Testing Conditions for Viewpoint Invariance in Object Recognition

William G. Hayward
Yale University

Michael J. Tarr
Brown University

Based on the geon structural description approach, I. Biederman and P. C. Gerhardstein (1993) proposed 3 conditions under which object recognition is predicted to be viewpoint invariant. Two experiments are reported that satisfied all 3 criteria yet revealed performance that was clearly viewpoint dependent. Experiment 1 demonstrated that for both sequential matching and naming tasks, recognition of qualitatively distinct objects became progressively longer and less accurate as the viewpoint difference between study and test viewpoints increased. Experiment 2 demonstrated that for single-part objects, larger effects of viewpoint occurred when there was a change in the visible structure, indicating sensitivity to qualitative features in the image, not geon structural descriptions. These results suggest that the conditions proposed by I. Biederman and P. C. Gerhardstein are not generally applicable, the recognition of qualitatively distinct objects often relies on viewpoint-dependent mechanisms, and the molar features of view-based mechanisms appear to be image features rather than geons.

In recent years, a debate has emerged about the nature of visual representations of objects. Specifically, the debate has focused on the degree of viewpoint sensitivity of the representation. Two candidates have emerged: relatively viewpoint-invariant and relatively viewpoint-dependent approaches. Much has been written about each (e.g., see Marr & Nishihara, 1978; Flaut & Farah, 1990; Hummel, 1994; and Tarr, 1995). At their core, these representational alternatives differ on the specificity with which they encode the relationship between the viewer and an object. In a viewpoint-invariant representation (e.g., Biederman, 1987; Hummel & Biederman, 1992), the relationship between viewer and object is coded only coarsely. In Biederman's recognition-by-components (RBC) approach, for example, part-based object representations are stable over changes in viewpoint as long as the same parts remain visible. Thus, for rotations in depth, the RBC approach predicts reaction times (RTs) and error rates that do not vary with the magnitude of the rotation. An exception to this prediction occurs when the part configuration changes with viewpoint, either because of visible parts becoming occluded or occluded parts becoming visible. In this case, the representation will change and recognition performance will cease to be invariant.

By contrast, approaches based on viewpoint-dependent representations (e.g., Bülthoff & Edelman, 1992; Tarr, 1995) suggest that the relationship between the viewer and the object is coded with high specificity. As a consequence, recognition performance will be affected by relatively small perturbations in viewpoint. In Tarr's multiple-views approach, for example, view-based object representations depict the appearance of objects from specific viewpoints. In this model, unfamiliar viewpoints are matched to familiar viewpoints in visual memory through normalization procedures (such as in Poggio & Edelman, 1990; Shepard & Metzler, 1971; Tarr & Pinker, 1989). Crucially, this normalization process is time-consuming and predicts RTs and error rates that increase monotonically with increasing changes in viewpoint.

Findings of both viewpoint invariance and viewpoint dependence have been reported using a variety of stimuli and experimental paradigms (for reviews, see Biederman & Gerhardtstein, 1993; Bülthoff, Edelman, & Tarr, 1995; Jolicoeur, 1990). The current debate has been fueled largely by arguments about the validity of the stimuli and recognition tasks in which each type of finding occurs. To clarify situations in which viewpoint invariance is predicted, Biederman and Gerhardtstein proposed three conditions: The stimuli must (a) have readily identifiable parts that correspond to qualitatively defined shapes (i.e., configurations of features can be used to recover geons, or qualitatively defined 3-D volumes); (b) have parts that combine into qualitatively distinct configurations (at a minimum, different geons in similar configurations, as used by Biederman & Gerhardtstein, 1993, or similar geons in different configurations, as used by Liter, 1995); and (c) have the same parts visible over viewpoint changes (i.e., no changes in which visible parts are subsequently occluded). According to Biederman and Gerhardtstein, whenever these conditions are met recognition...
performance should be largely unaffected by changes in viewpoint.\(^1\)

**Experiment 1**

The first experiment provided two simple tests of the conditions for viewpoint invariance set forth by Biederman and Gerhardstein (1993). In the sequential matching task, participants saw one of the objects shown in Figure 1, a mask, a second object from Figure 1, and then the mask again (the same task as used in some of the experiments reported by Biederman & Gerhardstein, 1993). On each trial, the second object could be (a) the same as the first object and displayed in the same viewpoint, (b) the same as the first object but displayed in a new viewpoint, or (c) a different object. The task was to judge whether the two presentations were of the same object or two different objects regardless of any change in viewpoint. When the object was the same, the viewpoint change could be as extreme as a 30° rotation in depth between the two presentations. Of interest was whether such rotations would affect recognition performance in terms of both response latencies and error rates. Because of the limitations inherent in a sequential matching task (i.e., it measures explicit memory over relatively brief intervals and may be sensitive to iconic properties of the image; Ellis & Allport, 1986), we used the same objects in a naming task, in which participants were trained to associate names with the objects shown in Figure 1. On each trial, participants saw one of the objects until they named the object by pressing the appropriate key or 5 s had elapsed. After practice naming the objects shown from only a single viewpoint, shown in Figure 1, participants saw the same objects from new, unfamiliar viewpoints. Their task was again to name each object, disregarding any change in viewpoint. The viewpoint change for unfamiliar viewpoints could be as extreme as a 30° rotation in depth away from the familiar viewpoint. Of interest was whether such rotations would affect naming performance in terms of both response latencies and error rates.

