Journal of Experimental Psychology:
Human Perception and Performance
1995, Vol. 21, No. 6, 1506-1514

Viewpoint-Dependent Mechanisms in Visual Object Recognition: Reply to Tarr and Bülthoff (1995)

Irving Biederman
University of Southern California

Peter C. Gerhardstein
Rutgers University

I. Biederman and P. C. Gerhardstein (1993) demonstrated that a representation specifying a distinctive arrangement of viewpoint-invariant parts (a geon structural description, [GSD]) dramatically reduced the costs of rotation in depth. M. J. Tarr and H. H. Bülthoff (1995) attempt to make a case for viewpoint-dependent mechanisms, such as mental rotation. Their suggestion that GSDs enjoy no special status in reducing the effects of depth rotation is contradicted by a wealth of direct experimental evidence as well as an inadvertent experiment that found no evidence for the spontaneous employment of mental rotation. Their complaint against geon theory's account of entry-level classification rests on a mistaken and unwarranted attribution that geon theory assumes a one-to-one correspondence between GSDs and entry-level names. GSDs provide a representation that distinguishes most entry- and subordinate-level classes and explains why complex objects are described as an arrangement of viewpoint-invariant parts.

Consider the nonsense object in Figure 1. When first viewed, how did the reader know that the object was one never encountered previously? Why was the reader fairly confident that he or she would know what the object would look like if rotated 30°? The large central block would still look like a block and the vertical cylinder and wedge on top of the block would still be on top of the block. The zigzag cross brace connecting the tilting cylinder (ending in a cone) to the wedge would still enjoy the same relation if rotated 30°. These words denoting parts and relations are easily matched to the corresponding regions of the image.

Geon theory (Biederman, 1987; Hummel & Biederman, 1992) seeks to account for these readily evident capacities and characteristics of human object recognition by positing that objects are represented as an arrangement of simple viewpoint-invariant parts (geons) and relations, termed a geon structural description (GSD). The resultant viewpoint-invariant representation is designed to account for many of the entry-level shape-based classifications, such as distinguishing between a chair, an elephant, and a frying pan. The theory also provides an account of the vast majority of subordinate-level classifications that people readily make when they distinguish, for example, a square table from a round one or a four-legged chair from a swivel or a rocking chair.

Biederman and Gerhardstein (1993) presented evidence showing that when objects differ in their GSDs, recognition from a novel viewpoint can be readily achieved, as long as the novel viewpoint would activate the same GSD (i.e., as long as the same viewpoint-invariant parts and relations among these parts were apparent in the image). If a set of stimuli could not be distinguished by their GSDs, as would occur with a set of bent paper clips differing only in the angles between their segments (such as those used by Bülthoff & Edelman, 1992, and Edelman & Bülthoff, 1992) or a set of blocks all at right angles to each other and differing primarily in length and attachment direction (as in Tarr's objects, 1989), strong viewpoint dependency would be expected. To the extent that rotation in depth partially changed the GSD, as would occur, for example, if some of the parts were occluded and others revealed, weaker activation of the original unit representing the object would occur and consequently some cost in recognition performance would be expected (Biederman, 1987; Hummel & Biederman, 1992).

Tarr and Bülthoff (1995) take issue with Biederman and Gerhardstein's (1993) position, preferring an account that assigns a central role to *viewpoint-dependent mechanisms.* According to this account, participants learn a representation of the image projected by an object when viewed at its particular orientation. If an object is viewed from a new orientation, a mechanism (e.g., mental rotation, interpolation, extrapolation) is used that incurs a cost proportional to the angular disparity between original and tested views.

Unfortunately, Tarr and Bülthoff (1995) do not commit

Irving Biederman, Department of Psychology, University of Southern California; Peter C. Gerhardstein, Department of Psychology, Rutgers University.

This research was supported by a grant from the U.S. Army Night Vision and Electronic Sensors Directorate Army Research Office DAAH04-94-G-0065. We express our appreciation to Eric E. Cooper, Moshe Bar, and József Fiser for their helpful comments on this article.

Correspondence concerning this article should be addressed to Irving Biederman, Department of Psychology, Hedco Neuroscience Building, MC 2520, University of Southern California, Los Angeles, California 90089-2520. Electronic mail may be sent via Internet to ib@rana.usc.edu.

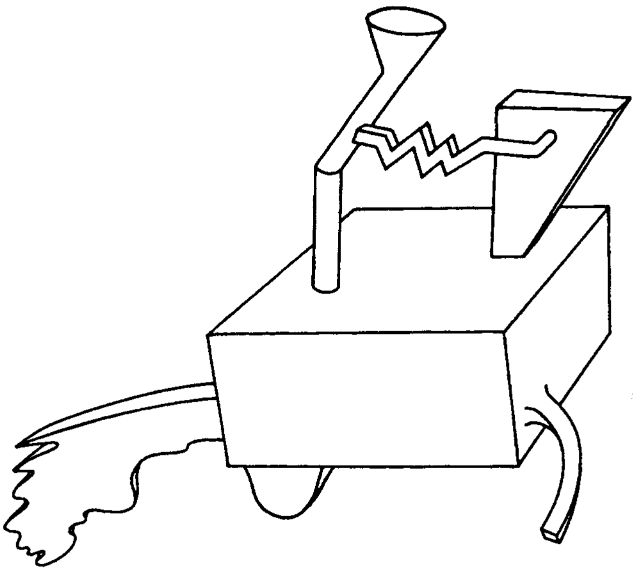

Figure 1. A nonsense object from Biederman (1987). People readily describe this object in terms of its simple parts and relations. There is strong intuition that it could be readily distinguished from most other objects even when viewed from a novel orientation in depth. From "Recognition-by-Components: A Theory of Human Image Understanding" by I. Biederman, 1987, *Psychological Review, 94,* p. 116. Copyright 1987 by the American Psychological Association.

