanations for these differences. First, there is evidence that for "difficult" discriminations, the putative rate of rotation increases with decreasing complexity (Folk & Luce, 1987). The objects used here may be both more complex and more difficult to discriminate than those used in prior studies, consequently producing slower rates. Second, because S. Shepard and D. Metzler (1988) employed only a single object, their estimated rates may have been idiosyncratically due to its 3-D structure. Third, the presence of a variable number of protruding parts to the sides, front, and back may influence the rate. Depending on the configuration of each object, rotations around different axes would pass through a correspondingly variable number of qualitatively different feature configurations. While an analysis investigating whether such factors influ-

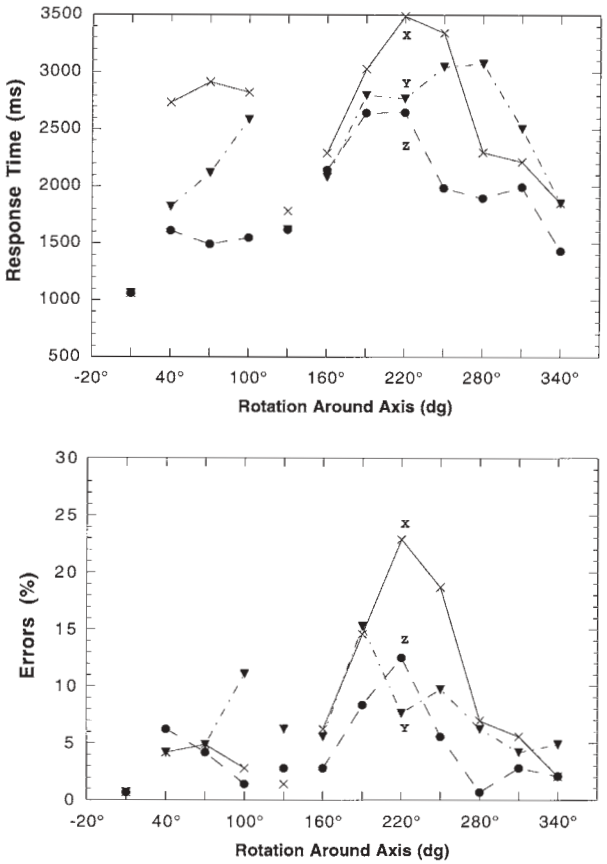

Figure 5. Mean handedness discrimination times and error rates collapsed over version for new, never-before-seen viewpoints (Block 13) in Experiment 1. Each axis of rotation is displayed separately, and familiar viewpoints (0º and 130º) are plotted as separate points.

Table 2
Putative Rates of Rotation (in Degrees/Second) From Mental Rotation Studies With 3-D Stimuli

	Slope		
Study	*x*-axis	*y*-axis	*z*-axis
J. Metzler & R. N. Shepard (1974)			
Experiment 1		64	46
Experiment 2 (mixed)		40	50
Experiment 2 (pure)		38	47
Jolicoeur et al. (1985)			
Experiment 1			38
S. Shepard & D. Metzler (1988)			
One-stimulus condition		343	
Two-stimulus condition		129	
Parsons (1987c)	67	42	35

enced the *pattern* of response times failed to find any reliable effects, it did suggest that the stimuli contained many more singularities and occlusions around the *x*-axis than around either the *y*-axis or the *z*-axis (where there are none). In this and in subsequent experiments employing the same objects, *x*-axis rotations are consistently slower than *y* or *z* rotations. A final contributing factor may be that R. N. Shepard and J. Metzler's (1971) objects did not contain a well-defined top or bottom, possibly necessitating the use of an additional viewpoint-dependent mechanism to locate it prior to normalization. The combination of these two viewpoint-dependent processes would produce cumulatively slower rates, because each process would contribute proportionally greater processing times for increasing changes in viewpoint (Tarr & Pinker, 1991). Depending on the visibility of the marker for the bottom at a given viewpoint, response times may have been differentially affected.

Multiple views. Experiment 1 also provides some support for multiple views. Most importantly, performance at unfamiliar viewpoints varied systematically with the distance from familiar viewpoints. This indicates that, with practice, subjects encode handedness- and viewpoint-specific representations of objects and normalize objects at unfamiliar viewpoints to these views. One alternative, that subjects encode handedness-specific viewpoint-invariant models, can be eliminated, because it predicts that the distance between the unfamiliar viewpoints and the familiar viewpoints should not have influenced performance. That the rates of rotation measured for unfamiliar viewpoints were comparable to those obtained in early practice trials suggests that the same mechanism was used in both instances. The fact that some effect of viewpoint remains even after extensive practice provides further evidence for familiarity-based views. The training viewpoint was observed three times as often as any other practice viewpoint, so it can be expected that objects seen in this viewpoint will be recognized better than objects seen in less frequently occurring viewpoints. This interpretation is also consistent with these residual slopes' providing a measure of canonicality. One factor in determining canonicality is the frequency with which an object is seen from a given viewpoint—here it seems likely that the training viewpoint is treated as canonical.

Is it possible to explain normalization to the nearest familiar viewpoint without appealing to multiple views? Following the rotation-for-handedness hypothesis, objects may be represented independently of both viewpoint and handedness (Biederman & Gerhardstein, 1993; Corballis, 1988; Hinton & Parsons, 1981). Objects would first be recognized without regard for viewpoint *and* handedness. Then, because handedness is defined only in the observer's egocentric frame of reference, handedness would be determined by the additional step of normalizing the object to this frame. However, to account for the viewpoint specificity observed after extensive practice, this model must also assume multiple viewpoint-specific reference frames in which left and right are defined. Although there is some evidence that egocentric frames may be dissociated from the upright for short durations (Jolicoeur, 1990b; McMullen, Hamm, & Jolicoeur, in press), there is no strong evidence that reoriented frames may be encoded in longer term representations (see Hinton & Parsons, 1981; and Robertson, Palmer, & Gomez, 1987). Therefore, the preferred explanation for the results of this experiment is that subjects encoded and used multiple views in judging the handedness of novel 3-D objects rotated in depth.

Deviations from linearity. Here and in subsequent experiments, there are some deviations from the linear prediction of normalization to the nearest familiar viewpoint. In particular, none of the four studies yielded linear patterns for the unfamiliar viewpoints from 40º to 100º. One possible post hoc explanation is that such deviations may be the result of foreshortening, occlusions, singularities, and other changes in the visible features of objects with rotations in depth (Biederman & Gerhardstein, 1993; Humphrey & Jolicoeur, 1993; Koenderink, 1987). To test this possibility, each of the objects was assigned a subjective rating of the degree of foreshortening and occlusion at each viewpoint for rotations in depth. These ratings were then used as predictors along with distance from the nearest familiar viewpoint and distance from the training viewpoint in multiple regression analyses on mean response times from Block 13 for each object. Analyses revealed that the ratings did not account for a reliable amount of the variance. Thus, the hypothesis that the variation in response times in Block 13 is due to changes in visible features is not supported. In contrast, analyses on mean response times from Block 13 with distance from the training viewpoint and distance from the nearest familiar viewpoint as predictors (the correlation between these predictors is .27) confirmed that distance from the nearest familiar viewpoint accounted for a reliable amount of the variance [$p < .05$; *x*, $F(1,9) = 9.9$; *y*, $F(1,9) = 31.3$; *z*, $F(1,9) = 5.3$]. For the *z*-axis, the distance from the training viewpoint also accounted for a reliable amount of the variance [$F(1,9) = 14.7$, $p < .01$], while not being reliable for either the *x*- or the *y*-axis. The latter result is consistent with Tarr and Pinker's (1989) finding that there are

instances in which normalization is to the upright despite the presence of a nearer familiar viewpoint. The upright is canonical and may "attract" misoriented objects to a greater degree than other familiar viewpoints (see also Robertson et al., 1987).

Another possible explanation for the nonlinear patterns of performance observed between 40º and 100º is that the normalization procedure used to judge handedness is based on 2-D view interpolation (Bülthoff & Edelman, 1992; Ullman & Basri, 1991) rather than alignment (Ullman, 1989). In a test of these two models, Bülthoff and Edelman found that recognition performance for unfamiliar views located between two familiar views (separated by 75º) was not linear as predicted by alignment models of normalization. In contrast, for unfamiliar views extrapolated past a familiar view, they found significant viewpoint dependency. These two conditions are analogous to the unfamiliar views used during the surprise phase of this experiment: the viewpoints between 40º and 100º fall between relatively closely spaced familiar views (separated by 120º), while the other viewpoints fall past the familiar views (separated by 240º). A view interpolation model would predict little effect of viewpoint for the former, but large effects of viewpoint for the latter—precisely the pattern found here. Indeed, given the replication of this pattern in each of the experiments presented in this paper, a 2-D view interpolation approach to normalization is the preferred explanation for how unfamiliar viewpoints are matched to multiple-views representations regardless of whether the task is handedness discrimination or recognition.

EXPERIMENT 2

What role does normalization play in recognition? Experiment 1 demonstrates only that viewpoint-dependent processes are used for handedness discriminations, not recognition judgments. Furthermore, although Tarr and Pinker (1989) demonstrated viewpoint dependence in the recognition of 2-D shapes, it is possible that such effects are restricted to the picture plane. Therefore, in Experiment 2, 3-D objects from Experiment 1 were used in a naming task. There are three crucial phases: when objects are unfamiliar; when objects are familiar and in familiar views; and when objects are familiar but in unfamiliar views. Multiple-views theory predicts that there initially will be an effect of viewpoint, that this effect will diminish with practice, and that it will then return for familiar objects in unfamiliar views. In contrast, viewpoint-invariant theories predict that practice will lead to equivalent performance at all familiar viewpoints and to no effect of viewpoint for familiar objects in unfamiliar viewpoints (or at least no effect over relatively large rotations in depth). Thus, the surprise phase of Experiment 2 distinguished between viewpoint-dependent and viewpoint-invariant theories of recognition.

Method

Subjects. Twelve students from the Boston area participated for pay.

Materials. All materials were identical to those in Experiment 1.

Procedure. The training procedure was the same as in Experiment 1 except that subjects were never shown the reversed versions. The four objects not used in the named set were presented as distractors at the same viewpoints as were the named objects (the distractors were not presented during training). These objects were included to minimize the possibility that subjects would find some feature that uniquely identified a named object.

The subjects responded via a three-key response board labeled "Kip," "Kef," and "Kor." They could use either hand or both hands to respond and were instructed to identify each object as quickly as possible regardless of viewpoint. In addition, the subjects were informed that they were to press a footpedal when objects that they had not been taught were displayed.

Design. Practice blocks consisted of 6 preliminary trials, followed by 96 trials composed of named and distractor objects in their standard versions at the training viewpoint and at rotations of 130º around the *x*-, *y*-, or *z*-axis (Figure 2). The trials corresponded to the named objects in the training viewpoint 12 times and 4 times in the other viewpoints and to the distractor objects in the training viewpoint 3 times and once in the other viewpoints. The surprise block consisted of 8 preliminary trials, followed by 576 trials. The named objects appeared in the training viewpoint 12 times and 4 times in each of the other 33 viewpoints (30º increments around the *x*-, *y*-, or *z*-axis). The distractor objects appeared in the training viewpoint 3 times and once in the other viewpoints. There was a self-timed break every 53 trials. As in Experiment 1, there were four sessions.

Results

Response time means were calculated as in Experiment 1. In Block 13, no reliable effect of either stimulus group (A, B, or C) or individual objects was found, indicating that both factors can be disregarded in further analyses. Putative rates of rotation for the first and last practice blocks are listed in Table 1. Over the course of the 12 practice blocks, overall response times decreased and rates of rotation, while remaining viewpoint dependent, became faster (error rates ranged from about 4%–31% in Block 1 to about 1%–3% in Block 12). These patterns were found to be statistically reliable for each axis. There was a reliable effect of block [$p < .001$; x, $F(11,121) = 54.1$; y, $F(11,121) = 37.1$; z, $F(11,121) = 33.0$], a reliable effect of viewpoint [$p < .001$; x, $F(1,11) = 72.1$; y, $F(1,11) = 16.9$; z, $F(1,11) = 30.3$], and a reliable block × viewpoint interaction [$p < .001$; x, $F(11,121) = 20.8$; y, $F(11,121) = 8.8$; z, $F(11,121) = 5.4$]. These interactions confirm that the effect of viewpoint diminished with practice.

