Psychological Science
1997, Vol. 8, No. 4, 282-289

To what extent do unique parts influence recognition across changes in viewpoint?

Michael J. Tarr
Brown University
Heinrich H. Bülthoff, Marion Zabinski, Volker Blanz
Max-Planck Institute, Yale University, Max-Planck Institute

We investigated how varying the number of unique parts within an object influences recognition across changes in viewpoint. The stimuli were shaded objects composed of five 3D volumes linked end-to-end with varying connection angles. Of the five volumes, zero, one, three, or five were qualitatively distinct (e.g., brick versus cone), the rest being tubes. Sequential-matching and naming tasks were used to assess the recognition of these stimuli over rotations in depth. Three major results stand out. First, regardless of the number of distinct parts, there was increasingly poorer recognition performance with increasing change in viewpoint. Second, the impact of viewpoint change for objects with one unique part was less than that for the other objects. Third, additional parts beyond a single unique part produced strong viewpoint dependency comparable to that obtained for objects with no distinct parts. Thus, visual recognition may be explained by a view-based theory in which viewpoint-specific representations encode both quantitative and qualitative features.

Introduction

One of the central questions in visual perception is how observers recognize objects across changes in viewpoint. This problem is particularly applicable to cases of rotation in depth in that different 3D rotations result in different two-dimensional images on our retinae (Figure 1). Regardless of the drastic changes that may occur with viewpoint, observers are able to recognize objects from almost any direction. To account for this performance, theories of visual recognition may be divided into two major families: *View-Based* approaches and *Structural-Description* approaches. View-Based theories propose that recognition relies on features tied to the input image or images, that is, represented in a viewpoint-specific frame of reference (Bülthoff & Edelman, 1992; Humphrey & Khan, 1992; Poggio & Edelman, 1990; Tarr, 1995). In contrast, Structural-Description theories propose that recognition relies on a hierarchy of elements, usually represented in a viewpoint-invariant frame of reference (Marr & Nishihara, 1978; Palmer, 1977). For example, the most prominent Structural-Description theory (Biederman, 1987) posits that the elements of the hierarchy are a small set of 3D volumes, referred to as "Geons," and that objects are recognized by encoding configurations of Geons that uniquely describe each object ("Geon-Structural-Descriptions" or GSDs).

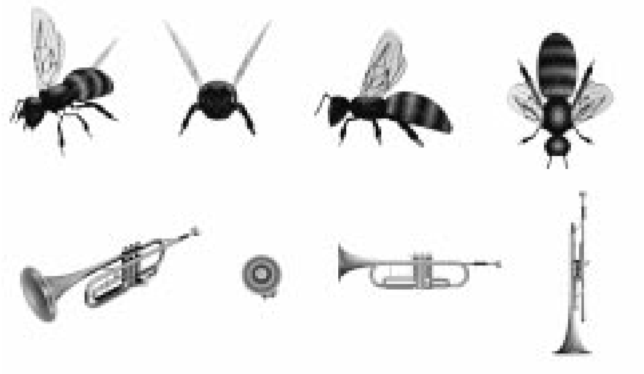

Figure 1. One of the central problems in visual object recognition is how observers recognize objects across changes in viewpoint. Despite dramatic changes in the image with rotations in depth we are able to recognize these 3D objects from almost any direction (barring intrinsically difficult viewpoints such as "accidental views", e.g., the head-on images of the bee and the trumpet). Illustration by Scott Yu.

These two types of theories make different predictions about recognition performance. View-Based theories predict that performance will be dependent on the distance from familiar viewpoints except in those instances when local diagnostic features are available. Biederman's GSD theory predicts that performance will be invariant across view-

We thank Jeff Liter and Alice O'Toole for their comments and, in particular Jeff for summarizing his thesis, and Jay Servidea for running many of the subjects. MJT was supported by a study grant from the Max-Planck Institut für biologische Kybernetik, Tübingen, Germany, and the Office of Naval Research, contract number N00014-93-1-0305. Some of the research reported here was done by MZ in partial fulfillment for the degree of B.A. in Psychology at Yale University. Please direct all correspondence to Michael J. Tarr, Department of Cognitive and Linguistic Sciences, Brown University, Box 1978, Providence, RI 02912, Tel: (401) 863-1148, Fax: (401) 863-2255, Email: *Michael_Tarr@brown.edu*, WWW: http://www.cog.brown.edu/brochure/people/mjt/tarr.html

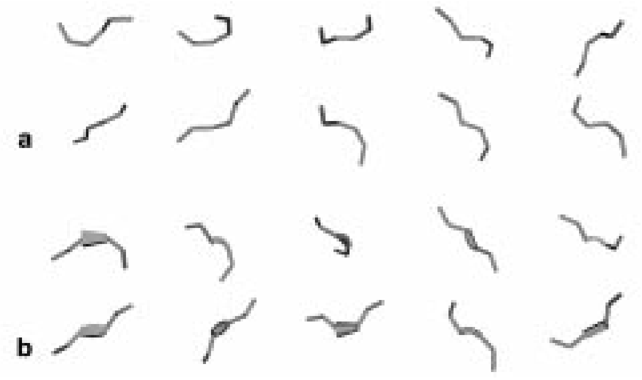

Figure 2. The paperclip-like objects used in the Baseline and 1Part conditions. Objects were either (a) "pure" paperclips as in the experiments reported by Bülthoff and Edelman (1992) (b) or created by inserting one distinctive Geon into the middle of each object as in the experiment reported by Biederman and Gerhardstein (1993).