themselves to any alternative representation to the one proposed by Biederman and Gerhardstein (1993)—hence the vague term in the previous paragraph "a representation of the image"—but the experiments they have performed, the analyses they have advanced, and the theories they favor (e.g., Bülthoff & Edelman, 1992; Poggio & Edelman, 1990) assume no special role for either part-based recognition or viewpoint-invariant properties. For example, Poggio and Edelman (1990) assumed that a bent paper clip can be represented as the centroid of a cloud of points, where each point marks one of the vertices of the clip. Not even the order of the points is modeled, despite the fact that a vastly different clip configuration would be produced by connecting the points in a different order. If applied to the nonsense object shown in Figure 1, the representation would be the centroid of a cloud of points where each point marks an extremum of curvature of the contours of the object.

Without a specification of a representation, Tarr and Bülthoff's (1995) advocacy of viewpoint dependency offers little more than the alternative to extrasensory perception and is not inconsistent with the position of Biederman and Gerhardstein in that any perceptual theory would have to assume that what is represented is determined by the image. The issue is, then, the extent to which experience with one view of an object can generalize to another view. What was apparent in Biederman and Gerhardstein's review of the literature was that different sets of stimuli and tasks could produce dramatically different magnitudes in the effects of rotation in depth. To account for this variability, Biederman and Gerhardstein proposed three conditions for viewpoint invariance.[1] The conditions essentially hold that strong generalization from one view to another will be manifested to the degree to which a set of stimuli has distinctive (different) GSDs and that the new view activates the same (or a highly similar) GSD as activated by the original view. Whereas the rotation costs for stimuli that fail the criteria, when expressed as rotation speeds (with lower speeds indicating higher costs), are in the order of 100–200°/s (Tarr, 1989), those that meet the criteria show costs in the range of 2,000–5,000°/s (Biederman & Gerhardstein, 1993).[2] This order of magnitude reduction in costs, due to the availability of distinctive GSDs, is actually an underestimate because typically participants have to be familiarized with stimuli that do not pass the criteria. Indeed, without such training they may show near chance recognition accuracy for rotations of such stimuli, as was evident in the Rock and DiVita (1987) experiment with bent wire constructions. With stimuli that meet the criteria, the high speeds of several thousand degrees per second are obtained instantly, without any familiarization with the stimuli. This striking interaction between rotation costs and the characteristics of a set of stimuli is the large effect in this domain, which all theories must address.

Because they do not propose a representation, Tarr and Bülthoff (1995) offer no general competing explanation on this critical point that the effects of rotation depend on the image characteristics of a set of stimuli. Tarr and Bülthoff describe a number of viewpoint-dependent *effects,* but these are precisely those effects that Biederman and Gerhardstein (1993) cited as having failed the conditions for viewpoint invariance (usually because the set of objects did not project

[1] Essentially the conditions were that for invariance to be manifested between two views of an object (a) the object may be capable of being represented as a specified arrangement of viewpoint-invariant parts (GSD condition), (b) each member of a set of stimuli must activate a different GDS (distinctive GSD condition), and (c) the different views must activate identical GSDs (identical GSDs over views condition). Instead of "viewpoint invariance," it would be more accurate to use the term "aspect invariance" (in the sense of Koenderink and van Doorn's 1976, definition of aspect) to describe the expected performance that met these conditions.

[2] There can be a weak and a strong form of the invariance implied by the conditions for invariance. Other things being equal, the weak form is that the performance costs (of reaction times [RTs] or errors) of rotation in depth for stimuli that meet these conditions will be smaller than for stimuli that do not meet these conditions. The reduction of costs from a few hundred degrees per second to several thousand degrees per second for stimuli that do and do not meet the criteria, respectively, supports this weaker form. The strong form would hold that the costs for stimuli that meet the conditions would be zero. That is, the function relating costs to rotation angle should have zero slope. How much of a departure from zero slope, in part for reasons discussed in the third section, would qualify as strong invariance has yet to be resolved. Specifically, to what extent can several thousand degrees per second as the cost of rotation with stimuli that meet the criteria be regarded as equivalent to a zero slope? This distinction and problem is similar to the one made between parallel (or preattentive) "pop out" and serial (or attentive) processing in search tasks.

distinctive GSDs as with a set of bent paper clips) or ones that produce a different GSD because of self-occlusion of an object's parts. Put simply, objects that differ in their GSDs will show much smaller effects of rotation in depth than objects that do not.