In the surprise block, the slopes for familiar viewpoints remained relatively flat, while the rates of rotation for unfamiliar viewpoints slowed to levels comparable to those obtained in Block 1 (Table 1). Figure 6 shows that the unfamiliar viewpoints between 40º and 100º do not appear to be rotated to the nearest familiar viewpoint, but rather exhibit patterns similar to the curves found in Experiment 1 over the same range of viewpoints.

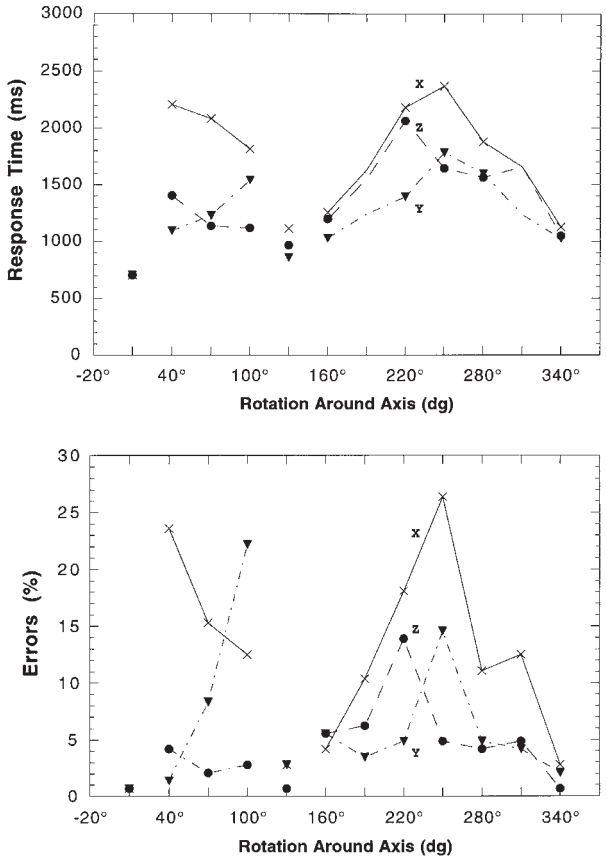

Figure 6. Mean recognition times and error rates for new, never-before-seen viewpoints (Block 13) in Experiment 2. Each axis of rotation is displayed separately, and familiar viewpoints (0º and 130º) are plotted as separate points.

The patterns of response times and error rates in the surprise block separated by the axes of rotation are shown in Figure 6. No evidence for a speed–accuracy tradeoff in recognition was found. As in the surprise block of Experiment 1, performance was related to the distance from the nearest familiar viewpoint. This may be seen for all three axes of rotation over the range of unfamiliar viewpoints from 160º to 340º. Minima appear at or near the familiar viewpoints of 10º and 130º. Peaks appear at the unfamiliar viewpoints of 250º for the *x*-axis, 250º for the *y*-axis, and 220º for the *z*-axis. Only the peak for *z*-axis rotations deviates from the "ideal" peak of 250º, the viewpoint farthest from a familiar viewpoint. Deviations from linearity fall at roughly the same viewpoints in Experiments 1 and 2. This was confirmed by multiple regressions on mean response times from Block 13 of Experiment 2, with the mean response times from Block 13 of Experiment 1 and distance from the nearest familiar viewpoint as predictors. These analyses revealed that the variation in response times in Experiment 1, beyond that correlated with distance from the nearest familiar viewpoint, accounted for a reliable amount of the variance in response times in Experiment 2 [$p < .05$; *x*, $F(1,9) = 11.9$; *y*, $F(1,9) = 7.9$; *z*, $F(1,9) = 11.4$]. Variation in the distance from the nearest familiar viewpoint uncorrelated with response times from Experiment 1 accounted for a reliable amount of the variance in Experiment 2 for the *z*-axis [$F(1,9) = 10.4$, $p < .05$], while not being reliable for either the *x*- or *y*-axis. Three multiple regression analyses on mean response times from Block 13 confirmed that the distance from the nearest familiar viewpoint accounted for a reliable amount of the variance in response times [$p < .01$; *x*, $F(1,9) = 12.0$; *y*, $F(1,9) = 23.1$; *z*, $F(1,9) = 16.2$]; the distance from the training viewpoint was not reliable for any axis.

Discussion

The results of Experiment 2 replicate the results of Experiment 1 for recognition rather than handedness judgments. The major findings are: (1) response times increased with the angular distance from the training viewpoint; (2) with practice, performance became nearly equivalent at all familiar viewpoints; and (3) at unfamiliar viewpoints, response times increased with the angular distance from the nearest familiar viewpoint. As shown in Table 1, the effect of viewpoint at unfamiliar viewpoints is comparable to that found in early trials of Experiment 2, to the effects of viewpoint obtained in Experiment 1, and to the effects reported in previous studies of normalization (J. Metzler & R. N. Shepard, 1974; Parsons, 1987c; S. Shepard & D. Metzler, 1988). These findings indicate that normalization mechanisms are responsible for effects of viewpoint not only in handedness discriminations, but in the recognition of 3-D objects rotated in depth. Supporting this argument, differences between the rates of rotation for each axis were similar in Experiments 1 and 2, as were deviations in linearity for each axis. Indeed, variations in response times from Experiment 1 (Figure 5) can predict the pattern found in Experiment 2. This provides strong evidence that the same mechanism was used in both tasks. Moreover, the deviations in linearity from 40º to 100º are again consistent with the predictions of a view-interpolation model (Bülthoff & Edelman, 1992). Thus, the most plausible explanation for viewpoint dependency in recognition is that unfamiliar viewpoints between familiar views are matched to these views by interpolation and that unfamiliar viewpoints beyond familiar views are matched by normalization through the shortest 3-D path.

Experiment 2 also provides evidence for multiple views. The initial effect of viewpoint indicates the presence of viewpoint-specific representations at the training viewpoint. The nearly equivalent performance with practice indicates that either multiple views or viewpoint-invariant representations have been acquired. In the surprise block, the systematic effect of distance from the nearest familiar view on recognition at unfamiliar views indicates the presence of multiple viewpoint-specific representations at each familiar viewpoint. This

degree of viewpoint specificity is inconsistent with both complete and restricted viewpoint-invariant theories, but consistent with and predicted by multiple-views theories.

Finally, the results of Experiment 2 rule out three alternative explanations for effects of viewpoint in recognition. First, the rotation-for-handedness hypothesis is not supported. In this experiment, handedness was irrelevant (since only one version of each object was used), yet response times still exhibited viewpoint dependency. Second, the "rotation-to-upright" hypothesis, that normalization procedures are used only under "ecologically unusual" conditions where top and bottom have been perturbed (see Rock et al., 1981), is not supported. Effects of viewpoint were obtained for rotations around the vertical axis where the positions of top and bottom with respect to the upright were preserved. Third, the hypothesis that viewpoint-dependent effects are due to explicit familiarity is not supported (Biederman & Cooper, 1991; Biederman & Gerhardstein, 1993). Viewpoint-dependent effects were obtained in a naming task and with response time as a dependent measure, both of which are generally considered to be characteristics of implicit memory tasks (see, e.g., Roediger et al., 1989). Taken together, these findings indicate that viewpoint-dependent recognition mechanisms are not a "special case" restricted to uncommon tasks, 2-D, or unusual viewpoints, but rather extend to the recognition of 3-D objects rotated in depth.

EXPERIMENT 3

One alternative not addressed by Experiment 2 is that recognition was viewpoint invariant, but that subjects surreptitiously normalized the objects to a frame of reference where handedness was defined. According to this version of the rotation-for-handedness hypothesis, representations are invariant with respect to viewpoint *and* handedness (Corballis, 1988; Corballis et al., 1978; Hinton & Parsons, 1981) and normalization procedures are used on the chance that the version of the observed object will not match the version represented in memory. Presumably this would occur because subjects are conservative: despite never having seen a reversed version, the possibility of such a version appearing and matching incorrectly prompted them to verify handedness.

Tarr and Pinker (1989) addressed this possibility by making handedness explicitly irrelevant, informing subjects of this fact and giving them practice in treating mirror pairs equivalently. Specifically, both handedness versions of an object were given the same name and classified as a single item in both training and recognition (this is analogous to learning the label "glove" for both right-handed and left-handed gloves). Because handedness is irrelevant to the task, normalization solely to verify handedness should no longer be necessary. In contrast, if normalization is used for recognition, viewpoint-dependent effects should be obtained. Across several experiments in which handedness was irrelevant, viewpoint-dependent effects were still found for the recognition of 2-D shapes. Such findings further disconfirm the rotation-for-handedness hypothesis. Experiment 3 was an attempt to replicate these results by making handedness explicitly irrelevant for the recognition of 3-D objects.

Tarr and Pinker (1989) also observed an effect of training that provides further support for view-specific object representations. In experiments utilizing both handedness versions of a shape, subjects were informed that the same name referred to both versions and initially subjects were shown both versions. During training, however, the subjects learned only the standard, and, as a consequence, presumably they encoded only this version. Only during initial practice trials were the reversed versions of these shapes seen extensively. Surprisingly, equivalent performance regardless of picture-plane orientation was obtained for the unfamiliar reversed versions. Although such a pattern is apparently viewpoint-invariant, Tarr and Pinker proposed that subjects were normalizing or "flipping" reversed shapes *in depth* around an axis lying within the picture plane. Such a rotation is the shortest path for aligning a mirror-reversed 2-D shape in *any* picture-plane viewpoint with its standard counterpart (see Parsons, 1987a, 1987b). The magnitude of this rotation will always be 180º, producing equivalent performance for all orientations. In contrast, a control experiment in which subjects were trained on *both* standard and reversed versions produced performance that varied with distance from the upright for both versions. Thus, when subjects were given the opportunity to learn both versions, they used a shorter picture-plane path rather than a flip in depth to recognize the reversed versions of the objects. Such effects are unlikely to occur as a result of determining handedness, since, regardless of training, there is always a shorter picture-plane rotation that will align the reversed shape with the upright. However, if reversed versions are recognized via normalization to view-specific representations, performance will vary, depending on whether these shapes are matched to standard versions or to reversed versions learned at the upright.

Experiment 3 permits a further test of this surprising result. Unlike with 2-D shapes, there is no rigid 3-D transformation that will align a 3-D object and its enantiomorph. At least two transformations do exist: a nonrigid deformation *pulling the object through itself*; and a rigid rotation in 4-D space. In the former case, a nonrigid deformation makes handedness, not viewpoint, congruent; therefore the practice/surprise manipulation is still diagnostic for recognition over rotations in depth. In the latter case, it seems unlikely that rigid 4-D transformations are computable within the human visual system (R. N. Shepard, 1984). Therefore, if equivalent performance is obtained for unfamiliar reversed versions of 3-D objects, it would suggest, contrary to the "flip in depth" hypothesis, that viewpoint-invariant mechanisms are used for recognition. In contrast, multiple-views pre-

dicts that unfamiliar reversed versions of 3-D objects will *not* be recognized in constant time. Instead, reversed versions might be aligned with familiar standard versions by normalizing them to a familiar viewpoint and then comparing noncongruent parts to see whether they appear in exactly opposite positions. Such a strategy would result in performance that varied with distance from the training viewpoint for both standard and reversed versions. In Experiment 3, these alternatives were tested by manipulating whether subjects were trained on both versions or only on standard versions.

Method

Subjects. Twenty-four students from the Boston area participated for pay (12 in each condition).

Materials. All materials were identical to those in Experiment 1.

Procedure. The training procedure was the same as in Experiment 2 except that subjects were shown the reversed versions of the objects as well as the standard versions and were told that the name of an object referred to both versions.

Two conditions were employed. In the both-versions condition, subjects were trained on the standard and reversed versions of each object, including duplicating and then building both versions from memory. In the one-version condition, subjects were shown both versions of each object, but were trained only on their standard versions.