point as long as the same configuration of Geons is available (when old Geons become occluded or new Geons come into view performance will cease to be viewpoint invariant; other Structural-Description theories make somewhat similar predictions – see Tarr, 1995). Relative to these two predictions, Biederman and Gerhardstein (1993) argued that the viewpoint dependency obtained in Bülthoff and Edelman's (1992; Edelman and Bülthoff, 1992) "paperclip" experiments was a spurious effect that resulted from a failure to include diagnostic parts (see Figure 2a for objects typical of those used in these experiments). To prove this point, Biederman and Gerhardstein (Exp. 5; 1993) ran an experiment in which subjects discriminated between paperclip objects where the middle segment was replaced by a qualitatively-distinct Geon (see Figure 2b). Not surprisingly, given the dissimilarities between both local and global features in Geons, they found complete viewpoint invariance. Although Biederman and Gerhardstein interpreted this result as evidence for the use of GSDs in recognition, Tarr and Bülthoff (1995) pointed out that this manipulation is equivalent to painting part or all of each paperclip with a distinct diagnostic color. Given that such colors would only be diagnostic in the limited context of the ten objects in the recognition set, Tarr and Bülthoff argued that simply inserting a locally distinctive feature or features is not typical of "real-world" recognition. In particular, the apparently diagnostic features or parts resultant from Geons may not be diagnostic when considered in the context of all familiar objects as the potential targets of recognition. In the two experiments reported here, we test this hypothesis by inserting three (Figure 3a) or five (Figure 3b) Geons into each object.

The specific predictions for increasing the number of distinctive parts vary with how one interprets the viewpoint invariance obtained when a single distinctive part is inserted. According to View-Based theories, viewpoint-invariant performance is the result of locally diagnostic features. Therefore, given the introduction of individual features that are locally confusable with each other (as is the case where the same distinctive parts reoccur in different configurations), View-Based theories predict that strong viewpoint dependency will be obtained. In contrast, according to GSD theory, viewpoint-invariant performance is the result of encoding unique *configurations* of parts. Additional parts may be incorporated into the structural description with little or no additional cost, therefore GSD theory predicts that recognition will be viewpoint invariant as long as the configuration of distinctive parts remains visible over a range of viewpoints[1].

General Methods

Materials. To test these predictions we created four new sets of paperclip-like objects. Each object was a realistically-shaded chain of five 3D volumes linked end-to-end. Objects were generated using OpenInventor on a SGI workstation. Images of each object were rendered in 24-bit color in four viewpoints and transferred to a Macintosh for conversion to a common 8-bit color palette. Of the five parts within each object, either 0, 1, 3, or 5 parts were qualitatively distinct from other members of the recognition set (e.g., brick versus cone). Ten different qualitatively-distinct parts of approximately the same size were used and were adapted from the ten Geons used by Biederman and Gerhardstein (1993). Non-distinct parts were cylindrical tubes similar to those used by Bülthoff and Edelman (1992). Independent of the number of distinct parts, the 3D angles between components were different for each object, thereby creating objects that could be discriminated on the basis of the angular relations between the parts as well as any part-shape differences. As shown in Figure 2a the objects in the Baseline set had no distinctive parts and corresponded closely to the objects used by Bülthoff and Edelman (1992). As shown in Figure 2b the objects in the 1Part set had a single distinctive part or Geon inserted into the middle of each object and corresponded closely to the objects used by Biederman and Gerhardstein (Exp. 5; 1993). As shown in Figure 3 the objects in the 3Parts set had three distinctive parts inserted into the middle of each object and the objects in the 5Parts set[2] had five distinctive parts inserted into the middle of each object. Individual parts appeared ap-

[1] Note that these predictions only address performance across variations in the *shapes* of individual parts. A more complete account would include the interaction between part shape and the relations between parts (Hayward & Tarr, 1995). This point will be returned to in the General Discussion.

[2] Because of a programming error the 5Parts set contained only nine different objects and the trials for one of the objects were repeated. However, since this repetition would not be expected to influence sequential-matching performance (Experiment 1) and the repeated object was only used as a distractor in the naming task (Experiment 2) there is no reason to believe that this error affected the results.

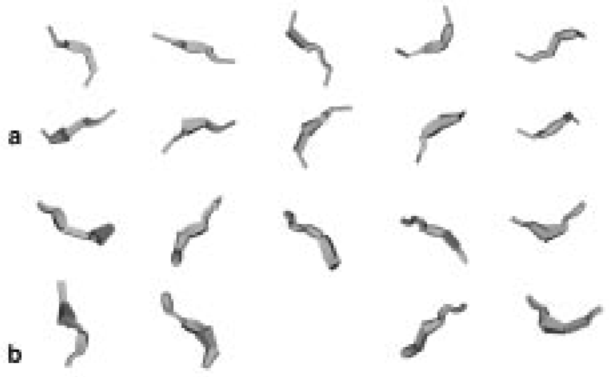

Figure 3. The Geon-String objects used in the 3Parts and 5Parts conditions. Objects were created by inserting (a) three or (b) five distinctive parts or Geons into the middle of each object (only 9 objects were used in the 5Parts condition).

proximately an equal number of times (although not exactly) in the 3Parts and 5Parts sets and to prevent subjects from simply focusing on the central part of each object, this part alone was never diagnostic within a given set. Although variation between objects was never based solely on the part ordering, no single part was diagnostic for the 3Parts and 5Parts sets. Thus, the configuration of parts uniquely specified an object and as such was diagnostic if observers use this type of information in recognition. Depending on the object, observers may have had to encode as many as three parts to differentiate it from other objects.

Subjects. Subjects were primarily drawn from the undergraduate introductory psychology course at Yale University and were provided with credit for their participation. Additional subjects were paid $5/hour for their time. A total of 120 subjects were run: 60 subjects in each of the two experiments (20 subjects per #Parts condition).