What is implied by Tarr and Bülthoff's (1995) critique—that there is no special status to viewpont-invariant properties (except under highly restricted conditions) or parts-based representation—is thus not consistent with the data nor with observations about everyday object recognition. Tarr and Bülthoff also take issue with the adequacy of GSD theory as a general account for entry-level and subordinate-level recognition. Much of this critique rests on a mistaken attribution that GSD theory assigns a one-to-one mapping of GSDs to entry-level classes. Consequently, despite the length of Tarr and Bülthoff's commentary and the vigor of its assertions, with closer analysis there is virtually nothing in it that challenges the major conclusions of Biederman and Gerhardstein (1993).

We hasten to add that our critical analysis of Tarr and Bülthoff's (1995) commentary does not mean that we believe that GSD theory is a done deal with respect to object recognition in general and depth-rotation effects in particular. GSD theory, for starters, is certainly incomplete, and one's best guess is that details of any current implementation will certainly be incomplete, if not wrong, as they would be for any model of such broad scope. (In fairness to Tarr & Bülthoff, we note that the viewpoint-dependent models can also undergo development and evolution). Indeed, Hummel and Biederman (1992) listed a number of cases where their implementation failed on simple images. Thus, a full implementation that would achieve automatic object classification from gray level images is still a distant goal (as it is for any theory). However, what we can do is consider the general assumptions of GSD theory and compare them against the alternative viewpoint-dependent proposals that would deny any special status to viewpoint-invariant contour differences or parts-based representation in favor of a metric representation of the stimulus sensitive to the length, angle, and degree of curvature of the contours.

GSDs as an Account of Entry- and Subordinate-Level Classification

Tarr and Bülthoff (1995) argue that GSD theory's account of entry-level shape classification has difficulty in explaining those cases where (a) different GSDs would be activated by different subordinate-level instances, as would occur, for example, with a rectangular table and an elliptical table, and (b) members from different entry-level classes are not well distinguished by different GSDs, as would occur, for example, with a coyote and a jackal. However, it is Tarr and Bülthoff who have erroneously attributed to GSD theory a one-to-one correspondence between different entry-level *names* and different GSDs. GSD theory provides an account of which shape distinctions will be easy and viewpoint *invariant and which will* be difficult and viewpoint dependent. In a great many cases, but not all, entry-level class boundaries define different GSDs. In many cases there are GSD differences among members of a subordinate class that, for reasons such as common functionality, a given culture has not defined as a different entry-level class. Nonetheless, GSD theory makes a clear prediction as to where the entry-level boundary would be placed if the boundary were extended to distinguish among subordinate-level members: When the differences in GSDs and aspect ratios are equated for physical similarity,[3] the boundary would be placed between different GSDs. Moreover, it is much easier to distinguish the rectangular tables from the elliptical ones at new orientations in depth than to distinguish among tables differing in aspect ratio (Biederman & Bar, 1995).

With respect to the latter claim of Tarr and Bülthoff (1995), that different entry-level classes may not lead to different GSDs, as would occur with similar shaped animals such as a jackal and a coyote, to our knowledge these are precisely the cases where people (and monkeys) have difficulty in making the classifications (Gaffen & Heywood, 1993). Indeed, neurological lesions whose effects were previously thought to be semantic in that they caused confusions among living things have been shown to be a consequence of the high interclass perceptual similarity of living things (Gaffen & Heywood).

When will we be faster at knowing the difference, say, between a cow and a horse? From the perspective of GSD theory, it will be when the distinctive parts are in view, such as horns and udders. GSD theory thus offers one basis for deriving which views of an object might be canonical (Palmer, Rosch, & Chase, 1981). Objects made up of only a single part are actually quite difficult to identify (Biederman, Hilton, & Hummel, 1991). Often they require the addition of small features and texture (and color), as in distinguishing among a peach, plum, and nectarine.

In not assuming a one-to-one correspondence between GSD and entry-level name, GSD theory can thus provide an account of the considerable variability that is evident in human performance in distinguishing among entry-level and subordinate-level classes.

Tarr and Bülthoff's Use of Unique Features as an Alternative to GSDs

To demonstrate the power of a single distinctive GSD to actually confer viewpoint invariance to a set of stimuli, Biederman and Gerhardstein (1993, Experiment 5) modi-

[3] How can the differences in GSDs be equated for physical similarity to differences in aspect ratio? The scaling can be done both behaviorally and theoretically: behaviorally, by the data from a same–different task for physical identity; theoretically, by a model of wavelet similarity that approximates the multiscale and multiorientation activation of hypercolumns of V1 simple cells. Cooper and Biederman (1993) found that aspect ratio differences that were slightly larger according to both methods of scaling were far less salient than geon differences in a same–different task for judging whether two images had the same name (and entry level concept).

fied a set of bent paper clip stimuli that had been used by Edelman, Bülthoff, and Weinshall (1989) by substituting a different geon for one of the segments in each of 10 bent paper clip objects, so that the middle segment (of five) was a brick in one, a cone in another, etc. Given one such "charm bracelet" as a target studied at one orientation, participants could readily distinguish that object at new orientations in depth from all others in the set. Without this distinguishing viewpoint-invariant difference, it had been shown that it was extraordinarily difficult to learn to distinguish one clip from the others at new orientations (Edelman, Bülthoff, & Weinshall, 1989).