Design. Trials in the both-versions condition were organized into practice and surprise blocks similar to those used in Experiment 2. However, trials for the three target objects were divided equally between standard and reversed versions. Trials for the distractor objects were similarly divided, except that to preserve the 3:1 ratio of named objects to distractors, in the practice blocks each distractor was presented in one version at the training viewpoint and the other version at the 130º viewpoints. In the surprise block, each distractor was presented in one version at 60º increments beginning at 10º (10º, 70º, …) and the other version at 60º increments beginning at 40º (40º, 100º, …). Which version was presented at even or odd viewpoints was also varied by the axis of rotation. Target objects were presented in both versions at all viewpoints. Subjects were given a self-timed break every 53 trials. The subjects were run in a total of four sessions as in Experiment 1.

In the both-versions condition, the practice (10º and 130º) and surprise viewpoints were identical to those used in Experiment 2. In the one-version condition, the objects were displayed in 13 new practice viewpoints (the training viewpoint and 40º, 70º, 130º, and 190º around each axis) and in the same 34 surprise viewpoints used in Experiment 2. New practice viewpoints were introduced primarily to investigate the effects of a decreased range between practice viewpoints and were not part of the manipulation of training between conditions.

Trials in the one-version condition were organized into practice blocks consisting of 6 preliminary trials, followed by 240 experimental trials. The three target objects appeared at the training viewpoint six times in the standard version and six times in the reversed version. At the other 12 viewpoints, the target objects appeared two times each in both versions. The remaining 60 trials were composed of distractor objects. To preserve the ratio of targets to distractors, each distractor was presented in one version at select practice viewpoints, and the other version, at the remaining practice viewpoints. Surprise blocks were identical to those used in the both-versions condition. The subjects were given a self-timed break every 62 trials. The subjects were run in a total of four sessions identical to those in Experiment 1.

Results

Both-versions condition. Response time means were calculated as in Experiment 1. Putative rates of rotation for the first and last practice blocks are listed in Table 1. In the first block, there were no reliable main effects of handedness version, and only a reliable interaction between viewpoint and version for the *x*-axis [$F(1,11) = 5.0$, $p < .05$]. In this instance the rates of rotation for both versions were still relatively slow, with a slope for standard versions of 159º/sec and for reversed versions of 77º/sec. Collapsing over Blocks 1–12 reveals a reliable effect of handedness version for the *y*-axis [$F(1,11) = 6.8$, $p < .05$], as well as a version × viewpoint interaction [$F(1,11) = 5.2$, $p < .05$]. No reliable effects were found for version for either the *x*- or *z*-axis, so the factor of version was omitted in all subsequent analyses. Over the course of the 12 practice blocks, overall response times decreased and rates of rotation, while remaining viewpoint dependent, became faster (Table 1; error rates ranged from about 5%–23% in Block 1 to about 1%–5% in Block 12). These patterns were found to be statistically reliable for each axis. There was a reliable effect of block [$p < .001$; *x*, $F(11,121) = 44.1$; *y*, $F(11,121) = 39.8$; *z*, $F(11,121) = 36.1$], a reliable effect of viewpoint [$p < .001$; *x*, $F(1,11) = 46.7$; *y*, $F(1,11) = 32.4$; *z*, $F(1,11) = 85.7$], and a reliable block × viewpoint interaction [$p < .001$; *x*, $F(11,121) = 16.5$; *y*, $F(11,121) = 4.9$; *z*, $F(11,121) = 19.1$]. These interactions confirm that the effect of viewpoint diminished with practice.

In the surprise block, the slopes for familiar viewpoints collapsed over version remained relatively flat, while the rates of rotation for unfamiliar viewpoints slowed to levels comparable to those obtained in Block 1 (see Table 1). The patterns of response times and error rates in the surprise block are shown in Figure 7. No evidence for a speed–accuracy tradeoff in recognition was found in any block of either condition. As in the surprise blocks of the previous experiments, recognition times generally increased with the distance from the nearest familiar viewpoint. For the both-versions condition, this may be seen in the response time curves over the range of unfamiliar viewpoints from 160º to 340º. Minima appear at or near the familiar viewpoints of 10º and 130º. Peaks appear at the unfamiliar viewpoints of 250º/280º for the *x*-axis, 250º for the *y*-axis, and 220º for the *z*-axis. Only the peak for *z*-axis rotations deviates from the "ideal" peak of 250º, the viewpoint farthest from a familiar viewpoint. Unfamiliar viewpoints between 40º and 100º do not appear to be rotated to the nearest familiar viewpoint, but rather exhibit deviations from linearity similar to those obtained in Experiments 1 and 2. This was partly confirmed by three multiple regressions on mean response times from Block 13 of the both-versions condition with the mean response times from Block 13 of Experiment 1 and distance from the nearest familiar viewpoint as predictors. These analyses revealed that the variation in response times in Experi-

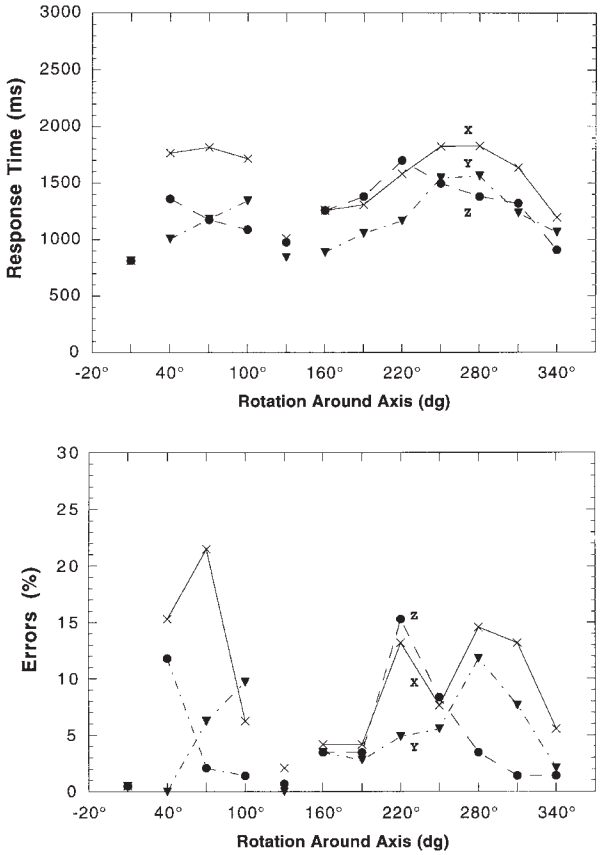

Figure 7. Mean recognition times and error rates collapsed over version for new, never-before-seen viewpoints (Block 13) in the both-versions condition of Experiment 3. Each axis of rotation is displayed separately, and familiar viewpoints (0º and 130º) are plotted as separate points.

ment 1, beyond that correlated with distance from the nearest familiar viewpoint, accounted for a reliable amount of the variance for the *z*-axis [$F(1,9) = 10.6, p < .01$], while not being reliable for the *x*- or *y*-axes. Variation in the distance from the nearest familiar viewpoint uncorrelated with response times from Experiment 1 also accounted for a reliable amount of the variance in the both-versions condition for the *z*-axis [$F(1,9) = 9.8$, $p < .05$], while not being reliable for the *x*- or *y*-axes. These findings suggest that the response time functions display highly similar variations across viewpoint. In addition, three multiple regression analyses on mean response times from Block 13 with distance from the nearest familiar viewpoint and distance from the training viewpoint as predictors confirmed that the distance from the nearest familiar viewpoint accounted for a reliable amount of the variance in response times [$p < .01$; *x*, $F(1,9) = 10.4$; *y*, $F(1,9) = 22.6$; *z*, $F(1,9) = 17.5$]; the distance from the training viewpoint was not a reliable predictor for any axis of rotation.

One-version condition. The effects reported for the both-versions condition were replicated in the one-version condition. Putative rates of rotation for the first and last practice blocks are listed in Table 1. In the first block there were no reliable main effects of handedness version or reliable interactions between viewpoint and version for any axis. Collapsing over Blocks 1–12 reveals a reliable effect of handedness version for the *y*-axis [$F(1,11) = 11.4, p < .01$] and a version × viewpoint interaction for the *y*-axis [$F(4,44) = 4.1, p < .01$] and the *x*-axis [$F(4,44) = 4.0, p < .01$]. Thus, viewpoint dependency varied from one handedness version to the other. However, these differences were unsystematic and do not appear to reflect the use of different mechanisms in the recognition of different versions. In particular, response times for both handedness versions generally increased with distance from the training viewpoint. Version was omitted from all subsequent analyses. Over the 12 practice blocks, overall response times decreased and rates of rotation, while remaining viewpoint dependent, became faster (Table 1). These patterns were found to be statistically reliable for each axis. There was a reliable effect of block [$p < .001$; *x*, $F(11,121) = 53.2$; *y*, $F(11,121) = 45.9$; *z*, $F(11,121) = 46.8$], a reliable effect of viewpoint [$p < .01$; *x*, $F(4,44) = 16.9$; *y*, $F(4,44) = 12.6$; *z*, $F(4,44) = 13.6$], and a reliable block × view-

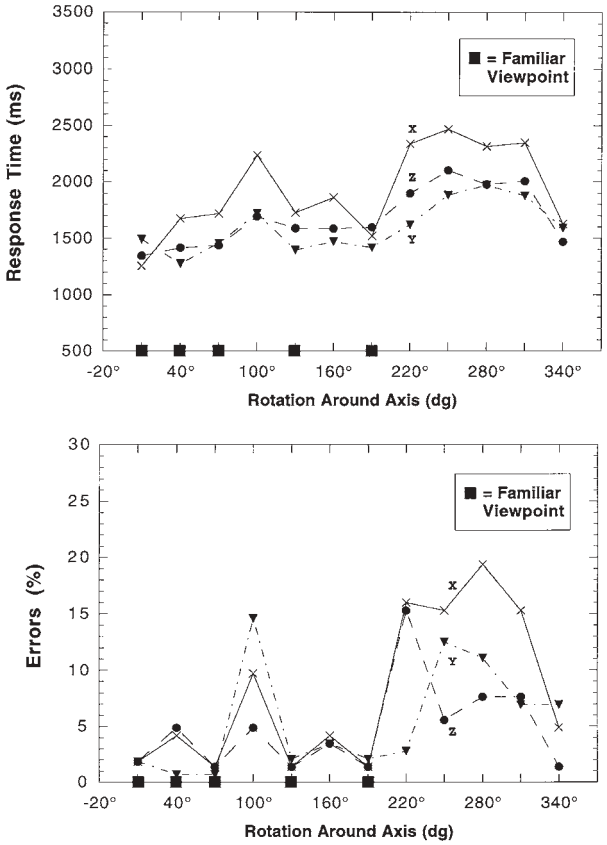

Figure 8. Mean recognition times and error rates collapsed over version for new, never-before-seen viewpoints (Block 13) in the one-version condition of Experiment 3. Each axis of rotation is displayed separately, and familiar viewpoints are marked by the squares along the horizontal axis.

point interaction [$p < .001$; x, $F(44,484) = 4.1$; y, $F(44,484) = 2.2$; z, $F(44,484) = 2.3$]. These interactions confirm that the effect of viewpoint diminished with practice.

In the surprise block, the slopes for familiar viewpoints collapsed over version remained relatively flat, while the rates of rotation for unfamiliar viewpoints slowed to levels comparable to those obtained in Block 1 (Table 1). Recognition times and error rates generally increased with the distance from the nearest familiar viewpoint, and, particularly, over the range of unfamiliar viewpoints from 220º to 340º (Figure 8). Multiple regression analyses on mean response times from Block 13 with distance from the nearest familiar viewpoint and distance from the training viewpoint as predictors (the correlation between these predictors is −.02) confirmed that the distance from the nearest familiar viewpoint accounted for a reliable amount of the variance in response times [$p < .001$, for each axis; x, $F(1,9) = 20.1$; y, $F(1,9) = 75.2$; z, $F(1,9) = 30.9$]; while the distance from the training viewpoint was not reliable for any axis. Minima are found at the familiar viewpoints of 10º and 190º. Peaks are generally near the ideal midpoint of 280º. As in the other experiments, the unfamiliar viewpoints between closely spaced familiar viewpoints (here at 100º and 160º) do not appear to be normalized to the nearest familiar viewpoint. These deviations from linearity are consistent with the predictions of view-interpolation recognition mechanisms (Bülthoff & Edelman, 1992), whereby effects of viewpoint are obtained most clearly when unfamiliar viewpoints fall beyond familiar viewpoints, rather than between them.