The predictions of the differing approaches are straightforward. Each object was easily differentiated from all others on the basis of qualitative differences in shape. Crucially, no pair of viewpoints resulted in a change in the configuration of visible parts (no previously unseen parts became visible and no previously visible parts became occluded). Thus, according to the RBC approach (Biederman & Gerhardstein, 1993), recognition performance should not be affected by the viewpoint changes used in this experiment. Specifically, the recognition of viewpoints separated by 30° rotations (and all smaller rotations) was predicted to be approximately equivalent to the recognition of the identical viewpoint shown twice. Similarly, the naming of objects in viewpoints up to a 30° rotation from a familiar viewpoint (and all smaller rotations) was predicted to be approximately equivalent to the naming of the familiar viewpoint. By contrast, the view-based approach predicts that recognition performance is mediated by normalization between unfamiliar and familiar viewpoints (Bülthoff & Edelman, 1992; Poggio & Edelman, 1990; Tarr, 1995). Therefore, recognition judgments for viewpoints separated by 30° rotations and naming judgments for unfamiliar viewpoints generated by 30° rotations were predicted to be slower and less accurate than recognition judgments for the identical viewpoint shown twice and naming judgments for the familiar viewpoint. Smaller rotations were predicted to yield performance that was in between these two cases, with RTs and error rates increasing monotonically with larger viewpoint changes.\(^2\)

**Method**

**Participants.** Forty undergraduates at Yale University participated in the sequential matching task in return for course credit and 25 undergraduate and graduate students at Brown University.

---

\(^1\) In particular, Biederman and Gerhardstein (1993) claimed that studies reported by Bülthoff and Edelman (1992; Edelman & Bülthoff, 1992) and Tarr (1995) failed to satisfy these conditions, and, consequently, they obtained viewpoint-dependent patterns of performance. Although the generic claim may be valid in that increasing object similarity may produce an increase in the degree of viewpoint dependency, Tarr and Bülthoff (1995) disputed whether these conditions form a reasonable characterization of "everyday" recognition.

\(^2\) It may be possible for the recognition-by-components approach to accommodate small effects of viewpoint. This concession was made recently by Biederman and Gerhardstein (1995). Although they offered no mechanisms to account for small effects of viewpoint that might arise under conditions in which relative viewpoint invariance is predicted, these effects would not be expected to relate systematically to the degree of viewpoint difference between a memorial and input representation. Rather, such small viewpoint effects might be tied to the degree of difficulty in recovering invariant descriptions of particular parts or residual iconic differences between pairs of images. The effects predicted by the view-based approach, although not specifically quantitative, are expected to yield putative rates of normalization slower than 1,000 deg/s (Cohen & Kubovy, 1993; Tarr & Pinker, 1991) and, qualitatively speaking, are expected to be related systematically to the viewpoint difference over rotations of 3-D shape (Tarr, 1995).
participants in the naming task as volunteers. No participant served in any other study using the same stimulus items.

Materials. Five objects were created on a Macintosh computer using 3-D modeling software (Alias Research, Toronto, CA). The objects are shown in Figure 1. Each object had two parts: a large, central volume and a small volume attached to one face of the large volume. All 10 parts were qualitatively different shapes. Consequently, although each object contained two parts, knowledge of the shape of either part alone was sufficient for recognition. Each object was rendered in 10° increments rotated in depth around a vertical axis through the center of the large principal volume from a 0° viewpoint (the small volume pointed directly toward the viewer) to a 90° viewpoint (the small volume pointed directly to the left). This resulted in 10 viewpoints per object (Figure 1 shows the 40° viewpoint). Because all objects were rotated in only one direction from 0°, there were no rotations that ended up as mirror images of a previous viewpoint (e.g., a 20° rotation from -10° to 10°). Both parts were visible in all of the viewpoints for all the objects (thus, there were no part occlusions in this experiment). The objects were surface rendered with realistic overhead lighting but no cast shadows.

Design and procedure: Sequential matching. Participants were informed that two objects would appear in quick succession and that they should quickly decide whether the two objects were the same or different. They were told that the objects may be presented in different viewpoints but that recognition decisions should be made solely on the basis of object identity regardless of any rotation in depth. Each trial began with a fixation cross for 500 ms, followed by the first object for 200 ms, a 750-ms mask (a repetitive pattern derived from features of all the objects), the second object for 100 ms, and the same mask again for 500 ms. Response latencies were recorded from the onset of the second object, and a response deadline of 1,500 ms was imposed. Trials in which the participant did not respond by this limit were discarded.

Each of the five objects in the recognition set appeared in 34 “same” trials and 34 “different” trials, for a total of 340 trials per participant. “Same” trials were generated by pairing each of the 10 viewpoints per object with itself and with the six adjacent viewpoints of ±10°, ±20°, and ±30°, with the exception that rotations that produced viewpoints outside of the 0°–90° range were excluded. This design resulted in 10 “same” trials in which the two viewpoints were identical, 9 “same” trials in which the separation between images was 10°, 8 “same” trials in which the separation was 20°, and 7 “same” trials in which the separation was 30°. “Different” trials were generated by pairing each of the 10 viewpoints per object with one of the four other objects in the recognition set. These distractor objects were matched in pose to each viewpoint separation as there were targets. Trial order was randomly determined for each participant. Participants did not receive feedback about their performance. Breaks were scheduled randomly throughout the experiment.

Naming. Participants were shown a sheet of paper depicting all five objects from the 40° viewpoint along with their corresponding names (“Kip,” “KeF,” “Kor,” “Kal,” and “Kym”) and the response keys associated with each name (1, 2, 3, 4, and 6 on the numeric keypad). They were told to study the objects briefly (about 2 min) and that they would be trained to recognize the objects on the computer. Each trial began with a fixation cross for 500 ms, which was followed by an object displayed until a response was made or 5 s had elapsed. Trials in which participants did not respond by this limit were discarded. Response latencies were recorded from the onset of the object, and feedback for an incorrect response was provided in the form of an audible beep.