Tarr and Bülthoff (1995) argue for the use of unique features to account for this result. The principle holds that viewpoint invariance would be expected if a restricted and known set of objects could be discriminated by a single viewpoint-invariant feature (or a few such features). Their account is essentially equivalent to the account offered by Biederman and Gerhardstein's (1993) GSD theory for their experiment in which the objects differed by only a single geon. Tarr and Bülthoff argue that because a single feature or a small number of features would be inadequate for distinguishing among a large number of object classes, some other basis for shape classification—presumably one that is viewpoint dependent—must be invoked.

The appeal to viewpont-invariant features as embodied in the use of unique features is a striking departure from Tarr and Bülthoff's (1995) experiments, theories, and analyses, none of which specified such features. To the extent that such features would be present in everyday shape classifications, there would then be little or no need to invoke the viewpoint-dependent representations that they champion.

Tarr and Bülthoff (1995) claim that unique features (a) are insufficient for distinguishing among the vast majority of entry- and subordinate-level shape classifications that we typically make, (b) are not present in everyday shape classifications, and (c) are only specified for a familiar and known set of stimuli. With respect to sufficiency, although a single feature would be inadequate for distinguishing among a large number of object classes, Tarr and Bülthoff's declaration that GSD theory is indistinguishable from a routine that used unique features—and hence also inadequate as an account of familiar classification—loses much of its force once we consider that a small number of viewpoint-invariant features *in specified relations* could very well be sufficient for distinguishing among familiar classes. Specifically, if the features were a viewpoint-invariant characterization of parts of the object and if their invariant relations were specified (i.e., if they constituted a distinctive GSD), then two or three viewpoint-invariant parts in only a few classes of specified relations could be sufficient for representing over a billion object classes on the basis of their shape (Biederman, 1987). The critical distinction here between Tarr and Bülthoff's principle of unique features and GSD theory is that the latter specifies the arrangement of invariant features into parts and invariant relations among parts, rather than just a single feature or a few features.

Do we have to know the set of to-be-distinguished entities to use viewpoint-invariant features? Can we only use one (or just a few) feature(s) at a time? Consider Figure 2. Is it the same as Figure 1? We suspect that most readers would have little trouble in realizing that these are different objects. The use of unique features would hold that the routine itself either would not have been invoked when looking at Figure 1, because the viewer did not know what other objects were in the set of to-be-distinguished objects, or it would have been represented by a few simple features, such as a straight line, a curve, and an L vertex, in which case it would have been extremely difficult to appreciate that Figure 2 was not Figure 1.

In arguing for the use of unique features, Tarr and Bülthoff (1995) appear to have erroneously identified Biederman and Gerhardstein's (1993) demonstration that only a single distinctive GSD was sufficient for viewpoint invariance (in their Experiment 5 that added a different geon to each of 10 bent paper clip stimuli) with the maximal power of GSDs to characterize the difference among objects. Everyday object classes often differ in many parts and relations, and such differences are readily expressed as different GSDs. In our judgment, the differences between the objects shown in Figure 1 and 2 are not uncharacteristic of the variability in object shapes found between randomly selected familiar object classes. As a general system of object representation capable of expressing the perceptually salient shape characteristics of novel objects, GSD theory provides an account of entry-level classification that is highly compatible with Tversky and Hemenway's (1984) observation that entry-level shape concepts are to a large extent expressed in terms of part representations. These part representations, according to Tversky and Hemenway, also account for the functionality of many of those classes.

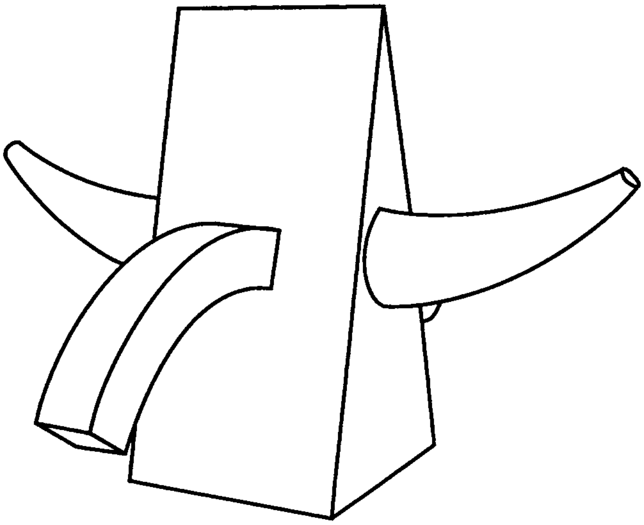

Figure 2. A nonsense object (a Viking metronome). Is this object the same as the one shown in Figure 1? Most people readily distinguish the difference between the two objects even though the two objects share many features and the reader, when viewing Object 1, could not anticipate how it would differ from Object 2.

Tarr and Bülthoff (1995) are essentially arguing that viewpoint-invariant differences are only important with a restricted and known set of stimuli, but that viewpoint-dependent representations are involved in everyday object recognition. We would argue the opposite. To the extent that there is a contribution from viewpoint-dependent aspects of an experience, the employment of such information would only be useful with a restricted set of instances. If a stapler projects an elongated and 1.5° image to the left of fixation, it is not clear that elongation, 1.5°, and to the left can be used as general attributes for the identification of staplers.