Discussion

The inclusion of both standard and reversed versions of an object, both assigned the same name, did little to alter subjects' behavior from earlier experiments. In particular, the fact that viewpoint-dependent effects were obtained in a task where handedness was explicitly irrelevant indicates that the rotation-for-handedness hypothesis cannot account for viewpoint-dependent recognition performance. The results of both conditions are consistent with multiple views and replicate the basic pattern of results found in Experiments 1 and 2. First, performance in initial practice trials was dependent on the distance from the training viewpoint. Second, these effects diminished with practice until performance was approximately equivalent at all familiar viewpoints. Third, the effects of viewpoint returned for unfamiliar viewpoints, but now with performance dependent on the distance from the nearest familiar viewpoint. Once again, the latter effect exhibited a pattern similar to the surprise block of Experiment 1, where it is uncontroversial that normalization procedures were used.

The results of the second manipulation unique to Experiment 3, the differential training of handedness versions in the both-versions and the one-version conditions, also support the existence of multiple views. In contrast to the results of earlier 2-D studies, during initial recognition both familiar and unfamiliar reversed versions of 3-D objects were normalized to the training viewpoint. This finding rules out the use of viewpoint-invariant mechanisms in Tarr and Pinker's study, since such a mechanism would be equally effective for unfamiliar 2-D or 3-D reversed versions of objects. Rather, normalization procedures appear to be used to align 2-D shapes by a flip in depth to familiar versions and to align 3-D objects by a 3-D rotation to the familiar training viewpoint; both are shortest path transformations.[7]

EXPERIMENT 4

The training manipulation used in Experiment 3 is predicated on the assumption that unfamiliar viewpoints seen during initial practice are normalized to views learned during training. However, it is possible that viewpoint-invariant representations only develop over experience (Jolicoeur, 1985, 1990a). In Experiment 4, this possibility was tested by investigating whether unfamiliar versions of highly familiar objects would be recognized through viewpoint-invariant mechanisms. In contrast to the procedure in Experiment 3, reversed versions were withheld throughout both training and practice. In a similar manipulation, using 2-D shapes, Tarr and Pinker (1989) found equivalent performance for the reversed versions. This was interpreted as further evidence for the use of a 180º flip in depth. Here the use of 3-D objects is predicted to result in performance related to the nearest familiar viewpoint. In contrast, both complete viewpoint-invariant and restricted viewpoint-invariant theories predict that performance will be invariant over handedness as well as viewpoint (see, e.g., Biederman & Cooper, 1991; Cooper et al., 1992).

Method

Subjects. Twelve students from the Boston area participated for pay.

Materials. All materials were identical to those in Experiment 1.

Procedure. The training procedure was the same as in Experiment 2. The subjects were never shown the reversed versions of the objects.

Design. The practice blocks were identical to those in Experiment 2; no reversed versions were presented. The surprise blocks were identical to those in the both-versions condition of Experiment 3; the trials for both named objects and distractors were divided equally between standard and reversed versions. The subjects were given a self-timed break every 53 trials. The subjects were run in four sessions, as in Experiment 1.

Results

Response time means were calculated as in Experiment 1. Putative rates of rotation for the first and last practice blocks are listed in Table 1. Over the course of the 12 practice blocks, overall response times decreased, and rates of rotation, while remaining viewpoint dependent, became faster (error rates ranged from about 11%–45% in Block 1 to about 0%–2% in Block 12).

These patterns were found to be statistically reliable for each axis. There was a reliable effect of block [$p < .001$; x, $F(11,121) = 29.1$; y, $F(11,121) = 22.8$; z, $F(11,121) = 23.5$], a reliable effect of viewpoint [$p < .01$; x, $F(1,11) = 33.4$; y, $F(1,11) = 25.0$; z, $F(1,11) = 27.5$], and a reliable block × viewpoint interaction [$p < .001$; x, $F(11,121) = 6.7$; y, $F(11,121) = 4.9$; z, $F(11,121) = 4.9$]. These interactions confirm that the effect of viewpoint diminished with practice.

In the surprise block, the slopes for both the standard and the reversed versions in familiar viewpoints remained relatively flat, while the rates of rotation for unfamiliar viewpoints slowed to levels comparable to those obtained in Block 1 (Table 1). The patterns of response times and error rates in the surprise block are shown in Figure 9. No evidence for a speed–accuracy tradeoff in recognition was found in any block. As in the surprise block of previous experiments, recognition times generally increased with the distance from the nearest familiar viewpoint. This may be seen in the response time curves over the range of unfamiliar viewpoints from 160º to 340º. With the exception of y-axis rotations for the reversed versions where the minima are 10º and 190º, minima appear near the familiar viewpoints of 10º and 130º. For the standard versions, peaks appear at the unfamiliar viewpoint of 250º for all three axes of rotation; for the reversed versions, peaks appear at 250º for the x-axis, 280º for the y-axis, and 220º for the z-axis. However, unfamiliar viewpoints between 40º and 100º do not appear to be rotated to the nearest familiar viewpoint, but rather exhibit deviations from linearity similar to those obtained in Experiments 1–3. This was confirmed by three multiple regressions on mean response times from Block 13 of Experiment 4 (collapsed across standard and reversed versions) with the mean response times from Block 13 of Experiment 1 and distance from the nearest familiar viewpoint as predictors. These analyses revealed that the variation in response times in Experiment 1, beyond that correlated with distance from the nearest familiar viewpoint, accounted for a reliable amount of the variance for the y-axis [$F(1,9) = 11.9$, $p < .01$] and the z-axis [$F(1,9) = 7.3$, $p < .05$], while not being reliable for the x-axis. Variation in the distance from the nearest familiar viewpoint uncorrelated with response times from Experiment 1 accounted for a reliable amount of the variance for the z-axis [$F(1,9) = 13.2$, $p < .01$], while not being reliable for the x- or the y-axis. These findings suggest that the response time

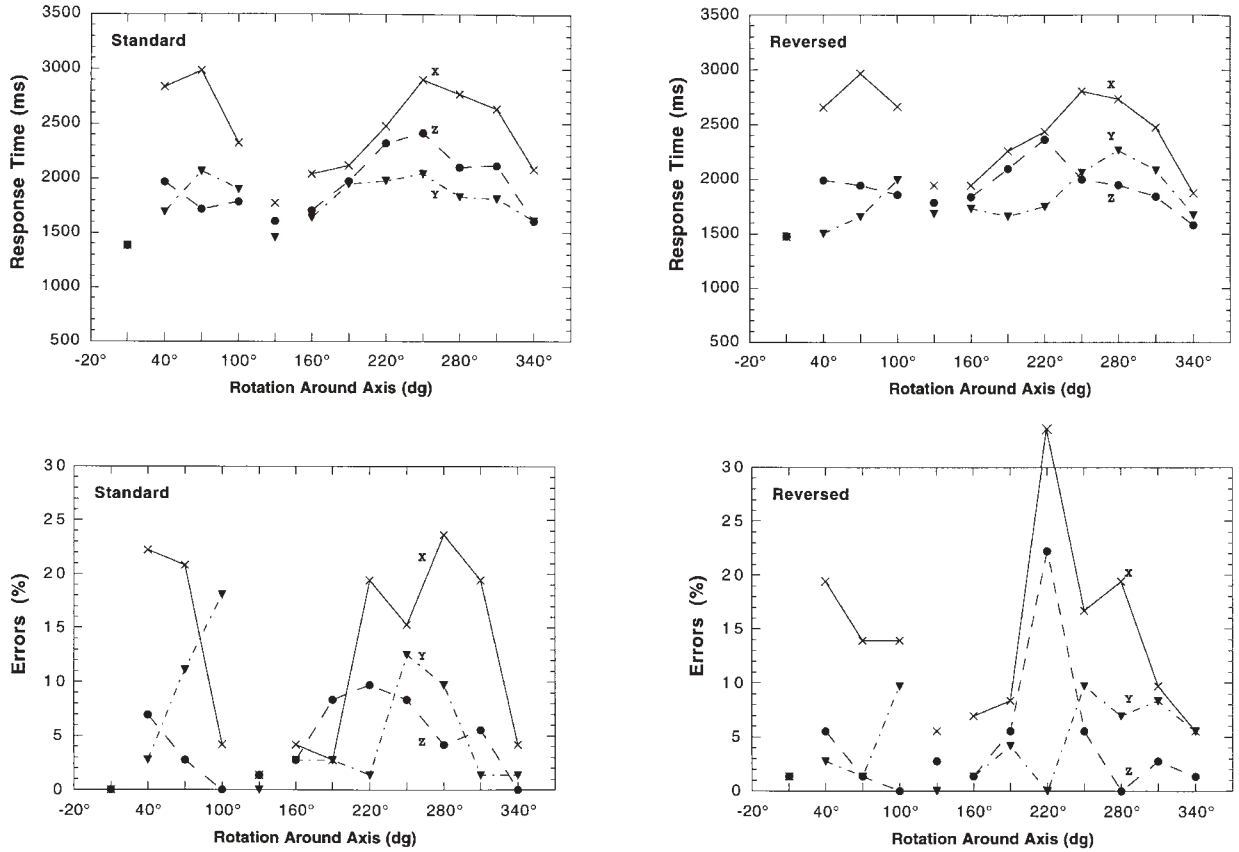

Figure 9. Mean recognition times and error rates, broken down by standard and reversed handedness versions for new, never-before-seen viewpoints (Block 13) in Experiment 4. Each axis of rotation is displayed separately, and familiar viewpoints (0º and 130º) are plotted as separate points.

functions display similar variations across viewpoints. In addition, six multiple regression analyses on mean response times from Block 13 with distance from the nearest familiar viewpoint and distance from the training viewpoint as predictors confirmed that for each version around all three axes of rotation the distance from the nearest familiar viewpoint accounted for a reliable amount of the variance in response times for the standard versions [$p < .01$; x, $F(1,9) = 12.4$; y, $F(1,9) = 16.1$; z, $F(1,9) = 33.8$] and for the reversed versions [$p < .05$; x, $F(1,9) = 7.8$; y, $F(1,9) = 5.5$; z, $F(1,9) = 7.4$]. For reversed versions rotated around the z-axis, the distance from the training viewpoint also accounted for a reliable amount of the variance [$F(1,9) = 5.4, p < .05$], while the distance from the training viewpoint was not a reliable predictor for either version for the x- or the y-axis. This finding is consistent with the results of Experiment 1, in which z-axis rotations exhibited two components: rotation to the training viewpoint and rotation to the nearest familiar viewpoint. Finally, a comparison of the slopes for the standard and reversed versions in the surprise block (Table 1), the crucial manipulation in this experiment, reveals little difference for either the x- or the y-axis; however there is a large difference for the z-axis, where the standard versions exhibited a rate of rotation over twice as slow as that for the reversed versions.

Discussion

The same basic pattern of results found in Experiments 1–3 was found in Experiment 4. First, in initial practice trials, performance was related to the distance from the training viewpoint. Second, these effects diminished with practice at all familiar viewpoints. Third, performance was related to the distance from the nearest familiar viewpoint when the same objects appeared in unfamiliar viewpoints. Moreover, the similarity between the patterns of performance observed here and in Block 13 of Experiment 1 suggest that the same normalization procedures were used in both tasks.

Crucially, the introduction of reversed versions of the objects in Block 13 produced few changes in response time patterns; as is shown in Figure 9, for the x- and y-axes of rotation, it appears that reversed versions were normalized to the nearest familiar viewpoint. There is one discrepancy in these results: reversed versions rotated around the z-axis exhibited a faster rate of rotation than that found for standard versions. A possible explanation for this finding is that subjects sometimes may have ignored the fronts and backs of the objects, treating them as flat shapes, and used a 180º flip in depth to align the reversed versions with the standard versions. An example of this is the aligning of one's left and right hands by holding them out in front and then rotating the left hand 180º around the vertical axis: the fronts and backs are different, but the 2-D contours match. In Experiment 4, this strategy could have been most readily used for picture-plane misorientations, because such rotations preserve the visible 2-D image structure projected by each object; in contrast, rotations in depth alter the image structure. Several experiments by Biederman and Gerhardstein (1993) support this interpretation: they found that performance was roughly equivalent when rotations in depth resulted in a mirror reversal of outline shape, but that performance was degraded when rotations in depth resulted in a change in outline shape.