Training consisted of each of the five objects shown twice in conjunction with the appropriate name and key. For these training trials, the 5-s response deadline was not imposed and participants were free to study each object for as long as they wished. After training, participants practiced naming the five objects in four blocks in which each object was shown five times without the name or key. Between blocks, participants were given a short break. Throughout both training and practice, the objects were shown only in the 40° viewpoint. After practice, participants were “surprised” with the now-familiar objects shown once from the familiar viewpoint and once from each of six unfamiliar viewpoints generated by ±10°, ±20°, and ±30° rotations from the familiar viewpoint (resulting in a total of 35 surprise trials per block).

Results

Sequential matching. The mean response latencies for correct “same” trials and error rates for “same” trials are shown in Figure 2. For latencies, recognition performance was equivalent between the 0° and 10° rotation conditions and progressively poorer in the 20° and 30° rotation conditions. For errors, recognition performance was progressively poorer throughout the range, decreasing systematically from 0° to 30°. These differences were statistically reliable, as determined by one-way analyses of variance (ANOVAs). A reliable main effect for viewpoint difference was obtained for both latencies, F(3, 117) = 13.6, p < .001, and errors, F(3, 117) = 9.10, p < .001. Moreover, these effects were not due to the idiosyncratic influence of a subset of objects: Item analyses using individual objects as the random variable (as opposed to participants) revealed reliable main effects for viewpoint difference for both latencies, F(3, 12) = 7.73, p < .005, and errors, F(3, 12) = 11.5, p < .001. Finally, differences occurred evenly across the entire distribution of recognition latencies; analyses also were reliable when performed on the 25th percentile, F(3, 117) = 6.31, p < .001, and 75th percentile, F(3, 117) = 6.96, p < .001, rather than the mean of participants’ responses.3 To assess whether RTs and error rates increased monotonically with increasing viewpoint differences, we performed linear contrasts. These revealed reliable linear effects for both latencies, F(1, 117) = 36.3, p < .001, and errors, F(1, 117) = 25.2, p < .001. To assess the magnitude of this linear trend, we regressed mean RTs against the viewpoint difference. This analysis revealed a slope of 847 deg/s.

3 It is doubtful that these findings can be explained by participants overemphasizing accuracy at the expense of both speed and viewpoint invariance (Biederman & Gerhardstein, 1995). Even excluding trials in which participants failed to respond within 1,500 ms, we obtained mean error rates between 2.5% and 7.0%. Including such trials, which effectively are no-go responses, yields error rates of between 7.05% and 16.1%, which are comparable to the error rates obtained in go/no-go tasks by Biederman and Gerhardstein (1993).
rotations in depth, with easily differentiable objects, reliable viewpoint-dependent performance costs were obtained.

**Naming.** The mean response latencies for correct trials and error rates for the first and second blocks of surprise trials are shown in Figure 3. A three-way ANOVA on latencies using direction of rotation, viewpoint difference, and block number as variables revealed no statistically reliable differences between pairs of unfamiliar viewpoints equidistant from the familiar viewpoint and no interactions; therefore, equidistant pairs (±10°, ±20°, ±30°) were collapsed in all further analyses and in Figure 3. For both latencies and errors, naming performance was progressively poorer throughout the range, decreasing systematically from 0° to 30° regardless of the direction of rotation. Using the collapsed data for statistical analyses, we found that most of these differences were statistically reliable in two-way ANOVAs using viewpoint difference and block number as variables. A reliable main effect for viewpoint difference was obtained for latencies, $F(3, 72) = 7.88, p < .001$, but not for errors; a reliable main effect for block number was obtained for latencies, $F(1, 24) = 19.8, p < .001$, and for
errors, $F(1, 24) = 5.19, p < .05$; and a reliable interaction was obtained for latencies, $F(3, 72) = 3.88, p < .05$, but not for errors. Moreover, these effects were not due to the idiosyncratic influence of a subset of objects. Item analyses on latencies using individual objects as the random variable (as opposed to participants) revealed reliable main effects for viewpoint difference $F(3, 12) = 8.32, p < .005$, and block number, $F(1, 4) = 25.5, p < .01$, as well as a reliable interaction, $F(3, 12) = 5.65, p < .05$. Again, analyses were performed on the 25th and 75th percentiles of participants’ response latencies to ensure that the effects occurred evenly across the entire distribution. Viewpoint difference—25th percentile, $F(3, 72) = 6.17, p < .001$; 75th percentile, $F(3, 72) = 3.37, p < .05$—and block number—25th percentile, $F(1, 24) = 14.08, p < .001$; 75th percentile, $F(1, 24) = 11.39, p < .001$—were reliable for both analyses, although the interaction was reliable only for the 75th percentile analysis, $F(3, 72) = 4.11, p < .01$; 25th percentile, $F(3, 72) = 2.23, p = .09$. To assess whether RTs and error rates increased monotonically with increasing viewpoint differences, we ran linear contrasts separately on the collapsed data from the first and second blocks of surprise trials. For the first block, these revealed a reliable linear effect for latencies, $F(1, 72) = 24.8, p < .001$, but not for errors; for the second block, these revealed no reliable linear effects for latencies or errors. To assess the magnitude of these trends, we regressed the mean RTs against viewpoint difference. This analysis revealed a slope of 225 deg/s for the first block and a slope of 2,921 deg/s for the second block. This pattern of trends across blocks suggested that participants initially used normalization procedures to recognize the objects in unfamiliar viewpoints but that they quickly learned to recognize the objects equally well in multiple viewpoints once such viewpoints were familiar.