Design. Both experiments used three test conditions corresponding to the number of distinctive parts inserted into each paperclip (1, 3, or 5). Subjects in all conditions also ran in a Baseline condition of “pure” paperclips (no distinctive parts). The order of the Baseline and the #Parts conditions were counterbalanced across subjects – the only effect of order was an overall speeding up of responses, most likely due to general practice; therefore, all analyses collapsed over block order.

Experiment 1

Design and Procedure. Subjects had to judge whether two sequentially presented objects were the same or different. Subjects ran in one #Parts condition (1Part, 3Parts, 5Parts) and the Baseline condition. Each condition consisted of 4 practice trials using objects not relevant to the current study and 320 test trials in which the 10 objects from the appropriate stimulus set were shown. Each of the 10 objects appeared at four viewpoints separated by 30° rotations in depth around the vertical axis. Objects appeared equally often at each viewpoint (32 times each). Each trial was composed of a fixation cross for 500 ms, an image of an object for 200 ms, a visual mask for 750 ms, a second image of an object for 100 ms, and the same mask for 500 ms (Figure 4). The subjects’ task was to judge as quickly and as accurately as possible whether the two objects in a given trial were the same or different regardless of any change in viewpoint. One half of the trials paired an object with itself (same response) in one of four viewpoints (appearing equally often) leading to viewpoint differences of 0°, 30°, 60°, and 90°, and one half of the trials paired an object with a different object (different response). Subjects responded by pressing one of two keys on a keyboard. Responses and response times (RTs) were recorded using RSVP software running on an Macintosh LC475. The Macintosh was also used to control stimulus presentation at a resolution of 72 dpi on an Apple 13” Color Monitor. Subjects viewed the objects binocularly from a distance of approximately 60 cm from the screen resulting in images (which were not presented in stereo) that subtended a region of approximately 7° x 7° of visual angle. Images were presented in synchronization with the refresh of the screen and were preloaded into computer memory so that the entire image appeared in one refresh cycle. Subjects received feedback in the form of a beep when their response was incorrect or they did not respond in 7.5 s. This deadline was used to encourage subjects to respond in a reasonable time span and to reduce outliers in the RTs. Presentation order of the trials was randomized for each subject and subjects received two rests at random intervals during a condition and a longer rest between conditions. The entire experiment took less than one hour.

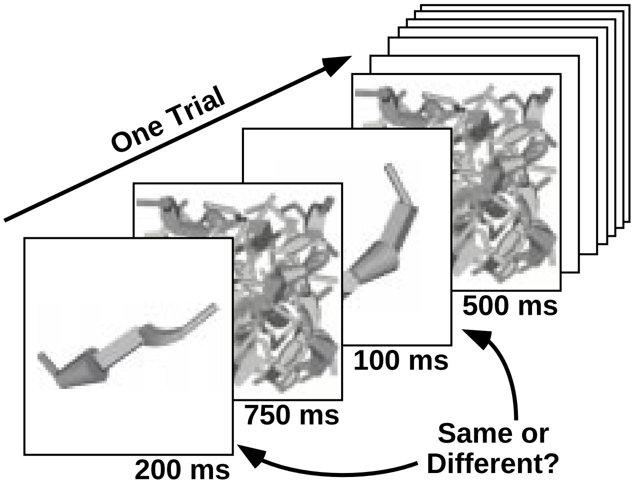

Figure 4. Experiment 1. The procedure used in the sequential-matching task.

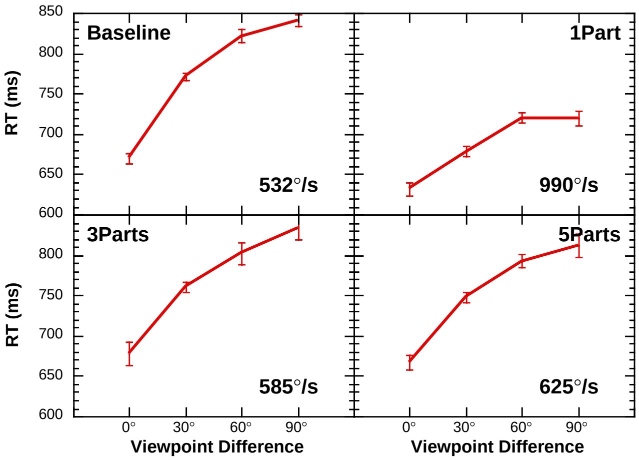

Figure 5. Experiment 1. Mean RTs for correct same trials in the sequential-matching task. The putative rate of normalization is shown for each condition. Error bars indicate the normalized standard error.

Results and Discussion

For the purposes of computing mean RTs, practice and incorrect responses were discarded. No adjustments were made to correct for outliers in that the RTs were normally distributed. Mean RTs were computed for same and different trials for each Condition (same/different): Baseline, 757 ms/800 ms; 1Part, 685 ms/713 ms; 3Parts, 759 ms/806 ms; and 5Parts, 748 ms/822 ms. Further analyses concerned only the mean RTs for same trials broken down by View difference (View0, View30, View60, and View90; Figure 5). The similarity in same and different RTs also allowed us to compute a sensitivity measure, d_L, from the correct same trials (hits) and the incorrect different trials (false alarms) for each Condition and each View (Figure 6). Note that d_L is functionally equivalent to d', but is computed using logistic distributions (Snodgrass & Corwin, 1988). To confirm the similarity of observer behavior across the #Parts conditions, an ANOVA was run on the RTs from the Baseline condition corresponding to the 1Part, 3Parts, and 5Parts conditions; #Parts and View were the independent variables. This analysis revealed no main effect or interaction for #Parts ($F<1$ for both effects) and a reliable main effect for View, $F(3,171)=90.9$, $p<.001$. Because there were no differences in performance in the Baseline condition across #Parts, data from the Baseline condition for all three #Parts conditions are plotted as the mean across all 60 subjects in Figures 5 and 6. This lack of a difference in Baseline suggests that comparisons between the #Parts conditions are valid and that any differences between conditions may be attributed to the #Parts manipulation.