Overall, these findings are in agreement with the one-version condition of Experiment 3, in which unfamiliar *untrained* reversed versions were apparently normalized to the familiar training viewpoint. In Experiment 4, unfamiliar *unpracticed* reversed versions were apparently normalized, but this time to the nearest familiar viewpoint. This finding provides further support for the multiple-views hypothesis, in particular implicating normalization mechanisms in the recognition of unfamiliar reversed versions of 3-D objects. Moreover, these results indicate that normalization mechanisms were also responsible for the equivalent performance observed in Experiments 3 and 4 of Tarr and Pinker (1989).

GENERAL DISCUSSION

The mechanisms used in human object recognition are almost certainly a product of many factors, including the task, the learning and retrieval contexts, and the functional and visual relationships between objects both encoded in memory and observed in the environment (a characterization consistent with current theory in memory research; e.g., Roediger et al., 1989). While the present study has focused on the question of whether discriminating between visually similar 3-D objects is viewpoint dependent or viewpoint invariant, it has implications for a more general understanding of the recognition process. In particular, the manipulations used in this study indicate that viewpoint-dependent mechanisms may not be dismissed as the byproduct of rotation for handedness, explicit familiarity, "contamination" from non–recognition-based processing systems, or experimental confounds such as prior exposure or diagnostic features within a restricted stimulus set. This study also addressed an asymmetry in what can be concluded from viewpoint-invariant, as opposed to viewpoint-dependent, patterns of performance. Viewpoint-invariant patterns may be plausibly interpreted as evidence for viewpoint-invariant object representations or for multiple views. This problem was addressed by familiarizing subjects with objects in a controlled set of viewpoints and presenting the now-familiar objects in unfamiliar viewpoints. The results of this manipulation may be summarized as follows:

1. In initial trials, when objects were first seen rotated in depth, performance was related monotonically to the distance from the trained viewpoint. This pattern of viewpoint dependency was found even when handedness was explicitly irrelevant to the recognition task.

2. With extensive practice at recognizing the objects from several viewpoints, the effect of viewpoint diminished to nearly equivalent performance at each familiar viewpoint.

3. When the now-familiar objects were presented in unfamiliar viewpoints, the diminished effects of viewpoint for familiar views did not transfer to unfamiliar views. Performance was once again viewpoint dependent, yet now related to the distance from the nearest familiar view.

These results support the multiple-views hypothesis (Bülthoff & Edelman, 1992; Tarr & Pinker, 1989). During initial trials, the trained objects are recognized at unfamiliar viewpoints by a normalization to the training viewpoint (the only familiar view). Following practice, viewpoint-specific representations are encoded at each familiar viewpoint and the familiar objects are recognized at familiar viewpoints without normalization. Finally, familiar objects at new unfamiliar viewpoints are recognized by a normalization to the nearest familiar viewpoint. Note especially that several pieces of converging evidence support the thesis that the same normalization procedure is used for recognition throughout all phases of each experiment:

1. Similar putative rates of rotation were obtained whenever familiar objects were recognized at unfamiliar viewpoints (initial trials following training and surprise trials following practice). These rates are comparable to those obtained for handedness discriminations on the same objects (Experiment 1) and to those found in other 3-D normalization studies (J. Metzler & R. N. Shepard, 1974; Parsons, 1987c; S. Shepard & D. Metzler, 1988; see Tables 1 and 2).

2. The deviations from linearity within response time functions and the relative ordering of rates of rotation across axes obtained for recognition (Experiments 2–4) were somewhat consistent with the variations and ordering of slopes found for handedness discriminations on the same objects (Experiment 1). For handedness discriminations, the relative ordering of slopes around each axis was (slowest to fastest) x, y, z (with one exception—the one-version condition of Experiment 3); for all experiments involving recognition judgments, the relative ordering in Block 1 was x, z, y. In addition, with the same exception, for all experiments in both Blocks 1 and 13, rotations around the x-axis were always the slowest.

3. Response time functions were generally consistent with normalization around the axis originally used to generate the rotation, indicating that the shortest path in 3-D was used to align percepts with object representations.

4. When subjects did not have an opportunity to learn mirror-reversed versions of 3-D objects, which cannot be aligned with their enantiomorphs by any rigid 3-D rotation, performance still exhibited the same pattern of viewpoint dependency as it did with standard versions (with the exception of y-axis rotations for unfamiliar viewpoints in Experiment 4). This indicates that unfamiliar mirror-reversed versions of *3-D objects* are recognized via normalization. By implication, this suggests that unfamiliar mirror-reversed versions of *2-D shapes* are recognized by similar viewpoint-dependent mechanisms. Thus, the equivalent performance across viewpoints found by Tarr and Pinker (1989) for unfamiliar mirror-reversed versions are plausibly due to a 180º flip in depth, rather than to viewpoint-invariant mechanisms.

Although these findings are inconsistent with an *exclusively* viewpoint-invariant theory of object recognition/representation, it is possible to portray them as exceptional and as the result of recognition conditions atypical of "normal" conditions. Indeed, Biederman and Gerhardstein (1993) argued that viewpoint-dependent performance was obtained here because the present experiments violate at least one of three conditions for obtaining immediate viewpoint invariance: (1) objects must be decomposable into readily identifiable parts; (2) objects must be discriminable via distinctive part-based structural descriptions; and (3) the same structural description must be recoverable over different viewpoints. Biederman and Gerhardstein assume that the spatial relations encoded in structural-descriptions are too coarse to distinguish between the objects used in these studies (a violation of Condition 2). Yet there is evidence that even given distinct structural descriptions, viewpoint dependency is still obtained. For instance, Bartram (1976), Lawson (1993), and Srinivas (1993) employed familiar common objects that were clearly decomposable into parts and that did not violate the condition that objects have distinct structural descriptions (stimuli were generally members of different categories). However, such studies may have still violated the condition that all tested viewpoints give rise to the same structural description. In particular, restricted viewpoint-invariant theories, in which features are invariant over a limited range of viewpoints, might predict an effect of viewpoint, given averaging across different structural descriptions. The present experiments and those by Bülthoff and Edelman (1992) addressed this issue by employing relatively small changes in viewpoint (e.g., 15º or less) that are unlikely to have violated this condition. The systematic patterns across finely measured changes in viewpoint obtained in the present experiments and in Bülthoff and Edelman (1992) provide evidence for viewpoint-dependent mechanisms, regardless of qualitative changes in part structure.

Nonetheless, the present results and those of Bülthoff and Edelman (1992) are subject to the second condition—that because all the stimuli give rise to similar structural descriptions, they can only be discriminated by using viewpoint-dependent information. To the extent that such effects are dismissed as atypical, such an analysis offers a somewhat narrow framework for human object recognition. In many instances, objects that share similar parts in similar spatial relationships must be discriminated (e.g., animals or models of cars). Additionally, there are well-documented instances where viewpoint-dependent patterns have been obtained in tasks where objects are qualitatively dissimilar from one another (typically entry-level judgments; see Jolicoeur, 1985, 1988; Jolicoeur & Milliken, 1989; McMullen et al., in press). Thus, viewpoint-dependent mechanisms seem unlikely

to be limited only to cases in which visually similar exemplars must be discriminated (typically subordinate-level judgments).

The results presented here are also inconsistent with two specific proposals that viewpoint dependence is an exceptional case: the rotation-for-handedness hypothesis, which suggests that mental rotation is used only when handedness is possibly relevant to the judgment; and the hypothesis that normalization procedures are used only when the top and the bottom of an object must be located. In none of the experiments did determining handedness facilitate recognition, yet effects of viewpoint were found consistently. Furthermore, even making handedness explicitly irrelevant failed to eradicate effects of viewpoint. In addition, recognition of objects in viewpoints that preserved the position of the top of an object with respect to gravity (*y*-axis rotations) also failed to eradicate effects of viewpoint.

In contrast, the results of all of the experiments presented here are consistent with multiple views. This theory accounts for initial effects of viewpoint, diminished effects of viewpoint with practice, the lack of transfer between familiar and unfamiliar viewpoints, and the increasing performance cost with increasing distance from the nearest familiar viewpoint. Cumulatively these findings demonstrate that the representations involved in object recognition under at least some conditions are viewpoint specific and that normalization is used to compensate for this specificity.

Computational Issues in Multiple-Views Theory

Normalization procedures. Theories of object recognition address the fundamental question of how objects are recognized despite differences in 2-D image structure that arise from novel viewpoints. Multiple-views theory provides an answer by proposing that objects are represented in visual memory as a collection of viewpoint-specific representations (referred to as "viewer-centered" if the coordinate system is determined by the perspective of the viewer; see Marr & Nishihara, 1978). Thus, an object will be recognized directly if it is observed at a familiar viewpoint that matches a viewpoint-specific representation. However, since it is impossible to encode views at every possible viewpoint, there can never exist a multiple-views representation that encompasses every view of an object. To recognize unfamiliar and otherwise unencoded views, multiple-views theory includes normalization procedures to align the percept with a familiar view in memory. The addition of normalization greatly reduces the need for a large number of views, thus making multiple-views theory more parsimonious. However, the inclusion of normalization procedures also allows the extreme case of a single view sufficient for recognition over any viewpoints that preserve the image structure of an object. Although there is some empirical evidence suggesting that this may sometimes occur (Tarr & Pinker, 1989, Experiment 1), the computational efficiency gained by minimizing the magnitude of each transformation suggests that more than one view of an object is likely to be encoded. Furthermore, with the exception of all but the simplest objects, changes in viewpoint result in changes in the structure of the 2-D image, thereby leading to new views.

How are the direction and the magnitude of transformations determined? One of the most persuasive arguments against the use of normalization procedures is the paradox of needing to identify an object in order to know the correct direction and distance of the transformation needed to align it with a target representation (Corballis et al., 1978; R. N. Shepard & Cooper, 1982).[8] One solution to this paradox is that only a small portion of the information available in the percept is used to determine the transformation. For example, Huttenlocher and Ullman (1987; Ullman, 1989) present a computational theory of object recognition that, similar to multiple-views theory, relies on the normalization of input shapes to object models. Ullman (1989) suggests that recognition is dependent on "alignment keys"—cues to the pose of an object that are independent of the identity of the object. It is demonstrated that if three non-collinear landmark features are located in both the percept and the representation, the 2-D coordinates of these landmarks are sufficient to compute the direction and the magnitude of the transformation (as well as the translation and size scaling) necessary to bring the two into alignment.

Huttenlocher and Ullman (1987) suggest that the correct target representation for a given transformation may be determined by comparing the alignment keys of the observed object with all object representations, performing the necessary alignment, and then comparing all possible matches in parallel. However, considering that humans must encode an extremely large number of representations, this is not an entirely satisfactory solution. No attempt is made to reduce the search space of representations prior to comparison; the sheer number of possible matches is handled simply by appealing to parallel mechanisms. A more efficient alternative is to use an overconstrained alignment key (Ullman, 1989). Whereas three landmarks on an observed object can always be aligned exactly with three landmarks in representations, four or more landmarks may only be aligned approximately (unless of course there is perfect match)—for instance, by a least-squares algorithm. The resulting goodness-of-fit measure provides an indication of whether the actual normalization is worth carrying out, thereby providing a technique for reducing the search space. Huttenlocher and Ullman (1987) have offered a specific implementation of object recognition by alignment, demonstrating that viewpoint normalization can occur prior to recognition.