Discussion

Both of the tasks used in this experiment fulfilled all three of the conditions for viewpoint invariance proposed by Biederman and Gerhardstein (1993), yet we observed an increase in both response latencies and error rates with increasing disparity between viewpoints or between familiar and unfamiliar viewpoints. The magnitudes of these viewpoint effects, 847 deg/s for sequential matching and 225 deg/s for naming (when unfamiliar viewpoints were first encountered), were comparable to those obtained in a variety of other studies and well within the 1,000 deg/s limit cited by Cohen and Kubovy (1993) for interpreting viewpoint dependency in RTs as arising from normalization procedures. In our laboratory, we have replicated this finding with other qualitatively distinct objects, including some adapted from Biederman and Gerhardstein (1993) using matching-to-sample, sequential matching, and naming tasks (Hayward, in press; Tarr, Hayward, Gauthier, & Williams, 1994; Tarr, Büllhoff, Zabinski, & Blanz, in press). The results are remarkably consistent. For example, using similar sequential matching tasks and rotations in depth around the vertical axis, putative rates of normalization ranged from 750 to 1,183 deg/s in Tarr et al. (1994) and from 532 to 990 deg/s in Tarr et al. (in press); for similar naming tasks and rotations around the vertical axis, rates ranged from 731 to 895 deg/s in Tarr et al. (1994) and from 257 to 893 deg/s in Tarr et al. (in press).

The putative rate for the initial naming of familiar objects in unfamiliar viewpoints in the naming task, 225 deg/s, was comparable to the 141–333 deg/s rates obtained for the initial naming of visually similar objects in unfamiliar viewpoints (Tarr, 1995). Unlike the objects used in the current experiments, the objects used by Tarr were not qualitatively distinct but differed only in the spatial relations between parts. Importantly, in this latter study, several predictions of the multiple-views theory were confirmed. There was an initial linear cost for naming objects in unfamiliar viewpoints. With practice, naming became equivalent across familiar viewpoints, and, with the introduction of new, unfamiliar viewpoints, the linear cost for naming was returned but was now related to the distance from the nearest of several familiar viewpoints (Tarr, 1995). Such results leave little question that multiple viewpoint-specific representations were used. Similarly, in the naming task of this experiment, we obtained an initial linear cost for naming the objects in the first surprise block, but, with practice, naming performance became near-equivalent in the second surprise block. This result, although consistent with observers encoding multiple views (Tarr & Pinker, 1989), cannot directly distinguish between multiple-view and view-invariant accounts for the flattening of performance with practice. However, given other tests of this effect (Bülthoff & Edelman, 1992; Tarr, 1995; Tarr & Pinker, 1989), our preferred interpretation is that participants encoded multiple viewpoint-specific representations.

Importantly, patterns of viewpoint dependency cannot be accounted for by the idiosyncratic influence of a small number of accidental viewpoints, the idiosyncratic influence of more difficult object discriminations, or averaging over viewpoint pairs that preserve the same visible parts and those that do not. First, in Experiment 1, compensating for potential costs due to accidental viewpoints did not change the pattern of viewpoint dependency. Second, item analyses revealed that the viewpoint-dependent pattern of performance was obtained for all objects in the recognition set, and error rates were high enough to indicate that participants were not overemphasizing accuracy in “finer” distinctions at the expense of viewpoint-invariant processes. Third, in the current study, rotations in depth never changed the visible parts, yet there was a cost in performance. Thus, across different tasks and contexts, recognition performance for highly discriminable objects differing in the qualitative shapes of component parts was found to degrade roughly proportionally with the degree of viewpoint change. Here, we found that rotations in depth as small as 10° were sufficient to obtain effects of viewpoint. This body of viewpoint-dependent results cannot be easily accommodated under the conditions set forth by Biederman and Gerhardstein (1993), nor can they be explained by the viewpoint-invariant, part-based RBC approach (Biederman, 1987; Hummel & Biederman, 1992). On the other hand, our results are consistent with view-based approaches to object
recognition (e.g., Bulthoff & Edelman, 1992; Tarr, 1995). In particular, progressively poorer recognition performance with increasing viewpoint disparity is thought to be the empirical signature of normalization procedures used to match percepts with object representations in visual memory (Cohen & Kubovy, 1993).

One criticism of Experiment 1 is that although each object in our recognition set was distinct in terms of part shapes, the spatial relations between the parts of each object were highly similar. Thus, one could argue that the second condition proposed by Biederman and Gerhardstein (1993) was not satisfied by our objects in that they did not contain entirely unique configurations. Several points may be made about this issue. First, each object in our recognition set contained two qualitatively distinct parts. Because each individual part was sufficient for recognition, the combination of the two such parts should increase the dissimilarity between objects even given similar part configurations. This may be seen by considering that objects composed of a single part each would have identical configurations and therefore would be discriminable only on the basis of part shape. Objects composed of two distinct parts in similar configurations also would be discriminable only on the basis of part shape, but in this instance two parts are unique for each object. Thus, in a similarity space defined by geon-structural descriptions (e.g., Hummel & Biederman, 1992), objects composed of two unique parts would be more dissimilar to each other than objects composed of single unique parts. Second, in their own test of their proposed conditions, Biederman and Gerhardstein (1993, Experiment 3) used novel objects that were comparable to the objects used here (i.e., objects composed of similar configurations of differently shaped parts). Third, the majority of studies reported by Biederman in support of the RBC approach (Biederman, 1987; Biederman & Gerhardstein, 1993) focused on how qualitative differences in part shape lead to viewpoint invariance, but they have not examined the role of qualitative differences in object configuration (the predictions made by Hummel & Biederman, 1992, remain untested). Thus, the current experiment was designed specifically to address these studies. Fourth, Liter (1995) found that qualitative differences in spatial relations between objects did not diminish the effect of viewpoint, although their presence did improve overall recognition accuracy. Consequently, the introduction of differing spatial relations between stimuli used in the current experiment would not be expected to alter the degree to which recognition was viewpoint dependent.