The Biederman and Gerhardstein (1993) investigation was not designed as a test of whether objects are represented as an arrangement of invariant parts. In claiming that their theory of unique features was indistinguishable from GSD theory, Tarr and Bülthoff (1995) ignore the considerable evidence for part-based, rather than feature-based or global shape, representations. For example, the brief presentation of a member of a complementary pair of contour-deleted images of an object, where each member shared no features (lines or vertices) in common with the other member but did share the same parts, primed the other member as strongly as it primed itself (Biederman & Cooper, 1991b). That is, object priming is predicted by overlap of GSDs, not image features. Biederman's earlier (1987) experiment with contour-deleted images, where the parts could or could not be recovered, showed that if the geons could not be recovered, recognition was virtually impossible. The nonrecoverable images in that study had plenty of contour that could have served as input to the viewpoint-dependent models that Tarr and Bülthoff favor. Cooper and Biederman (1993) showed that differences in nonaccidental properties of an object's part (yielding different geons) were much more salient than differences in aspect ratio in an object classification task (but not in making same–different judgments of physical identity), indicating that it is the nonaccidental properties that are important in object classification, as assumed by geon theory. None of these effects are handled by the class of theories that Tarr and Bülthoff propose nor were they addressed in their commentary.

The Interpretation of Viewpoint-Dependent Effects

Tarr and Bülthoff (1995) raise the reasonable question as to how stimuli are ultimately recognized when they do not meet the conditions for viewpoint invariance. The fundamental issue here is whether one has to posit a representation that does not assume GSDs, such as a template. Biederman and Gerhardstein (1993) listed three possible reasons why viewpoint-dependent effects might be found, especially with stimuli that do not meet their conditions for viewpoint invariance, which we will elaborate on here. All of them seem much more plausible than the metric templates theories favored by Tarr and Bülthoff.

Viewpoint-Dependent Processes, Not Representations

First, a participant may use a process (not a representation) that is viewpoint dependent, such as searching for a distinctive GSD at a small scale. The account of chicken sexing offered by Biederman and Shiffrar (1987); or the use of the logo or printed name to classify the make of a car would be examples of such a process. The process (viz., searching) that confirms or finds a GSD in this case could be viewpoint dependent, but the resultant representation could be viewpoint invariant. It is the initial search that would be producing the rotation costs, not the representation. Presumably, one could instruct a naive participant as to the where and the what of the critical information and the rotation costs should be dramatically reduced, if not eliminated. Of course, this would constitute a kind of unique feature, except that the particular feature that is used is dependent on an initial classification of the stimulus. We do not look for car logos in trying to distinguish an African from an Asian elephant. Some of these cases (e.g., a small or accidental GSD) would presumably be represented in a manner that would be similar to a GSD that did not require search or discovery, but the processes that led to their discovery—search or mental rotation—that would be viewpoint sensitive are not part of the representation of shape. It is possible that the marked effects of rotation of Tarr's (1989) block stimuli were produced by such search.

Accidental GSDs

Second, for stimuli that invite accidents, like bent paper clips, one might use distinctive GSDs that would hold for a small range of viewpoints, as described by Biederman and Gerhardstein's (1993) characterization of the recognition of the bent paper clip stimuli, except by definition (because they are accidents) more GSDs would be needed over the different aspects, yielding higher rotation costs. For example, a given bent paper clip could look like a W at one orientation, two loops at another, and a teepee at a third. Each of these representations would have to be mapped (learned) onto the same object unit in the Hummel and Biederman (1992) neural net. Large view-specific effects of learning should then be evident, as demonstrated, for example, by Tarr (1989) and Edelman and Bülthoff (1992). Over rotation angles where the same qualitative description could distinguish the stimulus from the other possible stimuli (e.g., as long as it looked like a W), little or no effect of rotation would be expected. From this perspective, the smooth monotonic increase in recognition costs (in reaction time or error rates) as a function of rotation angle shown in many reports of viewpoint dependency are an artifact of averaging over rotations where there was little or no cost (because the GSD did not change) with rotations in which there was a large cost (because the GSD did change).

Fine Metric Coding: A Dorsal System Function?

Third, a representation that was sensitive to metric variation and complex relations could be used, the account favored by Tarr and Bülthoff (1995), as with the model proposed by Poggio and Edelman (1990), which self organizes object units to the outputs of early filters. In only this last case would the representation of the stimulus be fundamentally viewpoint dependent, in the sense that the representation is inextricably tied to the image such that if played back it could reproduce the image, with no decomposition into parts and no priority given to viewpoint invariant properties. We are not sure that any case of entry-level access requires such representation. Moreover, the vast majority of subordinate-level classifications, such as the difference between round and square tables, with the exception of the special case of face individuation, would not appear to require such a representation either. Whether object classifications that required the specification of viewpoint-dependent information are represented in a manner equivalent to those that can be distinguished by GSDs is not definitively known. The conjecture of Biederman and Cooper (1992) was that they are not. Biederman and Cooper argued that viewpoint-dependent representations for pose information, as well as size, reflection, and position information, are likely specified in the dorsal pathway, an interpretation consistent with that advanced by Kosslyn (1994). Insofar as this information is sufficiently available to affect old–new judgments (e.g., Cooper, Biederman, & Hummel, 1992), it is not impossible that it might be used on some trials in an object recognition task, particularly when it would be difficult to use distinctive GSDs.