Other computational theories of recognition also dispense with the need to establish identity before normalization occurs. For instance, Bülthoff and Edelman (1992, 1993) propose a multiple-views theory of recognition in which normalization is accomplished through

view interpolation (Ullman & Basri, 1991). In their psychophysical experiments, Bülthoff and Edelman test the subtle predictions of the view-interpolation theory against the predictions of Ullman's alignment theory, as well as viewpoint-invariant theories. Their results not only provide evidence for multiple-views, but also are consistent specifically with that pattern of view generalization predicted by view interpolation. Although Bülthoff and Edelman focused on patterns of performance in error rates, their model may be extended to accommodate response times and may be able to simulate many of the patterns found in the experiments reported here. In particular, the patterns of performance observed at unfamiliar viewpoints located between closely spaced familiar views (40º–100º) are consistent with the predictions of normalization through view interpolation. Thus, the view-interpolation model may be considered as a demonstration not only that normalization may precede recognition, but also that there are plausible algorithms for implementing multiple views. Recent results extend this model even further. First, work has been done to implement similar recognition processes as neurally plausible networks (Poggio & Edelman, 1990). Second, there is evidence from behavioral and cell recording studies in monkeys that multiple-views mechanisms are used in the recognition of tube-like objects similar to those employed in Bülthoff and Edelman's studies (Logothetis, Pauls, Bülthoff, & Poggio, 1994).

Which viewpoints are encoded? Another computational issue concerns what criteria should be used by a multiple-views mechanism to determine when new views are encoded. This problem arises from there being far more viewpoints for even a single object than can be encoded in memory. The question is how to differentiate between a new view that is dissimilar enough from known views for it to be advantageous to encode it and a new view that is similar enough to a known view for it to remain unencoded.

Probabilistic factors. Tarr and Pinker (1989) address this problem by proposing that views of objects are encoded at any frequently seen and therefore familiar viewpoint of an object. Whether a view is actually encoded or not is probabilistic: the more often an object is seen from a particular viewpoint, the greater the likelihood that it will serve as a target for recognition at nearby viewpoints (this is also advantageous in that such views may be more likely to be encountered in the future). Thus, the most common views of an object will be the views recognized most easily by a direct match or by a minimal normalization to an encoded view. Such frequency- or exemplar-based learning may generalize across object class—an important consideration if unfamiliar exemplars of familiar classes are to be recognized. From an empirical standpoint, Jolicoeur and Milliken (1989) presented results that indicated that familiar views transfer to new objects observed under similar conditions (e.g., observing one car from a novel viewpoint may have an impact on the views used in representing cars in general). From a computational standpoint, Vetter et al. (1994) have presented methods for the derivation of "virtual views"—that is, for how view-specific representations may be generated for objects previously unseen in such viewpoints. The idea that the saliency of views is based on probabilistic factors is consistent with exemplar-based theories of human memory (see, e.g., Hintzman, 1986). Such models hypothesize that essentially all percepts are incorporated into memory, but that certain recurring events, objects, and/or concepts end up contributing the most weight to representations and, as such, appear most salient or prototypical. Moreover, viewpoint-specific representations are also consistent with exemplar-based theories in that they retain the early viewpoint-specific encoding of percepts, rather than derive abstractions such as structural descriptions.

Representing objects in familiar viewpoints makes adaptive sense—one efficient strategy for recognition is to concentrate on doing a good job at recognizing objects in their most commonly observed viewpoints, as when one faces the keypad on a telephone. This argument is supported by several pieces of evidence. Psychophysically, the results of Bülthoff and Edelman (1992), Tarr and Pinker (1989), and the present study demonstrate that familiar objects at unfamiliar views are recognized by normalizing them to the nearest familiar viewpoint, indicating that humans represent objects at familiar viewpoints. From a neuroscientific perspective, Kendrick and Baldwin (1987) found that some neurons in monkeys are maximally responsive to upright monkey faces and that other neurons are maximally responsive to upside-down monkey faces, but that neurons in sheep are maximally responsive only to upright sheep faces. They argued that the difference is related to the fact that monkeys, which are arboreal, often view other monkeys upside down, but that sheep almost never view other sheep upside down. Additionally, Perrett, Mistlin, and Chitty (1987) found that there exist separate cells in monkeys maximally sensitive to full-face views of monkey faces and other cells maximally sensitive to profiles. More recently, Logothetis et al. (1994) have established that monkeys learned to recognize novel 3-D objects by developing multiple view-specific representations of each object as they became increasingly familiar. Finally, Rock (1973, 1983) has pointed out that humans have difficulty recognizing objects at unusual orientations in the picture-plane. One reason for this is that humans frequently observe objects from an upright position or because many objects themselves have a common viewpoint with respect to gravity. Thus, there is evidence from both human performance and from studies of other species that views of an object are encoded on the basis their frequency of observation in the environment.

Geometric factors. A completely probabilistic mechanism for determining which views are encoded may not be entirely satisfactory when extended to 3-D. The sheer number of common views might easily overwhelm the

representational system (imagine how often one walks around a car). A somewhat different mechanism is needed for determining when a view of an object is unique enough to warrant encoding. One constraint that such a mechanism might exploit is the fact that most 3-D objects are self-occluding. Unlike views of 2-D shapes misoriented in the picture plane, no single view will depict all of the surfaces of a 3-D object—or put another way, views of 3-D objects will vary in their image structure. Consequently, determining unique views may rely on when the image structure of a viewpoint is qualitatively unique relative to known views. Only such "characteristic" views (Freeman & Chakravarty, 1980) may be considered as *immediately* warranting encoding.

What defines a qualitative view? One possibility is that qualitative configurations of image features such as cusps and T-junctions are extracted. Koenderink (1987, 1990) has proposed such a model, suggesting that 3-D objects are represented as a set of 2-D views, referred to as an *aspect graph* and classified by the way in which topological properties of surfaces vary with viewpoint. For instance, Koenderink suggests that a house might be represented by eight views—four for the views from each corner and four for each side alone (ignoring the top and bottom of the house). Such views are considered *generic*, in the sense that a range of adjacent viewpoints will give rise to the same qualitative view. Each generic view may be classified by a unique configuration of image features derived from a small topologically defined set of such configurations (six in all: the "language" of feature configurations is actually composed of three local and three multilocal types). Viewpoints sharing a familiar configuration for a given object will be considered simply as an instance of a known characteristic view. One recurrent criticism of Koenderink's work has been the failure to demonstrate that aspect graphs are computationally tractable. Indeed most implementations have concentrated on aspect graphs for polyhedral objects, rather than more naturally appearing curved objects. However, recent advances have allowed the computation of aspect graphs for some simple smoothly curved objects (see, e.g., Kriegman & Ponce, 1990), and there is some research exploring methods for reducing the large number of views that may arise if qualitative changes are considered at all image scales.

One implication of Koenderink's (1987, 1990) theory is that no more than one view-specific representation need exist for each generic view of an object. However, aspect graph theory is not a performance model; rather it is a competence model of the minimal number of *qualitative* views necessary for one to completely represent an object. Evidence that views are learned on the basis of probabilistic factors indicates that perceivers are also sensitive to *quantitative* changes in image structure. Therefore, although aspect graphs may be used as a framework for defining the qualitative boundaries between views (much as there are qualitative boundaries between phonemes), there is little data to suggest that perceivers exclusively encode generic views. All rotations in the picture plane are qualitatively equivalent in terms of image structure, yet there is ample evidence indicating that variations in picture plane orientation are instantiated as discrete views (Tarr & Pinker, 1989).

The representation of 3-D structure via characteristic views. The characteristic views of an object may be organized into a "graph structure" that describes the spatial relations between views and, in doing so, the 3-D structure of the object (for all known views; see Figure 10). Thus, when a qualitative change in image structure is encountered—that is, when a visual event occurs—a new view may be activated by following the graph representing the particular object. On the basis of this type of representation, Koenderink and van Doorn (1979) speculated that one variable in R. N. Shepard and J. Metzler's (1971) mental rotation experiments other than the angle of rotation may be the number of visual events through which a rotating object must pass (J. Metzler & R. N. Shepard, 1974, found large effects of changes in viewpoint both when singularities were crossed and when singularities were not crossed). Notably, this framework does not address how it was possible for the present subjects and those in other experiments (R. N. Shepard & J. Metzler, 1971; Parsons, 1987c) to normalize novel objects of uncertain 3-D structure through unseen views (since transitions in the graph may only occur between known views). One possibility

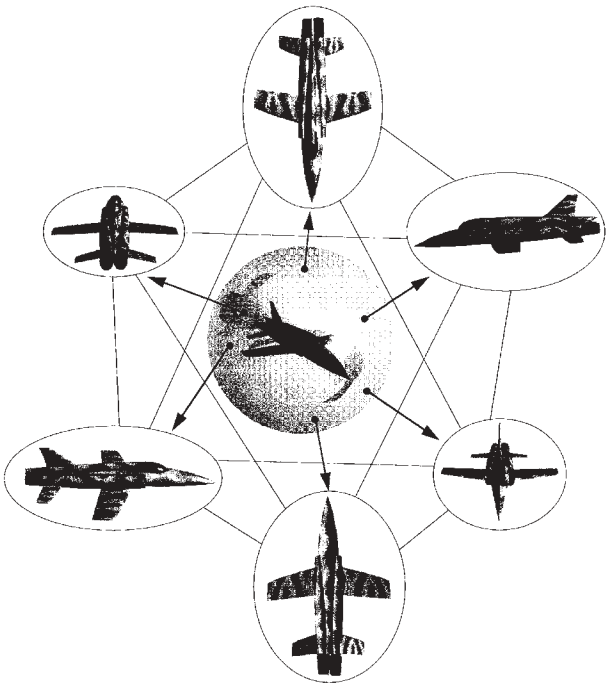

Figure 10. Multiple views. Imagine sitting on the surface of a bubble and observing a 3-D object from a variety of viewpoints. As your position on the view sphere or the pose of the object changes, new surfaces will become visible and remain relatively stable for some range, forming distinct, characteristic "views." Multiple-views theory proposes that visual memory for 3-D objects is composed of a connected set of such views.

is that all these experiments employed objects constructed from strings of cubes, connected at right angles. The 3-D structure of these objects was therefore highly predictable, even to the extent that subjects were able to predict new unfamiliar views on the basis of class similarity or object symmetries (Vetter et al., 1994). Such virtual views are most likely to play a role in the recognition of relatively familiar object classes and symmetrical objects. Indeed, Rock et al. (1989) found that less regular objects, such as wire forms, cannot be readily normalized into unseen views, and Bülthoff and Edelman (1992) found somewhat greater view specificity when they used highly asymmetrical tube-like objects.

Why Multiple-Views Is Not a Template Theory

Multiple-views theories may appear to advocate viewpoint-specific representations for every slight variation in viewpoint, of which there are an infinite number. This would seem to be a new example of the classic template-matching problem, which the field long ago dismissed as impractical. As discussed above, one solution to this problem is to posit that each viewpoint-specific representation has at least some generality for objects that vary in viewpoint or shape. Most complete or restricted viewpoint-invariant theories (Biederman, 1987; Hummel & Biederman, 1992; Marr & Nishihara, 1978) attempt to circumvent this problem by abstracting away from the initial percept. This abstraction, realized as a structural description, produces a viewpoint-invariant model composed of qualitatively defined 3-D volumes that approximate not only a particular input object, but many objects that share the same qualitative configuration of parts. No further mechanism for generality in either viewpoint or shape is needed—many members of a class of objects at many viewpoints lead to like representations. In contrast, any theory that relies on multiple views must specify the limits of generality in view-specific representations. There are several components to this, each generally based on the point that multiple-views models do not posit *structureless* representations, only *information-rich* representations. First, generality across viewpoint may be defined by using qualitative changes in image structure that arise because of variations in object geometry. Second, generality within characteristic views may be defined by the indexing of identity-independent features to establish an alignment between the percept and familiar views. Third, and perhaps most important, viewpoint-specific representations may be considered as exemplar-based entities in which many perceived views contribute to the response characteristics of a seemingly single familiar view and many perceived objects contribute to the response characteristics of a single known object or class (Tarr & Bülthoff, in press). Such exemplar-based multiple-views representations have been employed in several simulations of viewpoint-dependent recognition mechanisms (Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992; Poggio & Edelman, 1990).