To examine possible interactions between #Parts and View we performed two types of analyses: a #Parts x View ANOVA and three #Parts/Baseline x View ANOVAs comparing one of the #Parts conditions to the corresponding Baseline data.

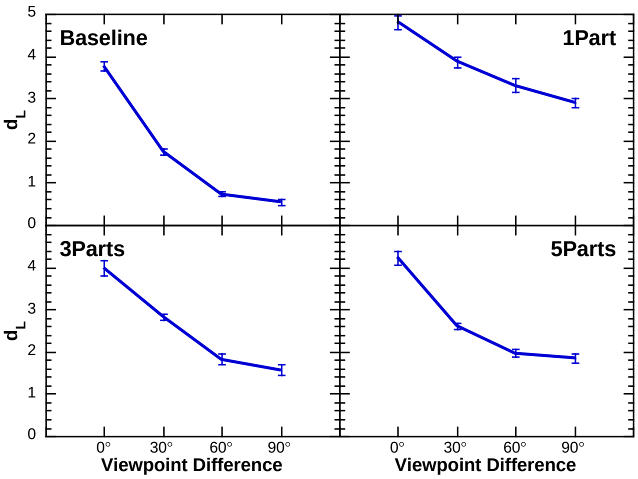

Figure 6. Experiment 1. Mean sensitivity computed using logistic distributions (d_L) in the sequential-matching task. For a discussion of this measure of sensitivity, see Snodgrass and Corwin (1988). Error bars indicate the normalized standard error.

Response Times. Across the #Parts conditions there was no effect of Condition, $F<1$, a reliable effect of View, $F(3,171)=77.1$, $p<.001$, and no interaction, F=1.23 (although a Condition x Linear interaction would most likely be reliable).

Comparing the #Parts conditions to their corresponding Baselines, there was a reliable #Parts/Baseline effect for the 1Part, $F(1,19)=5.15$, and 5Parts conditions, $F(1,19)=5.97$, both $p<.05$, a reliable effect of View for all three conditions, $F(3,57)=49.9$, $F(3,57)=25.3$, $F(3,57)=53.4$, all $p<.001$, and a reliable interaction only for the 1Part condition, $F(3,57)=5.65$, $p<.005$. The putative rates of normalization for each condition were 532°/s for Baseline as compared to 990°/s, 585°/s, and 625°/s respectively for 1Part, 3Parts, and 5Parts. These rates are comparable to rates obtained using both homogeneous object classes (Bülthoff & Edelman, 1992; Humphrey & Khan, 1992) and qualitatively distinct objects (Tarr, Hayward, Gauthier, & Williams, 1994) and fall below the 1000°/s upper bound cited by Cohen and Kubovy (1993) for normalization processes (for a review, see Tarr, 1995).

Sensitivity (d_L). Across the #Parts conditions there was a reliable effect of Condition, $F(2,57)=9.15$, a reliable effect of View, $F(3,171)=130$, both $p<.001$, and no interaction, $F=1.81$ (again a Condition x Linear interaction would most likely be reliable).

Comparing the #Parts conditions to their corresponding Baselines, there was a reliable #Parts/Baseline effect for the 1Part, $F(1,19)=124$, 3Parts, $F(1,19)=21.7$, and 5Parts conditions, $F(1,19)=15.3$, all $p<.001$, a reliable effect of View for all three conditions, $F(3,57)=96.7$, $F(3,57)=96.0$, $F(3,57)=166$, all $p<.001$, and a reliable interaction for all three conditions, $F(3,57)=9.97$, $F(3,57)=6.70$, $F(3,57)=6.00$, all $p<.005$. These effects and interactions reflect the fact that sensitivity was always higher and the impact of viewpoint always smaller in the #Parts conditions as compared to the Baseline condition.

Two results stand out in Experiment 1. First, for the Baseline and the 1Part conditions, the two conditions that are directly comparable to those used in Bülthoff and Edelman's (1992) and Biederman and Gerhardstein's (1993) studies, we replicated the overall pattern of performance across conditions. That is, the introduction of a single unique part significantly reduced viewpoint dependency for both RTs and sensitivity. Indeed, although the 1Part condition was not a perfect replication of Biederman and Gerhardstein's Experiment 5, in that we did obtain some effect of viewpoint, Biederman has repeatedly emphasized that the crucial prediction of his theory is that there is a dramatic decrease in viewpoint dependency when a distinct Geon is present (Biederman & Gerhardstein, 1995; Biederman & Bar, 1995). Moreover, Biederman and Gerhardstein (Exp. 5; 1993) used a match-to-sample task that allowed subjects to emphasize speed at the expense of sensitivity. Here we used a sequential-matching task that did not permit this strategy[3]. Indeed, using a sequential-matching task and objects with unique parts, Biederman and Gerhardstein (Exp. 3; 1993) did obtain small effects of viewpoint; conversely, using single Geons in a match-to-sample task with RT and accuracy feedback (conditions that closely match those used by Biederman & Gerhardstein), Tarr et al., (1994) were able to obtain near viewpoint-invariant performance.

Second, in terms of the predictions of the two theories it is clear that our overall results are highly consistent with View-Based theories of recognition. Specifically, regardless of the number of distinctive features or parts, including even the 1Part condition, recognition performance was viewpoint dependent (for related results see Tarr et al., 1994). Moreover, weaker viewpoint dependence was obtained only for the 1Part condition where unique local features are available. If observers had represented (and recognized) the objects as viewpoint-invariant configurations of parts, that is, GSDs, then viewpoint invariance (or at most weak viewpoint dependence) would have been found for the 3Parts and 5Parts conditions, which obviously was not the case. These results suggest that in many instances, recognition is view-based, relying on viewpoint-specific representations and time-consuming normalization procedures (Bülthoff, Edelman, & Tarr, 1995). It is also important to note that some of our current results are not consistent with a "pure" View-Based model in which undifferentiated two-dimensional images form the representation. Indeed, we have consistently argued for feature-based views, not simple image templates (Tarr & Bülthoff, 1995). In the present study the effect of including distinct features manifests itself in the sensitivity data, where overall performance was better in the 3Parts and the 5Parts conditions relative to the Baseline condition (Figure 6). Thus, although there is little evidence for multiple parts being represented in diagnostic configurations of Geons, the presence of additional distinct features does influence, and in this case facilitate, view-based recognition mechanisms.