Small Effects of Rotation With Stimuli That Possess Distinctive GSDs

Tarr and Bülthoff (1995) wish to make a case for viewpoint-dependent mechanisms to account for the small (and nonsignificant, $Fs < 1.00$) effect of rotation with the nonsense objects in the no-parts-change condition of Biederman and Gerhardstein's (1993) Experiment 3. (The objects resembled the one shown in Figure 2 and, though the members of the set were highly similar to each other, the set met the conditions for viewpoint invariance.) Biederman and Gerhardstein (1993, p. 1180) had proposed that such small effects could have been the consequence of uncontrolled partial foreshortening or occlusion or an occasional employment of a strategic variant that was viewpoint dependent.[4]

With respect to partial foreshortening, from the perspective of geon theory, a distinctive GSD allows invariance as long as the information activating it can be resolved (in the allowed time). But no account would hold that the effect should be all-or-none: With rotation in depth that produces self-occlusion or extreme foreshortening, obviously that information would be more and more difficult to resolve. With the highly similar stimuli in Biederman and Gerhardstein's (1993) Experiment 3 (the 10 unfamiliar five-part objects had identical relations and only 10 different geons for the 50 parts), rotation costs could then be expected. Most entry-level classes with complex objects have several differences from their near neighbors so rotation in depth would result in much less of a cost. A full account of object recognition, as noted by Biederman and Gerhardstein, would necessarily have to include scale effects.

If uncontrolled partial foreshortening or occlusion could account for the small effect in Biederman and Gerhardstein's (1993) Experiment 3, its inclusion as evidence for a viewpoint-dependent mechanism would render such a mechanism too general to have any empirical content. Imagine an analogous experiment on translation invariance that resulted in an effect of translation when an object was moved from a lighted area into a shaded area so it could not be adequately resolved. Would Tarr and Bülthoff (1995) argue for a translation-dependent mechanism in such a case? This extreme example is presented to indicate that (a) Tarr and Bülthoff assume that any effect of rotation is evidence for a viewpoint-dependent mechanism and (b) that they have not committed themselves to any particular mechanism. To the extent that participants might occasionally use some viewpoint-dependent strategy in this experiment, its effects appear to be small relative to the enormous gain from having distinctive GSDs.

Whatever the argument for a viewpoint-dependent mechanism that Tarr and Bülthoff (1995) wish to invoke, they would have to accommodate the other documented invariances for size, translation, reflection, and contour features in object naming. They would also have to account for why these same variables exert large effects on old–new recognition memory. GSD theory readily accounts for both the invariances when naming (Hummel & Biederman, 1992) and, in assuming a role of the dorsal representations in familiarity-based judgments, why old–new recognition memory does not show the invariances characteristic of naming.

[4] Tarr, Hayward, Gauthier, and Williams (1994) recently reported larger rotation costs for stimuli that met the conditions for invariance than those reported by Biederman and Gerhardstein (1993). Tarr et al.'s overall RTs were dramatically longer than those of Biederman and Gerhardstein's. In a go/no-go task for verifying single geons, for example, whereas the mean RTs for Biederman and Gerhardstein's participants were approximately 300 ms, the RTs for Tarr et al.'s participants were over 450 ms. Biederman and Gerhardstein's participants did have higher false-alarm rates and speed and accuracy feedback after every trial. It is likely that the speed feedback induced Biederman and Gerhardstein's participants to forego the fine distinctions required to distinguish near-accidental views of some of the distractors from the target in that the false alarms in their experiment primarily came from stimuli that were highly similar to the target. As described in the previous section, the search for distinguishing features in a rotated view might have resulted in the higher slope in the Tarr et al. study. Whatever the rotation costs are for stimuli that do differ in GSDs, they are dramatically smaller than the rotation costs for stimuli that differ only in viewpoint-dependent properties when performance on the two types of stimuli is evaluated under uniform conditions (Biederman & Bar, 1995).

Viewpoint-Dependent Effects in the Recognition of Everyday Objects

Tarr and Bülthoff (1995) attempt to explain away Biederman and Gerhardstein's (1993) finding for viewpoint invariance in the perception of everyday objects as a case of "robust generalization" from previously experienced viewpoints. But the actual images used in Biederman and Gerhardstein's experiments were never seen by the participants prior to the experiment, yet instant viewpoint invariance in their visual priming (vs. name or concept priming) was obtained. It would be interesting to see a demonstration of how the kinds of models favored by Tarr and Bülthoff that do not posit viewpoint-invariant properties cope with these kinds of object classes. In any event, robust generalization is not invariance. Actually, there is strong evidence against the kind of model priming appealed to by Tarr and Bülthoff: An image composed of half of an object's parts does not prime an image containing the other half of that object's parts (Biederman & Cooper, 1991b).

Perhaps Tarr and Bülthoff (1995) have painted themselves into a corner in that the (never-explicated) viewpoint-dependent model shows perfect generalization over highly different shape classes (such as different kinds of chairs) but is completely incapable of generalizing from one view of a bent paper clip to another!