Experiment 2

The results of Experiment 1 suggest that recognition of qualitatively dissimilar objects is mediated by a normalization process and shows approximately linear effects of a change in viewpoint on recognition and naming. One inference from such results is that recognition performance is a constant linear function related to the degree of absolute rotation between two viewpoints. However, it also is possible that the magnitude or slope of this function varies not only with the degree of viewpoint change but also with the degree of geometric change. Biederman and Gerhardstein (1993) investigated this issue in their Experiment 3, in which they compared the recognition of objects where a rotation in depth did not alter the configuration of visible parts, a so-called "quantitative" change, with the recognition of objects where an equivalent magnitude rotation in depth altered the configuration of visible parts, a so-called "qualitative" change (i.e., a violation of their third condition). They found that there was only a small effect of viewpoint when the configuration of visible parts remained consistent but a relatively larger effect of viewpoint when the visible parts changed. Therefore, recognition costs were not a constant linear function related to rotation in depth.

Although Biederman and Gerhardstein (1993) interpreted their result as support for the RBC approach, we know of no current view-based theory that proposes a strict linear relation between absolute rotation and recognition performance (although Shepard & Cooper's, 1982, approach to mental rotation does predict a strict linear relationship). In fact, a result similar to that observed by Biederman and Gerhardstein would be predicted by almost any extant theory of recognition. For example, alignment (Ullman, 1989) and view interpolation schemes (Poggio & Edelman, 1990) both require common "index features" to compute a match between two images. Thus, given two images that differ qualitatively (e.g., when a rotation produces a different configuration of visible parts), executing an alignment or view interpolation may be more difficult and hence more time-consuming relative to two images that differ only quantitatively (e.g., when a rotation does not produce a different configuration of visible parts).

To investigate these two alternatives, Hayward (in press) replicated Biederman and Gerhardstein's (1993) Experiment 3 using similar objects and an almost-identical task. However, Hayward posited that the difference between the quantitative and qualitative conditions would not be 3-D parts per se but image properties such as the shape of the silhouette that changed concurrently with changes in visible parts. For example, a qualitative change in the silhouette of a human head would be observed as it rotates around the vertical axis from facing left to facing right and the contours defining the nose enter, exit, and then reenter the bounding contour. Such visual events (Koenderink & van Doorn, 1979) are considered significant only if one relies on a feature set based on the way contours in the line drawing of the image intersect (for a review, see Van Effelterre, 1994). By contrast, if one relies on 3-D parts as features, no qualitative change will have occurred in that the same set of visible parts (eyes, nose, and mouth) will have remained visible through the entire rotation. Hayward tested these

---

4 Biederman and Gerhardstein's (1993) experiment was roughly a replication of a subset of the conditions tested by Bulthoff and Edelman (1992). Specifically, Bulthoff and Edelman's "inter" condition, in which consistent index features between familiar and unfamiliar viewpoints were available, resulted in smaller effects of viewpoint relative to their "extra" and "ortho" conditions, in which consistent index features between viewpoints were not available.
proposals by comparing recognition using shaded images of the objects with recognition using silhouettes of the objects. Hayward found no reliable differences in the patterns of viewpoint dependency obtained with shaded images and silhouettes. This result offers some evidence that the changes in features located in the bounding contours of the objects, not changes in configurations of 3-D parts, were responsible for the differences between the quantitative and qualitative conditions observed by Biederman and Gerhardstein (1993).

One caveat on this result should be noted. Although the equivalent performance for shaded objects and silhouettes would seem to rule out 3-D parts as the molar features used in recognition, it crucially assumes that robust descriptions of 3-D parts could not be recovered from silhouettes. Indeed, two different implementations of the RBC approach would likely fail to recover a part description from silhouettes: In Hummel and Biederman's (1992) model, configurations of internal contours such as "Y" and "arrowhead" vertices play a crucial role in geon recovery, and, in the more recent model proposed by Hummel and Stankiewicz (1996), there would be no path to the intermediate surface representations necessary for geon recovery. These models, of course, represent only cartoon simulations of possible routes to part descriptions; it is possible that an alternative part-based model might predict these results (e.g., Zhu & Yuille, 1996). Thus, Hayward's (in press) experiment does not distinguish completely between feature-based and part-based alternatives.