Experiment 2

Although it may be possible to draw general conclusions based on the results of Experiment 1, the sequential-matching task primarily speaks to shorter-term representations and mechanisms that may mediate recognition (Ellis & Allport, 1986). Moreover, it is an explicit task and some researchers have argued that such tasks elicit sensitivity to stimulus features not found for implicit tasks (Biederman & Cooper, 1991; Cooper, Schacter, Ballesteros, & Moore, 1992). To examine the generality of our findings from Experiment 1 we used the same conditions in a naming task. Naming uncontroversially taps longer-term object representations and has a larger implicit component than sequential-matching. In particular, while the first experiment could have been performed by activating a entirely new object representation on each trial, accurate performance in a naming task requires both learning and the repeated activation of the learned representations.

Design and Procedure. Subjects had to name objects presented one per a trial. Subjects ran in one test condition (1Part, 3Parts, or 5Parts) and the Baseline condition. There were three phases to each condition: learning, practice, and test. First, subjects learned to associate nonsense names with four objects from each set shown in a single "canonical" viewpoint. The remaining objects served as distractors to which the correct response was "none-of-the-above." For 88 trials subjects viewed each of the four objects with the corresponding name, "tep" "tib" "tok" "tam" or "nil" (distractors) written at the bottom of the screen. Subjects saw each of the 4 named objects 16 times and each of the 6 distractors 4 times. On each learning trial the subject was to press the key corresponding to the name of the object. There was no time limit for responding. Subjects received feedback in the form of a beep when they made an incorrect response.

Second, subjects ran in a practice phase identical to the learning phase with the exception that the names of each object did not appear with the presentation of the object and there was a time limit of 7.5 s to respond. The subjects' task was to remember the name of the object and press the correct key. Feedback was provided for incorrect responses.

Third, in the test phase, subjects ran in two identical blocks of 154 trials (308 trials in all). Both the named objects and the distractors were shown in the canonical viewpoint and 6 new viewpoints generated by rotations in depth of 30°, 60°,

[3] Error rates in Experiment 1 were generally comparable to those obtained by Biederman and Gerhardstein (1993), for same trials in the 1Part condition they ranged from 4% to 20%.

and 90° clockwise and counterclockwise around the vertical axis from the canonical viewpoint (this design produced equal magnitude rotations in two directions – these are denoted by positive and negative viewpoint changes; such a distinction was meaningless in Experiment 1 where the direction of rotation was arbitrary). Within each block, subjects saw each of the 4 named objects 4 times at each of the 7 Views and each of the 6 distractor objects 1 time at each of the 7 Views. Each trial consisted of a fixation cross for 500 ms followed by an image of an object that was shown for up to 7.5 s or until the subject responded. Stimulus presentation and the collection of responses was the same as in Experiment 1. Subjects received two rests at random intervals during a condition and a longer rest between each phase. One condition was run during a one hour session and the other condition was run during a second one hour session on a different day.

Results and Discussion

Only the data for the test phase are considered in the results. For the purposes of computing mean RTs, incorrect and none-of-the-above responses were discarded. Mean RTs were computed for named targets and none-of-the-above trials for each Condition (named targets/none-of-the-above): Baseline, 1486 ms/2035 ms; 1Part, 1201 ms/1295 ms; 3Parts, 1673 ms/1576 ms; and 5Parts, 1633 ms/1732 ms. Further analyses concerned the mean RTs for named target trials broken down by View difference (View-90, View-60, View-30, View0, View30, View60, View90; Figure 7). The similarity in target and none-of-the-above RTs also allowed us to compute a sensitivity measure, again measured by d_L, for each Condition and each View difference (Figure 6). RTs for the Baseline condition are most likely uninterpretable because the d_Ls indicate that subjects could not recognize the objects when rotated away from the trained viewpoint (View0). Therefore, no statistics on RTs from the Baseline condition are reported. Data from all three #Parts conditions were, however, pooled for the Baseline and are plotted as the mean across all 60 subjects in Figures 7 and 8.

Again the results of most interest are the possible interactions between #Parts and View. The primary analysis was a #Parts x View ANOVA. Although, RTs analyses involving the Baseline condition were not be run because of low sensitivity, three #Parts/Baseline x View ANOVAs comparing one of the #Parts conditions to the corresponding Baseline were run on d_L.