Tarr and Bülthoff (1995) argue for a slight effect of viewpoint in Biederman and Gerhardstein's (1993) Experiments 1 and 2 with novel stimuli from familiar classes. But the argument fails for three reasons. First, the small effect of viewpoint when the effect of the different exemplar condition is subtracted out is dramatically smaller than anything observed with stimuli that fail the conditions for viewpoint invariance. From their own Figure 1, the slope from Experiment 1 is approximately 8,000°/s! In their plot of the data from their Experiment 2, Tarr and Bülthoff ignore Biederman and Gerhardstein's partition of that data according to whether GSDs did or did not change over rotation. Even without that partitioning, the resulting slope is approximately 1,800°/s. Any correction for a speed–accuracy tradeoff in Experiments 1 and 2 would have resulted in still smaller effects of rotation in that in both experiments the error rates declined with rotation. Note that the invariance here is immediate. Even with extensive prefamiliarization, nothing close to these slopes was obtained in all those experiments where the set of stimuli fail the conditions for viewpoint invariance.

Second, Tarr and Bülthoff (1995) argue that the minuscule effect of rotation in Biederman and Gerhardstein's (1993) Experiment 1 (with familiar object classes) may be a result of the 135° rotation value producing images similar to what would be produced with mirror reflection and that a number of studies have demonstrated invariance over mirror reflection. We agree. It is interesting that Tarr and Bülthoff did not draw the implications of such an effect for their viewpoint-dependent account as reflection invariance would appear to pose a serious challenge to their viewpoint-dependent mechanism (all the models of which assume a coordinate space rather than a structural description). Invariance over mirror symmetry was readily obtained in the Hummel and Biederman (1992) network by simply not specifying left–right. If they want to appeal to a symmetry computation such as that proposed by Vetter, Poggio, and Bülthoff (1994), they must wrestle with the difficult problem of accounting for why rotation effects are ever obtained with symmetric or near symmetric objects.

Third, Tarr and Bülthoff (1995) ignore the possible resolution effects discussed previously. These effects would increase and then decrease as the rotation approached 180° (for bilaterally symmetric stimuli), consistent with the striking degree of reflection invariance.

An Inadvertent Test of the Use of Viewpoint-Dependent Mechanisms

Do people spontaneously use mental rotation to determine object equivalence in everyday activities, as has been claimed by Tarr and Bülthoff (1995)? An inadvertent natural experiment allowed a test of this claim. Figure 3 is reprinted from the original Biederman and Gerhardstein (1993) article.

This original caption and text were wrong. These are three different objects. This erroneous figure and caption were in the manuscript submitted for review as well as in a preprint sent to several hundred individuals. The article has been frequently cited and has been the subject of intense scrutiny and commentary by several groups of investigators. The same figure was also shown as a slide in approximately 50 colloquia and conference presentations, always with the incorrect description.

Not a single individual, to our knowledge, ever found the error. (The authors discovered their error when requesting approval for reprinting the figure. The copy editor even neglected to substitute a corrected figure that we submitted,

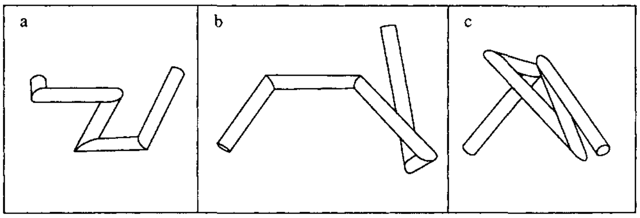

Figure 3. "Line drawings of three poses of an object like those in Figure 1 of Edelman, Bülthoff, and Weinshall (1989). (The poses differ by about 50°. This object has to be distinguished from nine others, all made up of the same set of five wires differing only in angle.)" That is what the published caption in Biederman and Gerhardstein (1993) read. The caption was wrong with respect to the originally published figure (presented above). These are three *different* objects. No one noticed. 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. 1165. Copyright 1993 by the American Psychological Association.

thinking that it was identical to the original. The correct figure was published in the succeeding issue of the journal as a correction and in the reprints.) Informal experiments reveal that when images do differ in viewpoint-invariant parts or relations individuals will never accept them as equivalent. Imagine someone showing you Figures 1 and 2 and asserting that the objects in the two images are identical but merely seen at different orientations. You would think such a claim is a joke. (No one ever laughed at the original error.) Different objects with the same viewpoint-invariant parts and relations are accepted as equivalent. We conclude that people do not spontaneously use viewpoint-dependent mechanisms to distinguish objects that differ only in viewpoint-dependent properties. It is not that people cannot do mental rotation or ultimately discriminate metric detail; it is just that they are satisfied that objects that have the same or highly similar GSDs may not generally be worth discriminating.

Differentiating Between GSD Theory and Other View-Restricted Theories of Recognition

We would agree with Tarr and Bülthoff's (1995) characterization of GSD theory as a class of aspect graph theories, but one that specifies a structural description, organizing viewpoint-invariant properties into an arrangement of parts. There can be enormous numbers of changes in the image features as an object rotates in depth, with little or no change in the GSD. To our knowledge, no previous account of aspect graph theory made these arguments for equivalence classes based on GSDs. It is not clear that just considerations of resolution scale would capture the contribution from the part level of organization to object representations. Moreover, there are considerable reductions in complexity in considering a part representation: The sum of the aspects for the aspect graphs for an object's individual parts has far fewer aspects than that for the complete object. The strong evidence for such part-based representations is derived from the Biederman and Cooper (1991b) and Cooper and Biederman (1993) investigations, which were never directly addressed by Tarr and Bülthoff. The complete invariance for complementary primed images in the Biederman and Cooper (1991b) experiment, where none of the features were common in the members of a complementary pair, poses a serious challenge to viewpoint-dependent accounts.