To further investigate the features used in object recognition, we used a sequential matching task similar to that used in Experiment 1. Here participants recognized the single parts shown in Figure 4A. These parts were similar to those used in Biederman and Gerhardstein's (1993) Experiment 4, although they used line drawings and here we used parts with photorealistically textured and shaded surfaces. Again, the task was to judge whether the two presentations were of the same part or two different parts regardless of any change in viewpoint. When the part was the same, the second presentation was either at the identical viewpoint or a viewpoint generated by a 45° rotation in depth. As illustrated in Figure 4B, two distinct 45° rotations were used: one in which only the quantitative structure as defined in the line drawing of the image changed and one in which the qualitative structure of the image changed. For instance, as defined by the methods of Koenderink and van Doorn (1979) and shown in Figure 4B, in the qualitative condition there was no change in the bounding contour from a smooth transition between the face of the cylinder and the edge of the body to an abrupt transition between the face of the cylinder and the edge of the body; in the quantitative condition the smooth transition remained in the bounding contour at both viewpoints. Of interest was (a) whether rotations in depth would affect recognition performance in terms of both response latencies and error rates and (b) whether there would be a difference in the degree of viewpoint dependency between the quantitative and qualitative conditions.

Crucially, for the qualitatively distinct single parts used in this experiment (geons), the RBC and view-based approaches make different predictions regarding both the overall impact of changes in viewpoint and the magnitude of viewpoint dependency. First, Biederman and Gerhardstein (1993) claimed explicitly that the recognition of single geons should be relatively viewpoint invariant. Indeed, this point is fundamental to the RBC approach in that overall viewpoint invariance may not be obtained without the viewpoint-invariant recovery of individual parts within the representation. Biederman and Gerhardstein tested their proposal using a matching-to-sample task in which participants were required to repeatedly recognize the same "target" part in a series of 18 trials. Although they found apparent viewpoint invariance over rotations in depth of up to 90°, it is difficult to interpret their results for two reasons: First, the matching-to-sample design provided opportunities for participants to learn multiple views (Tarr, 1995): Specific rotations away from the target viewpoint were repeated three times during a series, and the distractors in one series of trials were parts that were used as target parts in a later series. Second, because only a single part needed to be identified over many trials, participants might have selected local diagnostic features and used these repeatedly. Typical recognition conditions do not allow the continued use of preselected features for accurate recognition (for a discussion of this point, see Tarr & Bülhoff, 1995). Given these potential confounds, we opted to use the other explicit

![Figure 4. A: The 10 single parts used as stimuli in Experiment 2. Each of these 3-D volumes is qualitatively unique. From "Recognizing Depth-Rotated Objects: Evidence and Conditions for Three-Dimensional Viewpoint Invariance," by I. Biederman and P. C. Gerhardstein, 1993, Journal of Experimental Psychology: Human Perception and Performance, 19, p. 1176. Copyright 1993 by the American Psychological Association. Adapted with permission of the author. B: Example of equal magnitude rotations in depth that give rise to qualitative or quantitative changes in image structure (e.g., the shape of the bounding contour).](image-url)
recognition task used by Biederman and Gerhardstein (1993), that of sequential matching. As in Experiment 1, the RBC approach predicts only small costs for changes in viewpoint, whereas view-based theories predict a somewhat larger cost for changes in viewpoint.

Given the difficulties inherent in quantitatively predicting and interpreting the magnitude of viewpoint effects, much more critical in this experiment were predictions regarding the quantitative and qualitative change conditions. As exemplified in Experiments 3–5 of Biederman and Gerhardstein (1993), the molar features of the RBC approach are qualitatively defined parts. Thus, for changes in viewpoint across single parts, the RBC approach predicts little difference between conditions regardless of changes in image structure (with the exception of accidental views). Moreover, even if the RBC approach is able to accommodate small effects of viewpoint for the recognition of single parts (Biederman & Gerhardstein, 1995), it would not seem to predict differences across different pairs of viewpoints for each part. By contrast, view-based theories propose that the molar features of recognition are viewpoint-specific image features (e.g., Bricolo, Poggio, & Logothetis, in press; Bulthoff & Edelman, 1992; Poggio & Edelman, 1990; Tarr, 1995). Thus, for changes in viewpoint across single parts, these theories predict a difference between conditions, with smaller effects of viewpoint in the qualitative condition as compared with relatively larger effects of viewpoint in the qualitative condition.

Method

Participants. Twenty undergraduates at Yale University participated in return for course credit. No participant served in any other study using the same stimulus items.

Materials. Ten single parts were created on a Macintosh computer using 3-D modeling software. The parts are shown in Figure 4. Each part consisted of a single volume. All 10 parts were qualitatively different shapes and were adapted from Figure 12 of Biederman and Gerhardstein (1993). Each part was rendered in three viewpoints: the 0° viewpoint shown in Figure 4A and the ±45° viewpoints shown in Figure 4B that were generated by rotations in depth around a vertical axis through the center of the part from a 0° viewpoint. For each part one of the 45° rotations resulted in no qualitative changes in the shape of the bounding contour, whereas the other 45° rotation resulted in a qualitative change in the shape of the bounding contour (the particular direction of the rotation was different for each part). None of the viewpoints showed an accidental view. The objects were surface rendered with realistic overhead lighting but no cast shadows.

Design and procedure. The procedure was almost identical to the sequential matching task used in Experiment 1, except that participants were instructed to respond only when the two image pairs were perceived as the same object regardless of any change in viewpoint. If two different objects were presented, participants were instructed to do nothing and were told that the next trial would begin after 1,500 ms (a go/no-go response).

Each of the 10 objects in the recognition set appeared in 60 “same” trials and 60 “different” trials, for a total of 120 trials per participant. “Same” trials were generated by pairing the 0° viewpoint of an object with itself and with the ±45° viewpoints. Particular viewpoint pairs were repeated once during the experiment. This design resulted in 6 “same” trials per object: 2 in the identical viewpoint condition, 2 in the qualitative condition, and 2 in the quantitative condition. “Different” trials were generated by twice pairing each of the three viewpoints per object with one of the other nine objects. The distractor object for each trial was randomly selected, and the occurrence of particular objects was balanced so that each object and viewpoint served equally often as a distractor. Distractor objects were matched in pose to the viewpoint conditions used for same trials. Trial order was randomly determined for each participant. Participants were given a brief break midway through the experiment. Response latencies were recorded from the onset of the second object. Feedback for incorrect responses was not provided.