Response Times. There was a reliable effect of #Parts, $F(2,57)=9.66$, $p<.001$, a reliable effect of View, $F(6,342)=19.0$, $p<.001$, and a marginal Condition x View interaction, $F(12,342)=1.70$, $p=.06$. Putative rates of normalization for the #Parts conditions were (computed by collapsing over rotations symmetric around View0): 893°/s for Baseline (not meaningful given the poor performance in this condition) and 508°/s, 257°/s, and 282°/s respectively for the 1Part, 3Parts, and 5Parts conditions. These rates are comparable to those obtained in other studies (Bülthoff &

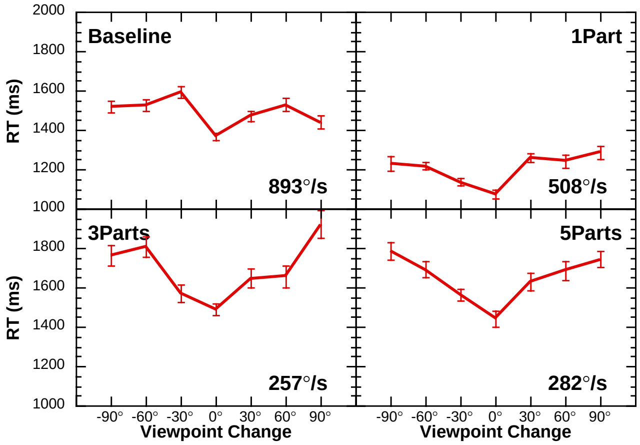

Figure 7. Experiment 2. Mean RTs for correct named trials in the naming task. For the Baseline condition, sensitivity measures indicate that subjects could not recognize the objects when rotated away from the trained viewpoint (View0), therefore no statistics on RTs from the Baseline condition are reported. The putative rate of normalization (collapsing over rotations symmetric around the 0° viewpoint) is shown for each condition. Error bars indicate the normalized standard error.

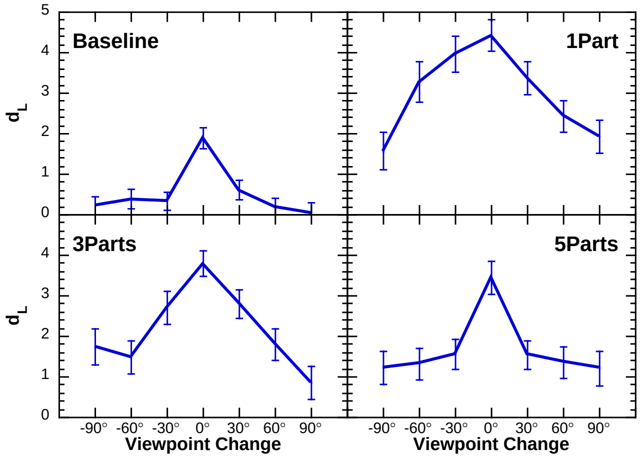

Figure 8. Experiment 2. Mean sensitivity measured on a logistic distribution (d_L) in the naming task. Sensitivity can be measured in this experiment because the task includes distractor objects for which a name response constitutes a false alarm. Error bars indicate the normalized standard error.

Edelman, 1992; Humphrey & Khan, 1992; Tarr et al., 1994) and fall well below 1000°/s (Cohen & Kubovy, 1993).

Sensitivity (d_L). There was a reliable effect of #Parts, $F(2,57)=7.21$, $p<.005$, a reliable effect of View, $F(6,342)=61.2$, $p<.001$, and a reliable #Parts x View interaction, $F(12,342)=5.59$, $p<.001$.

Comparing the #Parts conditions to their corresponding Baselines, there was a reliable #Parts/Baseline effect for the 1Part, $F(1,19)=44.0$, 3Parts, $F(1,19)=32.5$, and 5Parts conditions, $F(1,19)=56.2$, all $p<.001$, a reliable effect of View for all three, $F(6,114)=32.7$, $F(6,114)=30.8$, $F(6,114)=17.3$, all $p<.001$; and a reliable interaction for the 1Part, $F(6,114)=6.10$, and the 3Parts conditions, $F(6,114)=5.14$, both $p<.001$. These effects and interactions reflect the fact that sensitivity was always higher in each #Parts condition as compared to the Baseline. Notably, despite an elevation in sensitivity in the 5Parts condition, there is little difference in the pattern relative to the Baseline condition (Figure 8; this is supported by the high correlation, r = .98, between the d_Ls for the Baseline and 5Parts conditions across View). This result indicates that although recognition was poor at unfamiliar viewpoints for 5Part objects, subjects were sometimes able to use viewpoint-invariant local features to facilitate normalization.

The results of Experiment 2 essentially replicate the results of Experiment 1 with few exceptions. First, recognition performance, in terms of RTs and sensitivity, was clearly viewpoint dependent in the 3Parts and the 5Parts conditions. Once again the addition of distinct parts did not lead to viewpoint invariance, indicating that observers did not learn GSDs, but rather used viewpoint-specific representations. Second, for the 1Part condition, where unique local features are available, we obtained only weak viewpoint dependence in RTs, but strong viewpoint dependence in sensitivity. Third, the sensitivity advantage for the 3Parts and 5Parts conditions over the Baseline (Figure 8) offers salient evidence that although object recognition may be viewpoint dependent, it is not insensitive to distinct shape features encoded within long-term visual representations. Supporting this interpretation, Perrett, Oram, and Wachsmuth, (1996) have suggested that viewpoint dependency may arise from collections of viewpoint-specific features that contribute to the recognition of an object in a manner proportional to the strength of their response, i.e., based on their visual similarity to other features. Thus, with increasing similarity between local features, as is the case in the 3Parts and 5Parts conditions, each featural unit will require more evidence to reach a given threshold, as the "signal" is now smaller relative to the "noise" – thus, there is an increasing cost for changes in viewpoint. Overall, these findings provide further evidence that visual recognition, in both short-term *and* long-term tasks, typically relies on view-based mechanisms and representations.