Evidence for multiple-views object representations raises the issue of whether recognition includes the recovery of viewpoint-invariant 3-D representations. One possibility is that multiple-views representations encode no 3-D information or at most viewpoint-specific depth information. Supporting this argument, Witkin and Tenenbaum (1983) suggest that if one conceives the primary goal of early vision as a search for structure, surfaces and volumes are no longer distinguished as levels of representation; rather, they are alternative interpretations of structure within the image (such as that used in defining characteristic views). In contrast, several of the most influential accounts of human object recognition are based on the recovery of part-based 3-D descriptions and a subsequent viewpoint-invariant match to like representations in visual memory (Biederman, 1987; Hummel & Biederman, 1992; Marr & Nishihara, 1978). Such theories are referred to as structural-description theories because object representations consist of 3-D parts (e.g., cylinders or cubes) and the qualitative spatial relations between these parts (see also Cooper et al., 1992). Parts are recovered by locating configurations of features within the image that provide constraints on the 3-D structure of each part (e.g., a cylinder might be recovered by locating its axis of elongation through symmetry; see Marr & Nishihara, 1978). Viewpoint invariance is a consequence of the fact that, once recovered, a structural description of parts is stable over a range of adjacent viewpoints or all viewpoints. Such representations coincide with introspections about the 3-D nature of objects. It is not hard to imagine that we remember 3-D replicas of the objects we observe. Moreover, structural-description theories are appealing because most computer-aided-design systems (which implement the inverse problem of recognition: the 3-D representations used to produce 2-D images) employ part-based 3-D models as underlying representations.

Despite such attractive properties, a representational format sufficient for encoding 3-D structure does not automatically qualify as sufficient for recognition. Although there are few a priori reasons for hypothesizing that human object is based on 3-D parts, Marr and Nishihara (1978) argue that part-based structural descriptions meet their criteria of *sensitivity* and *stability* (although it is not entirely clear that other versions of structural descriptions, such as geon structural descriptions, satisfy either criteria; see Tarr & Bülthoff, in press). Specifically, structural descriptions capture shape information that is stable over similar objects (allowing novel objects to be recognized despite subtle differences), yet that is also sensitive to finer differences between similar objects (allowing individual exemplars to be differentiated). In Marr and Nishihara's view, because structural descriptions satisfy this and other criteria, they are well suited as representations for recognition. However, it should be noted that the properties of sensitivity and stability are not limited to structural descriptions; rather, any part-based description, either viewpoint invariant or

viewpoint dependent, may share these properties. In particular, the multiple-views theory advocated here is in no way necessarily "holistic" or structureless and may in fact include a hierarchical decomposition of features. Indeed, the problem of recovering 3-D parts is independent from the recognition problem and as such will play a role in recognition only if recognition mechanisms rely on this type of information.

Thus, as investigators, we are faced with two separate puzzles: our knowledge of the 3-D structure of objects and our ability to recognize objects. Many theorists attempt to solve the latter problem by first solving the former. However, as Marr and Nishihara (1978) point out, even if structural descriptions satisfy their criteria of stability and sensitivity, such representations must still satisfy their criterion of *accessibility*—whether the desired representation can be computed relatively efficiently from the image. Although Marr and Nishihara suggest that this is possible for the volumes used in their theory, and although Biederman (1987; Hummel & Biederman, 1992) argues likewise for geons, the recovery of 3-D parts is nevertheless a computationally difficult task, possibly more difficult than necessary for efficient recognition. Indeed, the results presented here indicate that this is a plausible argument, in particular because there are many established conditions under which successful recognition does not appear to require the prior recovery of 3-D parts.

Conclusions: Routes to Recognition

The multiple-views approach does not exclude other routes to recognition. It is probable that, depending on the task and context, human object recognition involves multiple mechanisms. Supporting this claim, many researchers have argued for at least two distinct recognition systems generally divided along the lines of coarse shape descriptions and metrically precise representations (Bülthoff & Edelman, 1993; Cooper et al., 1992; Farah, 1990; Jolicoeur, 1990a; Tarr & Pinker, 1990). Even Biederman's (Biederman & Gerhardstein, 1993) relatively exclusive approach is, by definition, limited to entry-level recognition (for instance, "chair" or "car," but not "Queen Anne" or "Colonial," or "Saturn" or "Toyota"). Thus, different recognition mechanisms, shown to be viewpoint dependent, must necessarily be used to discriminate between the cube objects shown in Figure 1 (which cannot be differentiated by Hummel and Biederman's, 1992, simulation).

The open question is whether a clear mapping can be made between the type of task and particular mechanisms. One possible answer is that there is a theoretical basis for allocating categorical tasks such as entry-level recognition to structural descriptions and exemplar-specific tasks to image-based multiple-views mechanisms. Specifically, many categories of objects are differentiable by a small set of parts in a unique configuration (Tversky & Hemenway, 1984). This observation forms the foundation for Biederman's (1987) part-based approach and the claim that entry-level access may be accomplished by the qualitative relations between no more than three geons. In contrast, when objects may be discriminated only by quantitative variations in shape or the complex relationships between features, coarse structural descriptions may be insufficient. This observation forms the foundation for the *image-based* approach (Bülthoff & Edelman, 1992; Tarr & Pinker, 1989) and the claim that exemplar-specific discriminations may be accomplished by metrically precise viewpoint-dependent representations, such as those posited in multiple views. However, the viability of this answer should be tempered by the fact that there is ample evidence for viewpoint-dependent performance in entry-level tasks (Jolicoeur, 1985, 1988, 1990a, 1990b; Lawson, 1993; Palmer et al., 1981), as well as exemplar-specific discriminations. Thus, although structural-descriptions may be *theoretically* appealing, there are few results to suggest that they form the basis of categorical recognition. In comparison, many results suggest that multiple views play an important role in recognition tasks ranging from the most extreme within-class discriminations to categorization at the entry level.

Why might human object recognition employ two or more mechanisms? On the basis of intuition, it may seem that most routine behavior necessitates establishing only categories of objects, as, for instance, a chair, a car, or a hungry carnivore (the stereotypical tiger in the jungle). If this were the case, a perceiver could rely on mechanisms restricted to general inferences about categories. This "default" recognition mechanism should be fast (it is rarely advantageous to be eaten) and robust, and it should filter out the majority of task-irrelevant information. This conception is consistent with how viewpoint-invariant recognition mechanisms based on either diagnostic features (e.g., Jolicoeur, 1990a) or parts (e.g., Biederman, 1987) have been characterized. However, there are also many circumstances in which it is crucial to differentiate between exemplars within a visually similar class. Here speed of recognition and restricting information content are not as imperative. There should be greater emphasis on accuracy, accomplished by preserving information so that an exact match may be made. This is consistent with how image-based multiple-views mechanisms have been characterized.

Why might discrete mechanisms have arisen during our evolutionary history? While it seems clear that categorization is important for survival (is it an apple or a rock?) and therefore adaptive, it is interesting to speculate on the circumstances under which the discrimination of specific identity may be adaptive as well. One possibility is that specific identity information was crucial to two important primate behaviors: tool making and kin identification. First, differentiating and consistently duplicating particular tools requires precise metric information. Supporting this conjecture, some anthropologists have speculated that early developments in tool technology and artistic expression may be linked to

increased competence in visual thinking (White, 1989). Second, visually identifying individuals within one's own species requires noticing subtle variations among similar parts (which is one reason why specialized face-processing mechanisms have been proposed). Thus, while it seems probable that many animals have evolved mechanisms for categorization—for example, for identifying all members of a given species by diagnostic features—only the demands of kin identification and tool replication render metrically precise mechanisms adaptive (much as other adaptive pressures led to the unique linguistic abilities of humans).

In summary, shape-based recognition in humans may be considered as a continuum bounded by at least two mechanisms: viewpoint-invariant processes that support primarily coarse category judgments, and viewpoint-dependent processes that support a range of judgments from categorical to exemplar specific. Several results support this hypothesis. First, findings from studies employing objects containing diagnostic features indicate that viewpoint-invariant mechanisms may be used for some categorical tasks (Biederman & Gerhardstein, 1993; Eley, 1982; Jolicoeur, 1990a). Second, findings of studies employing dissimilar objects falling into discrete categories also implicate viewpoint-dependent mechanisms in categorical tasks (Jolicoeur, 1985, 1988, 1990b; Lawson, 1993; Palmer et al., 1981). Finally, the findings presented here, along with the results of other studies employing configurally similar objects (Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992; Humphrey & Khan, 1992; Tarr & Pinker, 1989), offer strong evidence for the crucial role that viewpoint-dependent mechanisms play in exemplar-specific tasks. While such discriminations have sometimes been relegated to a circumscribed and infrequent role in "everyday" object recognition, differentiating between members within a class may be quite common. For numerous tasks, such as picking out a screwdriver from a toolbox filled with tools, a qualitative description based on a small set of features is insufficient. Rather, recognition must rely on more complex metric information encoded in view-specific representations. Thus, under many circumstances, including the recognition of both novel and familiar objects in both exemplar-specific and categorical tasks, recognition performance is viewpoint dependent, relying on multiple-views and normalization procedures.

REFERENCES

Bartram, D. J. (1974). The role of visual and semantic codes in object naming. *Cognitive Psychology*, **6**, 325-356.

Bartram, D. J. (1976). Levels of coding in picture-picture comparison tasks. *Memory & Cognition*, **4**, 593-602.

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

Biederman, I., & Cooper, E. E. (1991). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology*, **23**, 393-419.

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

Biederman, I., & Shiffrar, M. M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 640-645.

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

Bülthoff, H. H., & Edelman, S. (1993). Evaluating object recognition theories by computer graphics psychophysics. In T. A. Poggio & D. A. Glaser (Eds.), *Exploring brain functions: Models in neuroscience* (pp. 139-164). New York: Wiley.

Carpenter, P. A., & Just, M. A. (1978). Eye fixations during mental rotation. In J. W. Senders, D. F. Fisher, & R. A. Monty (Eds.), *Eye movements and the higher psychological functions* (pp. 115-133). Hillsdale, NJ: Erlbaum.

Cohen, D., & Kubovy, M. (1993). Mental rotation, mental representation, and flat slopes. *Cognitive Psychology*, **25**, 351-382.

Cooper, L. A., Schacter, D. L., Ballesteros, S., & Moore, C. (1992). Priming and recognition of transformed three-dimensional objects: Effects of size and reflection. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 43-57.

Cooper, L. A., & Shepard, R. N. (1973). Chronometric studies of the rotation of mental images. In W. G. Chase (Ed.), *Visual information processing* (pp. 75-176). New York: Academic Press.

Corballis, M. C. (1988). Recognition of disoriented shapes. *Psychological Review*, **95**, 115-123.

Corballis, M. C., & Nagourney, B. A. (1978). Latency to categorize disoriented alphanumeric characters as letters or digits. *Canadian Journal of Psychology*, **32**, 186-188.

Corballis, M. C., Zbrodoff, N. J., Shetzer, L. I., & Butler, P. B. (1978). Decisions about identity and orientation of rotated letters and digits. *Memory & Cognition*, **6**, 98-107.

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

Eley, M. G. (1982). Identifying rotated letter-like symbols. *Memory & Cognition*, **10**, 25-32.

Farah, M. J. (1990). *Visual agnosia: Disorders of object recognition and what they tell us about normal vision*. Cambridge, MA: MIT Press.

Folk, M. D., & Luce, R. D. (1987). Effects of stimulus complexity on mental rotation rate of polygons. *Journal of Experimental Psychology: Human Perception & Performance*, **13**, 395-404.

Freeman, H., & Chakravarty, I. (1980). The use of characteristic views in the recognition of three-dimensional objects. In E. S. Gelsema & L. N. Kanal (Eds.), *Pattern recognition in practice* (pp. 277-288). New York: North-Holland.

Gould, S. J. (1989). *Wonderful life*. New York: W. W. Norton.

Hinton, G. E., & Parsons, L. M. (1981). Frames of reference and mental imagery. In J. Long & A. Baddeley (Eds.), *Attention and performance IX* (pp. 261-277). Hillsdale, NJ: Erlbaum.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **93**, 411-428.

Horn, B. K. P., & Brooks, M. J. (1989). *Shape from shading*. Cambridge, MA: MIT Press.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, **99**, 480-517.

Humphrey, G. K., & Jolicoeur, P. (1993). An examination of the effects of axis foreshortening, monocular depth cues, and visual field on object identification. *Quarterly Journal of Experimental Psychology*, **46A**, 137-159.

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

Huttenlocher, D. P., & Ullman, S. (1987). Object recognition using alignment. In *Proceedings of the First International Conference on Computer Vision* (pp. 102-111). London: IEEE Computer Society Press.