Conclusions

Despite our questioning of the role of viewpoint-dependent mechanisms in adult object recognition to solve problems of depth rotation, particular images must necessarily play a role in the development of object recognition capacities. There is ample evidence that adult neural connectivity could not have been genetically determined (Cherniak, 1994). Although genetics provides a rough scaffolding for determining what statistics of images are going to affect connectivity, the actual organization must be activity dependent. Assuming that object recognition capacity develops from viewing objects and scenes, that experience, by definition is viewpoint dependent. From robust statistics of such images, general viewpoint-invariant capacities might develop. Hummel and Biederman (1992) described a simulation in which a general capacity to group collinear and parallel contours developed from viewing random patterns that was then independent of the previously viewed patterns. The HyperBF network described by Poggio and Girosi (1990) might offer one general scheme for how viewpoint-dependent activity might ultimately organize units that have a general viewpoint invariant capacity, such as those in the hidden layers of the Hummel and Biederman model. In our opinion, how these viewpoint-invariant capacities for object recognition develop from the statistics of viewpoint-dependent experiences is one of the great challenges in the developmental neurobiology of object recognition. We emphasize that the general formalism proposed by Poggio and Girosi is not a theory of representation unless the characteristics of the units are specified.

In the past decade, at least five classes of results in human object recognition have accumulated that, in our opinion, any theory must address. These include

1. The 100% difference in median identifiability between recoverable and nonrecoverable contour-deleted images (Biederman, 1987).
2. The equivalence in priming observed between complementary pairs of images that have no vertices or edges in common (but do allow perception of the same parts) and the lack of any priming between images that do not share parts (Biederman & Cooper, 1991b).
3. The greater sensitivity of object representations to viewpoint–invariant than to metric differences (Cooper & Biederman, 1993 Biederman & Bar, 1995).
4. The invariance in priming to changes in size, translation, reflection, and rotation in depth (Biederman & Cooper, 1991a, 1992; Biederman & Gerhardstein, 1993).
5. The tendency for objects to be described as an arrangement of parts (Tversky & Hemenway, 1984).

Geon theory offers an account of these results. They remain a challenge to alternative theories.

References

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

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

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

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

Biederman, I., & Cooper, E. E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 121–133.

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

Biederman, I., Hilton, H. J., & Hummel, J. E. (1991). Pattern goodness and pattern recognition. In J. R. Pomerantz & G. R. Lockhead (Eds.), *The perception of structure* (pp. 73–95). Washington, DC: American Psychological Association.

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

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences, 89,* 60–64.

Cherniak, C. (1994). Component placement optimization in the brain. *Journal of Neuroscience, 14,* 2418–2427.

Cooper, E. E., & Biederman, I. (1993, November). *Geon differences during recognition are more salient than metric differences.* Poster session presented at the 34th annual meeting of the Psychonomics Society, Washington, DC.

Cooper, E. E., Biederman, I., & Hummel, J. E. (1992). Metric invariance in object recognition: A review and further evidence. *Canadian Journal of Psychology, 46,* 191–214.

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

Edelman, S., Bülthoff, H. H., & Weinshall (1989). *Stimulus determines recognition strategy for novel 3D objects* (A.I. Memo No. 4). Cambridge, MA: MIT Press.

Gaffen D., & Heywood, C. A. (1993). A spurious category-specific visual agnosia for living things in normal human and nonhuman primates. *Journal of Cognitive Neuroscience, 5,* 118–128.

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

Koenderink, J. J., & van Doorn, A. J. (1976). The singularities of the visual mapping. *Biological Cybernatics, 24,* 51–59.

Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate.* Cambridge, MA: MIT Press.

Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and performance IX* (pp. 135–151). Hillsdale, NJ: Erlbaum.

Poggio, T., & Edelman S., (1990). A network that learns to recognize 3D objects. *Nature, 343,* 263–266.

Poggio, T., & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science, 247,* 978–982.

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

Tarr, M. J. (1989). *Orientation dependence in three-dimensional object recognition.* Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge MA: MIT Press.

Tarr, M. J., & Bülthoff, H. H. (1995). Conditions for viewpoint dependence and viewpoint invariance. What mechanisms are used to recognize an object? *Journal of Experimental Psychology: Human Perception and Performance, 21,* 1496–1507.

Tarr, M. J., Hayward, W. G., Gauthier, I., & Williams, P. (1994, November). *Geon recognition is viewpoint dependent.* Paper presented at the 35th annual meeting of the Psychonomics Society, St. Louis, MO.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General, 113,* 169–193.

Vetter, T., Poggio, T., & Bülthoff, H. H. (1994). The importance of symmetry and virtual views in three-dimensional object recognition. *Current Biology, 4,* 18–23.

Received August 15, 1994
Revision received December 28, 1994
Accepted February 24, 1995 ■