General Discussion

Our current results indicate that visual recognition is viewpoint dependent in all but the most simple cases, for example when there are highly unique local features such as single Geons or different colors (Tarr & Bülthoff, 1995). Supporting this argument, Tarr et al. (1994) found robust effects of viewpoint across sequential-matching and naming tasks for both line drawings and shaded images of "Geon objects" adapted from Biederman and Gerhardstein's (1993) stimuli. Importantly, Tarr et al. also found that effects of viewpoint could not be accounted for by changes in visible qualitative features or parts – the performance cost due to changes in viewpoint was dependent on quantitative changes (i.e., the magnitude of the rotation), but independent of qualitative changes (i.e., features or parts becoming visible or occluded, thereby leading to a new description). Regarding the strength of such effects, the similarities in RTs in the Baseline condition and in the 3Parts and 5Parts conditions indicate that the results of earlier "paperclip" experiments (Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992) generalize to the recognition of objects composed of distinct parts or Geons. This result undermines the criticism of Biederman and Gerhardstein (1993) that the viewpoint-dependent results obtained with paperclip stimuli lack generalizability. Indeed, the commonalities observed here suggest that the mechanisms studied in the recognition of the paperclips and other novel homogeneous stimulus classes are closely tied if not the same as the mechanisms used in the recognition of more common objects. The differences, however, in sensitivity between the Baseline and the 3Parts and 5Parts conditions indicate that distinctive shape features do play an important role in view-based recognition.

Related work

The experiments presented here and, in particular, the 1Part condition, are quite similar to Experiment 5 of Biederman and Gerhardstein (1993). Notably, there is a discrepancy between their results, which were viewpoint invariant, and our current results, which are viewpoint dependent. Despite the fact that Biederman has stated that the most important prediction of his theory is that viewpoint dependency is less when a distinct Geon is present (Biederman & Gerhardstein, 1995; Biederman & Bar, 1995), a pattern we obtained here, we offer several speculations for our inability to precisely replicate their findings. First, Biederman and Gerhardstein's stimuli were line drawings rather than shaded images. Although it has sometimes been assumed that the goal of early vision is to derive a line drawing of an image (e.g., Marr, 1982), line drawings represent a fairly specific rendering of surface discontinuities that presumes an exclusively geometric approach to recognition. In contrast, shaded images provide observers with surfaces, texture, and shading – all properties that may contribute to "everyday" recognition. Second, Biederman and Gerhardstein's task was match-to-sample rather sequential-matching or naming. The match-to-sample task requires that subjects identify only a single target over a series trials. Therefore, subjects may select and rely on features that are sufficient for establishing the presence of the target in a limited context, but that may not uniquely specify the object given additional distractors or the benefit of repeated recognition episodes (as noted earlier, Tarr et al., 1994, were able to obtain almost complete viewpoint invariance using a match-to-sample task). In contrast, sequential-matching and naming provide a somewhat closer approximation of "everyday" recognition. Specifically, both tasks require observers to encode or activate a new object representation on each trial, thereby making the selection of diagnostic features for a single object less advantageous. Third, Biederman and Gerhardstein provided subjects with accuracy *and* response time feedback on each trial. This manipulation may have prompted subjects to respond as rapidly as possible on the basis of limited local information. Indeed, false alarm rates in Biederman and Gerhardstein's experiment ran as high as 20% indicating that subjects were strongly biased to respond as if the single target was present without positive identification.

Two recent studies (Liter, 1995; Liu, Kersten, & Knill, 1995) used stimuli similar to those used here and by Biederman and Gerhardstein. The results of both studies support the claim that object recognition is typically viewpoint dependent. Liter (1995) examined the influence of qualitative and quantitative features on viewpoint dependency. His stimuli were also modeled on Bülthoff and Edelman's (1992) paperclip objects and consisted of four tubes or four qualitatively-varying parts connected end-to-end. Using a recognition memory task, generalization to new viewpoints was tested over rotations in depth for two conditions. The Quantitative condition was analogous to our Baseline condition in that objects varied only in the connection angles between tubes and their lengths. The Qualitative condition was analogous to our 3Parts and 5Parts conditions in that objects varied in the order of differently shaped components. Liter found that recognition in both conditions was viewpoint dependent. However, as in our comparison of recognition sensitivity for the 3Parts and 5Parts conditions relative to the Baseline, generalization over depth rotations was better when the objects differed qualitatively. This result reinforces our point that although recognition may remain viewpoint dependent, it is influenced by qualitative variations in object features (although not in a way that would be predicted by GSD theory).

Liter (1995) also investigated the influence of qualitative variations in the spatial relationships between parts. Here, all of the parts were tubes, but some of the connections were end-to-middle. Recognition performance was measured in: a Quantitative condition in which the pattern of connection types was the same for all objects, but the connection angles, the lengths of the parts, and the location of the connections varied; and, in a Qualitative condition in which the pattern of connection types varied. Similar to our 3Parts and 5Parts conditions, Liter obtained better overall recognition performance in the Qualitative condition, but no dimunition in the effect of viewpoint. Indeed, similar effects of viewpoint were found for the Quantitative condition. Thus, it appears that qualitative differences in spatial relations as well as object shape play a role in the recognition of objects across changes in viewpoint, but that neither difference is sufficient to support viewpoint invariance.

Liu et al. (1995) also used objects modeled on Bülthoff and Edelman's (1992) paperclip objects. In a 2AFC match-to-sample task using noisy images, recognition at novel viewpoints was poorest for objects composed of disconnected balls, intermediate for balls connected with thin wires, and best for balls connected with tubes. Accuracy increased in conjunction with the addition of constraints for inferring 3D structure from the image. This result suggests that it is not only qualitative differences in part shape or spatial relations that influence view-based recognition. In particular, any information that is stable over rotations in depth is likely to improve recognition performance.

Relations between features

As noted earlier, the experiments presented here focus on how variations in part shape affect recognition. This is also true for the experiments reported by Biederman and Gerhardstein (1993), some of those reported by Liter (1995), and those reported by Liu et al. (1995). One criticism of each of these studies is that apparent effects of viewpoint may be due to a failure to vary the spatial relations between parts. There are several specific points that address this critique.

First, as demonstrated by Liter, qualitative differences in spatial relations did enhance overall recognition performance, but *did not* diminish the degree of viewpoint dependency. Based on this result, it seems clear that the relations between features or parts, regardless of their shape, play a crucial role in visual recognition (Hayward & Tarr, 1995; Hoffman & Richards, 1984), but, as with qualitative differences in part shape, do not lead to viewpoint invariance.