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

Jolicoeur, P. (1988). Mental rotation and the identification of disoriented objects. *Canadian Journal of Psychology*, **42**, 461-478.

Jolicoeur, P. (1990a). Identification of disoriented objects: A dual-systems theory. *Mind & Language*, **5**, 387-410.
Jolicoeur, P. (1990b). Orientation congruency effects on the identification of disoriented shapes. *Journal of Experimental Psychology: Human Perception & Performance*, **16**, 351-364.
Jolicoeur, P., Gluck, M., & Kosslyn, S. M. (1984). Pictures and names: Making the connection. *Cognitive Psychology*, **16**, 243-275.
Jolicoeur, P., & Milliken, B. (1989). Identification of disoriented objects: Effects of context of prior presentation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 200-210.
Jolicoeur, P., Regehr, S., Smith, L. B. J. P., & Smith, G. N. (1985). Mental rotation of representations of two-dimensional and three-dimensional objects. *Canadian Journal of Psychology*, **39**, 100-129.
Kendrick, K. M., & Baldwin, B. A. (1987). Cells in temporal cortex of conscious sheep can respond preferentially to the sight of faces. *Science*, **236**, 448-450.
Koenderink, J. J. (1987). An internal representation for solid shape based on the topological properties of the apparent contour. In W. Richards & S. Ullman (Eds.), *Image understanding 1985–86* (pp. 257-285). Norwood, NJ: Ablex.
Koenderink, J. J. (1990). *Solid shape*. Cambridge, MA: MIT Press.
Koenderink, J. J., & van Doorn, A. J. (1979). The internal representation of solid shape with respect to vision. *Biological Cybernetics*, **32**, 211-216.
Koriat, A., & Norman, J. (1985). Mental rotation and visual familiarity. *Perception & Psychophysics*, **37**, 429-439.
Kriegman, D. J., & Ponce, J. (1990). Computing exact aspect graphs of curved objects: Solids of revolution. *International Journal of Computer Vision*, **5**, 119-135.
Lawson, R. (1993). *The achievement of visual object constancy: Evidence for view-specific representations*. Unpublished doctoral dissertation, University of Birmingham.
Logothetis, N. K., Pauls, J., Bülthoff, H. H., & Poggio, T. (1994). View-dependent object recognition in monkeys. *Current Biology*, **4**, 401-414.
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.
McMullen, P. A., Hamm, J., & Jolicoeur, P. (in press). Rotated object identification with and without orientation cues. *Canadian Journal of Experimental Psychology*.
McMullen, P. A., & Jolicoeur, P. (1992). Reference frame and effects of orientation on finding the tops of rotated objects. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 807-820.
Metzler, J., & Shepard, R. N. (1974). Transformational studies of the internal representation of three-dimensional objects. In R. L. Solso (Ed.), *Theories of cognitive psychology: The Loyola Symposium* (pp. 147-201). Potomac, MD: Erlbaum.
Murray, J. E., Jolicoeur, P., McMullen, P. A., & Ingleton, M. (1993). Orientation-invariant transfer of training in the identification of rotated natural objects. *Memory & Cognition*, **21**, 604-610.
Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and performance IX*. Hillsdale, NJ: Erlbaum.
Parsons, L. M. (1987a). Imagined spatial transformation of one's body. *Journal of Experimental Psychology: General*, **116**, 172-191.
Parsons, L. M. (1987b). Imagined spatial transformations of one's hands and feet. *Cognitive Psychology*, **19**, 178-241.
Parsons, L. M. (1987c). Visual discrimination of abstract mirror-reflected three-dimensional objects at many orientations. *Perception & Psychophysics*, **42**, 49-59.
Perrett, D. I., Mistlin, A. J., & Chitty, A. J. (1987). Visual neurones responsive to faces. *Trends in Neuroscience*, **10**, 358-364.
Pinker, S. (1984). Visual cognition: An introduction. *Cognition*, **18**, 1-63.
Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, **343**, 263-266.
Robertson, L. C., Palmer, S. E., & Gomez, L. M. (1987). Reference frames in mental rotation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 368-379.
Rock, I. (1973). *Orientation and form*. New York: Academic Press.
Rock, I. (1983). *The logic of perception*. Cambridge, MA: MIT Press.
Rock, I., & Di Vita, J. (1987). A case of viewer-centered object perception. *Cognitive Psychology*, **19**, 280-293.
Rock, I., Di Vita, J., & Barbeito, R. (1981). The effect on form perception of change of orientation in the third dimension. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 719-732.
Rock, I., Wheeler, D., & Tudor, L. (1989). *Can* we imagine how objects look from other viewpoints? *Cognitive Psychology*, **21**, 185-210.
Roediger, H. L., III, Weldon, M. S., & Challis, B. H. (1989). Explaining dissociations between implicit and explicit measures of retention: A processing account. In H. L. Roediger III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 3-42). Hillsdale, NJ: Erlbaum.
Shepard, R. N. (1984). Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review*, **91**, 417-447.
Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations*. Cambridge, MA: MIT Press.
Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, **171**, 701-703.
Shepard, S., & Metzler, D. (1988). Mental rotation: Effects of dimensionality of objects and type of task. *Journal of Experimental Psychology: Human Perception & Performance*, **14**, 3-11.
Shinar, D., & Owen, D. H. (1973). Effects of form rotation on the speed of classification: The development of shape constancy. *Perception & Psychophysics*, **14**, 149-154.
Simion, F., Bagnara, S., Roncato, S., & Umiltà, C. (1982). Transformation processes upon the visual code. *Perception & Psychophysics*, **31**, 13-25.
Srinivas, K. (1993). Perceptual specificity in nonverbal priming. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 582-602.
Tarr, M. J., & Bülthoff, H. H. (in press). Is human object recognition better described by geon-structural-descriptions or by multiple-views? *Journal of Experimental Psychology: Human Perception & Performance*.
Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, **21**, 233-282.
Tarr, M. J., & Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame? *Psychological Science*, **1**, 253-256.
Tarr, M. J., & Pinker, S. (1991). Orientation-dependent mechanisms in shape recognition: Further issues. *Psychological Science*, **2**, 207-209.
Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, **113**, 169-193.
Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, **32**, 193-254.
Ullman, S., & Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **13**, 992-1006.
Vetter, T., Poggio, T., & Bülthoff, H. H. (1994). The importance of symmetry and virtual views in three-dimensional object recognition. *Current Biology*, **4**, 18-23.
White, M. J. (1980). Naming and categorization of tilted alphanumeric characters do not require mental rotation. *Bulletin of the Psychonomic Society*, **15**, 153-156.
White, R. (1989). Visual thinking in the ice age. *Scientific American*, **261**, 92-99.
Witkin, A. P., & Tenenbaum, J. M. (1983). On the role of structure in vision. In A. Rosenfeld, B. Hope, & J. Beck (Eds.), *Human and machine vision* (pp. 481-544). New York: Academic Press.

NOTES

1. *Normalization* is intended as a generic term that encompasses any of several mechanisms by which two spatial representations may be compared (e.g., Bülthoff & Edelman, 1992; Shepard & Cooper, 1982; Ullman, 1989).

2. Biederman and Gerhardstein (1993) also reported three studies in which they used a sequential matching paradigm. A set of 10 qualitatively different objects (e.g., a brick vs. a cylinder) was employed in each experiment. While they obtained viewpoint-invariant performance across rotations in depth, as with Eley's (1982) study, they intentionally included diagnostic features in each object. Tarr and Bülthoff (in press) suggest that such features would not be diagnostic in a more ecological context.

3. Even with massive amounts of practice, response time functions usually fail to flatten out completely, displaying a small residual slope that is possibly due to rapid viewpoint-dependent feature location mechanisms (see Carpenter & Just, 1978; Tarr & Pinker, 1991).

4. Although it may seem that subjects must have known the handedness version of a shape in order to determine whether to rotate it in depth or in the picture plane (in the same way that it appears that they must have known the identity of the shape in order to know the direction and the magnitude of the rotation), handedness or identity information is actually irrelevant to computing the shortest path rotation. Recognition by alignment, such as that proposed by Huttenlocher and Ullman (1987), would compute both rotations in depth for aligning mirror reversals with their standards *and* rotations in the picture plane for aligning misoriented standards with standards.

5. All rotations are reported in degrees and are measured as follows: *x*-axis rotations, starting from the upper vertical and increasing with rotation toward the observer; *y*-axis rotations, starting from the right horizontal and increasing clockwise when looking down the *y*-axis from the positive end; and *z*-axis rotations, starting from the upper vertical and increasing clockwise when looking down the *z*-axis from the positive end (see Figure 2a). All rotations were around the center point of the imaginary box defined by the farthest reaching points of the object.

6. Tarr and Pinker (1989, 1991) hypothesize that diminished effects of viewpoint are due to the absence of rotation on some trials—those on which input shapes are matched directly against representations at familiar viewpoints—rather than to a speeding up of normalization procedures (although bimodal distributions were not found in early practice trials). A "speeding up" explanation would predict that an increased proficiency in rotating (still present for familiar viewpoints in the surprise block), reflected as faster rotation, would transfer to unfamiliar viewpoints, which it did not. Thus, it appears unlikely that the effect of practice on rate of rotation can be accounted for by an improvement in a task-specific skill (see Tarr & Pinker, 1991).

7. How did subjects recognize unfamiliar reversed versions of 3-D objects if no match to a familiar representation was possible? One explanation is that reversed versions are normalized to familiar viewpoints of standard versions that are partially congruent with the reversed versions. Once aligned, the two images may be searched for identical parts connected on opposite sides (at least 1 subject reported doing this). This strategy predicts that overall recognition times for unfamiliar reversed versions should be longer than for familiar standard versions, owing to the additional comparison. This prediction was not supported by the data: familiar and unfamiliar versions exhibit no reliable differences in overall response times. However, Figure 1 shows that enantiomorphs of the objects differ in the location of no more than one or two parts, suggesting that a comparison between them may occur quite rapidly (enantiomorphs may be seen by the reader by physically rotating Figure 1 around the vertical axis and holding the sheet to the light).

8. This "paradox" is actually one example of the more general problem of *indexing*, the process of finding a match between an input shape and candidate object representations. Implementations of computer-based recognition systems suggest that indexing is a fundamental problem regardless of the representational format chosen for objects. To locate a possible match for a given input shape, a small number of features in the input must be compared with features in known models. However, the combinatorics of such searches are staggering: even matching less than five features in an image to a slightly larger set of features in every model will yield huge numbers of search nodes. This problem applies to all current theories of recognition—whether the features are geons, qualitative configurations of image features, or simple edges. Even assuming that some changes in viewpoint introduce no additional variation—for instance, by relying on viewpoint-stable features—the magnitude of the indexing problem remains daunting.

APPENDIX
Construction of Stimuli

All stimuli were constructed with the Cubicomp ModelMaker300 3-D modeling system in perspective projection and were designed so that their spatial center coincided with the center of the modeling space. This ensured that all rotations were around the spatial center of the object. All misorientations were generated by rotations around the *x*-axis, the *y*-axis, and finally the *z*-axis. The basic set of seven objects, shown in Figure 1, were constructed from cubes connected at the faces. These objects are somewhat similar in appearance to those used by R. N. Shepard and J. Metzler (1971) and Parsons (1987c). Each of the objects shared a main vertical axis seven cubes high with a cube attached to each horizontal side of the bottom cube, thus forming a string of three cubes that clearly marked the bottom of the main axis. Cubes were attached to the main axis with the following constraints.

1. All objects were asymmetrical across the sagittal, frontal, and horizontal planes.

2. Each object contained a string of seven cubes that crossed the main vertical axis through either the sagittal or the frontal plane.

3. No other string of cubes on an object was longer than six cubes.

4. No cubes were attached to either the top or the bottom cube (other than the cubes marking the bottom of the main axis) of the main vertical axis.

5. No cube was attached to the face of a cube when either cube adjacent to that cube along the main axis had a cube attached to the same face.

Standard versions of each object were determined arbitrarily. Reversed versions (enantiomorphs) were generated by reflecting the object at upright through the sagittal plane (reversing left and right, but not front and back). Rotations of reversed versions were calculated following mirror reversal.

(Manuscript received March 23, 1994;
revision accepted for publication November 12, 1994.)