Results

The mean response latencies for correct “same” trials and error rates for “same” trials are shown in Figure 5. For latencies, recognition performance was poorer in the two rotation conditions than in the identical viewpoint condition. For errors, recognition performance was poorer only in the qualitative condition. These differences were statistically reliable in pairwise matched-pair t tests: The identical viewpoint condition versus the qualitative condition was reliably different for both latencies, t(19) = 7.68, p < .001,
and errors, \( t(19) = 3.40, p < .005 \); the identical viewpoint condition versus the quantitative condition was reliably different only for latencies, \( t(19) = 3.31, p < .005 \); and the qualitative condition versus the quantitative condition was reliably different for both latencies, \( t(19) = 2.61, p < .05 \), and errors, \( t(19) = 3.07, p < .01 \). Moreover, these effects were not due to the idiosyncratic influence of a subset of objects. Item analyses on latencies using individual objects as the random variable (as opposed to participants) revealed reliable differences for the identical viewpoint condition versus the quantitative condition for latencies, \( t(9) = 5.21, p < .001 \); the identical viewpoint condition versus the qualitative condition for latencies, \( t(9) = 5.09, p < .001 \), and marginally for errors, \( t(9) = 1.88, p = .09 \); and the qualitative condition versus the quantitative condition for latencies, \( t(9) = 2.92, p < .05 \), and marginally for errors, \( t(9) = 1.99, p = .07 \). Finally, all comparisons were conducted again using the 25th and 75th percentiles of participants’ recognition latencies rather than means. These analyses showed the same patterns of performance for the identical condition versus the qualitative condition—25th percentile, \( t(19) = 4.21, p < .001 \); 75th percentile, \( t(19) = 6.79, p < .001 \)—and the identical condition versus the quantitative condition—25th percentile, \( t(19) = 2.65, p < .05 \); 75th percentile, \( t(19) = 4.08, p < .001 \). The comparisons of the qualitative condition versus the quantitative condition, however, were not statistically significant—25th percentile, \( t(19) = 1.84, p = .08 \); 75th percentile, \( t(19) = 1.58, p = .13 \). To assess the magnitude of these effects, we regressed the mean RTs against viewpoint differences. This analysis revealed a slope of 523 deg/s for the qualitative condition versus a slope of 972 deg/s for the quantitative condition (Biederman & Gerhardstein, 1993, Experiment 3, obtained rates of 459 deg/s in their qualitative condition and 1,875 deg/s in their quantitative condition). Thus, across rotations in depth, with easily differentiable objects reliable viewpoint-dependent performance costs were obtained. Importantly, the magnitude of this effect varied with how the rotation altered the visible image structure. Qualitative changes produced larger effects of viewpoint relative to quantitative changes.

Discussion

The results of Experiment 2 are straightforward. First, recognition of depth-rotated, single-part objects was slower and less accurate when the viewpoint of the object changed between presentations than when the viewpoint of the object did not change between presentations. Second, for changes in viewpoint the extent of the cost in performance was not equal across all rotations; a rotation that produced a qualitative change in visible image structure resulted in a greater cost relative to a rotation that produced only a quantitative change. Such results are compatible with theories of object recognition that hypothesize that variations in viewpoint are compensated for through the normalization of viewpoint-specific features (e.g., Bricolo, Poggio, & Logothetis, in press; Bülthoff & Edelman, 1992; Poggio & Edelman, 1990; Tarr, 1995). By contrast, these results do not appear to be consistent with the RBC approach (Biederman, 1987; Hummel & Biederman, 1992), which posits that individual 3-D volumes will be recovered with minimal impact of viewpoint.

More recently, Biederman and Gerhardstein (1995) have suggested that small effects of viewpoint might be found across rotations in depth for one of three reasons: (a) partial occlusion of the object, (b) partial foreshortening of the object, and (c) "occasional employment of a strategic variant that was viewpoint dependent" (p. 1511). In the case of our Experiment 2, however, it appears unlikely that such explanations apply. First, the viewpoints for each 3-D volume were selected so that all features used in the recovery of geons, such as whether parallelism is present and the type of contour termination at a vertex, were available in every viewpoint. Thus, it is doubtful that the effects of viewpoint in this experiment were due to partial occlusion or foreshortening in any way that would be consistent with the RBC approach. Second, it is unclear exactly what recognition strategy qualifies as a "strategic variant," but both the absolute RTs and error rates, as well as the particular pattern of results, obtained in this experiment are consistent with the extensive body of research that shows an influence of viewpoint (Bülthoff et al., 1995). Therefore, there seems to be little reason to assume that performance here was mediated by a special-purpose strategy different from that used in numerous studies that provide evidence for view-based mechanisms.

General Discussion

The findings reported in this article appear to be incompatible with Biederman’s (1987) RBC model of recognition and specifically with many of the studies supporting the RBC model that seem to show immediate viewpoint invariance (e.g., Biederman & Gerhardstein, 1993). In particular, in Experiment 1 we used a methodology that was highly similar to the one Biederman and Gerhardstein (1993) used in their Experiment 3, but it showed systematic viewpoint dependence even though changes in viewpoint did not produce a change in the visible parts. One reason for this difference may be that in Biederman and Gerhardstein’s experiment, changes in viewpoint that retained the same visible parts resulted in roughly mirror-reflected image pairs. In our experiment, rotations in depth never crossed the bilaterally symmetrical front of the object, and so changes in viewpoint never resulted in two images of the object that were mirror reflections. This difference may be crucial because there is evidence that mirror-image transformations can be achieved with little or no cost to recognition performance (e.g., Biederman & Cooper, 1991; Cooper, Schacter, Ballesteros, & Moore, 1992). Recognition of unfamiliar views therefore may be particularly efficient across mirror-image transformations.