Second, exactly how qualitative differences in features and in relations interact remains a to-be-answered question. Although Biederman's GSD approach includes qualitative relations between parts (Biederman, 1987; Hummel & Biederman, 1992), these relations are not based on any specific principles and, as such, are relatively free parameters. Moreover, with regard to the present experiments, it is assumed that any model incorporating qualitative relations could encode the minimal case of the simple ABC ordering relationships present in the stimuli used here (Tarr & Pinker, 1990). Given that the part ordering within the structural description is not ambiguous (i.e., ABC is not confused with BCA), differences in part shape within this ordering should be sufficient for recognition. Such is the case for Hummel and Biederman's (1992) neural-net simulation of Geon-based recognition based on their inclusion of relations such as "beside," "above," and "below." This is particularly true in that no two objects in any of our stimulus sets contained the exact same subset of parts. Therefore, even given some confusion by a Geon-based description system in the ordering of parts, there should be sufficient information to uniquely distinguish each object.

Third, it should be pointed out that Biederman's experimental work (Biederman, 1987; Biederman & Cooper, 1991; Biederman & Gerhardstein, 1993) has omitted any systematic examination of the relations between parts or features. Instead, empirical support for the GSD approach has focused almost exclusively on demonstrating that differences in Geon shape form the basis for viewpoint-invariant recognition. The experiments presented here were designed to address these specific manipulations.

Conclusions

We tested visual recognition across changes in viewpoint in four conditions: one similar to Bülthoff and Edelman's (1992) paperclip experiments, one similar to Biederman and Gerhardstein's (1993) experiment in which they inserted unique Geons into paperclips, and two conditions in which we inserted additional distinct parts into paperclips. The additional-parts manipulation provided a method for investigating why viewpoint invariance was found in Biederman and Gerhardstein's experiment, as well as a test of the predictions of the View-Based and the GSD theories. We find that:

- Recognition is typically view-based.
- Viewpoint invariant recognition may be explained by local diagnostic features.
- The results of earlier experiments with homogeneous stimulus classes generalize to stimuli that are qualitatively discriminable.
- View-based mechanisms are sensitive to distinctive shape features.

In summary, we find little evidence to support the currently popular GSD approach (Biederman, 1987; Biederman & Gerhardstein, 1993), but strong evidence for View-Based theories of recognition (Bülthoff & Edelman, 1992; Humphrey & Khan, 1992; Poggio & Edelman, 1990; Tarr, 1995). Indeed, there is a remarkable uniformity of results pertaining to how observers recognize objects over changes in viewpoint – in almost every instance, across a wide range of tasks and stimuli, recognition has been found to be viewpoint dependent. The findings presented here lend further generality to this claim and illustrate some of the properties of view-based representations – in the present case, that qualitative differences in shape do not typically result in viewpoint invariance, but are part of the representation in that they influence overall recognition accuracy and the degree of viewpoint dependence.

References

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

Biederman, I. & Bar, M. (1995). One-shot viewpoint invariance with nonsense objects. In *36th Annual Meeting of the Psychonomic Society* Los Angeles, CA.

Biederman, I. & Cooper, E. E. (1991). Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, *20*, 585–593.

Biederman, I. & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(6), 1162–1182.

Biederman, I. & Gerhardstein, P. C. (1995). Viewpoint-dependent mechanisms in visual object recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1506–1514.

Bülthoff, H. H. & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Natl. Acad. Sci. USA*, *89*, 60–64.

Bülthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain?. *Cerebral Cortex*, *5*(3), 247–260.

Cohen, D. & Kubovy, M. (1993). Mental rotation, mental representation, and flat slopes. *Cognitive Psychology*, *25*(3), 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, and Cognition*, *18*, 43–57.

Edelman, S. & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, *32*(12), 2385–2400.

Ellis, R. & Allport, D. A. (1986). Multiple levels of representation for visual objects: A behavioural study. In A. G. Cohn & J. R. Thomas (Eds.), *Artificial intelligence and its applications*, pp. 245–247. New York: Wiley.

Hayward, W. G. & Tarr, M. J. (1995). Spatial language and spatial representation. *Cognition*, *55*, 39–84.

Hoffman, D. & Richards, W. (1984). Parts of recognition. *Cognition*, *18*, 65–96.

Hummel, J. E. & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*(3), 480–517.

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

Liter, J. C. (1995). *Features affecting orientation-invariant recognition of novel objects*. Ph.D. thesis, University of California, Irvine, CA.

Liu, Z., Kersten, D., & Knill, D. C. (1995). Structural organization improves object discrimination. Unpublished Manuscript, University of Minnesota.

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

Marr, D. & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Philosophical Transactions of the Royal Society of London, B, 200*, 269–294.

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Science*, *9*, 441–474.

Perrett, D. I., Oram, M. W., & Wachsmuth, E. (1996). Evidence accumulation in cell populations responsive to faces: An account of generalisation of recognition without mental transformations. Unpublished Manuscript, St. Andrews University, Scotland.

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

Snodgrass, J. G. & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34–50.

Tarr, M. J. (1995). Rotating objects to recognize them: A case study of the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin and Review*, *2*(1), 55–82.

Tarr, M. J. & Bülthoff, H. H. (1995). Is human object recognition better described by geon-structural-descriptions or by multiple-views?. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1494–1505.

Tarr, M. J., Hayward, W. G., Gauthier, I., & Williams, P. (1994). Geon recognition is viewpoint dependent. In *35th Annual Meeting of the Psychonomic Society* St. Louis, MO.

Tarr, M. J. & Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame?. *Psychological Science*, *1*(42), 253–256.