The results of our Experiment 2 reinforce this point, demonstrating that even for single-part objects, recognition over qualitatively different geometric features (e.g., when two images of an object are not mirror reflections) is much less efficient than recognition over qualitatively similar
features (e.g., when two images of an object are mirror reflections). This finding is entirely compatible with findings of viewpoint dependence with a change in an object’s visible parts (i.e., a qualitative feature change). In particular, any rotation of the object that either occludes or reveals parts will greatly affect the geometry of the image and, on the basis of the results of our second experiment, would be expected to result in a substantial cost in recognition. Crucially, however, as suggested by the results of Experiment 2, such costs are not restricted to part occlusions; recognition should be impaired by many different types of qualitative change in image structure.

One possible criticism of this view-based account (e.g., Bülthoff & Edelman, 1992; Bülthoff et al., 1995; Poggio & Edelman, 1990; Tarr, 1995) is that both the features of the representation and the processes used for matching inputs to stored representations are relatively underspecified. By contrast, Biederman’s (1987; Hummel & Biederman, 1992) RBC approach is highly specific in terms of both the features of the representation and the processes used to match inputs to stored representations. Biederman (1987) argued that object recognition is mediated by configurations of qualitatively defined 3-D parts that are recovered on the basis of co-occurring “nonaccidental” properties (e.g., parallel lines). Regardless of the degree of specificity currently found in each theory, it is clear that view-based models generally predict a linear relationship between the magnitude of rotation and recognition performance (the degree of viewpoint dependence, however, may be influenced by the presence or absence of qualitatively unique features; see Tarr & Bülthoff, 1995; Tarr et al., in press), whereas the RBC approach predicts a nonlinear relationship based on whether a rotation maintains the same configuration of qualitatively defined parts. Given these predictions, our current results have the following implications: First, they demonstrate that configurations of geons do not predict recognition performance in either sequential matching or naming tasks, despite using qualitatively dissimilar objects that have unique geon-structural descriptions. Second, the results of Experiment 2 demonstrate explicitly that viewpoint-dependent recognition mechanisms are not simply based on the magnitude of rotation between two viewpoints. Rather, they are influenced by the types of features available for matching (see also Tarr et al., in press). In this case, normalization between two viewpoints is related to qualitative changes in the outline shape of the image (a finding compatible with more recent implementations of view-based models, such as Bricolo, Poggio, & Logothetis, in press). Notwithstanding, it is still far from clear how normalization is achieved across the wide variety of recognition situations that confront the human visual system. In our view, however, the results of our experiments, along with the great many others that have investigated the recognition of depth-rotated objects, represent a firm basis on which to build the enterprise.

Conclusions

In this research there was a clear systematic effect of viewpoint on both recognition and naming performance for single-part and multipart objects. viewpoint-invariant approaches and, in particular, the RBC approach (Biederman, 1987; Hummel & Biederman, 1992), predict the opposite result. Moreover, based on the RBC approach, Biederman and Gerhardstein (1993) proposed three necessary conditions under which their theory holds. Our experiments satisfied all three conditions yet resulted in viewpoint-dependent recognition across two different recognition tasks. Such findings suggest that the RBC approach does not provide an adequate account of visual object recognition. By contrast, viewpoint-dependent approaches and, in particular, the view-based approach (as proposed by Bülthoff & Edelman, 1992, Poggio & Edelman, 1990, and Tarr, 1995), predict the obtained patterns of performance. First, in Experiment 1 we found that both sequential matching and naming across larger changes in viewpoint were slower and more error prone than across smaller changes in viewpoint. Second, in Experiment 2 we found that the degree of viewpoint dependency in the recognition of single parts depended on qualitative variations in image structure rather than on the parts themselves. These results, taken together with the results of other recent studies (Hayward, in press; Tarr et al., 1994; Tarr et al., in press), and the results of earlier studies (e.g., Bülthoff & Edelman, 1992; Ellis & Allport, 1986; Humphrey & Khan, 1992; Jolicoeur, 1985, 1988; Srinivas, 1993; Tarr, 1995), provide both a challenge for the RBC approach and support for view-based theories.

References


To be fair, view-based theories do posit specific mechanisms for recognition. Indeed, one of the earliest view-based models (Poggio & Edelman, 1990) proposed clear computational mechanisms for matching inputs to stored representations. This model was then expanded and tested against psychophysical data by Bülthoff and Edelman (1992; Edelman & Bülthoff, 1992) and, more recently, against neurophysiological data by Logothetis and Pauls (1995; Logothetis, Pauls, & Poggio, 1995). What has been less clear are what kinds of image-based features are used to compute a match, a question addressed recently by Tarr, Bülthoff, Zabinski, and Blanz (in press). Conversely, the specificity of the recognition-by-components (RBC) approach is sometimes overstated. For example, the predicates used to specify the spatial relations between parts are imprecise and, when spelled out by Hummel and Biederman (1992), appear somewhat ad hoc. Moreover, the RBC approach offers no clear mechanism for recognition in instances in which it does predict viewpoint dependence.


