Hummel, J. E. (2000). Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich and A. Markman (Eds.). *Cognitive Dynamics: Conceptual Change in Humans and Machines* (pp. 157 - 185). Hillsdale, NJ: Erlbaum.

Where View-based Theories Break Down:

The Role of Structure in Shape Perception and Object Recognition.

John E. Hummel

University of California, Los Angeles

Address correspondence to:
John E. Hummel
Department of Psychology
University of California, Los Angeles
405 Hilgard Ave.
Los Angeles, CA 90095-1563
jhummel@psych.ucla.edu

Preparation of this manuscript was supported by NSF Grant SBR-9511504. I am grateful to Steve Engel and Brad Love for their detailed and very helpful comments on an earlier draft of this manuscript.

Recent years have seen the growth of a movement in the object recognition community based on the idea that visual object recognition is mediated by the activation of template-like *views.* This chapter reviews the evidence and motivation for this view-based account of object representation, describes in detail the nature of this account, and discusses its logical and empirical limitations. I shall argue that view-based theory is fundamentally and irreparably flawed as an account of human shape perception and object recognition. I will then present the most plausible and common objections to my arguments and respond to each. Finally, I shall relate view-based theories of object recognition to other models based on formally similar types of knowledge representation, and conclude with a discussion of why such accounts must necessarily fail as an account of human perception and cognition.

Structural Descriptions, the Problem of Viewpoint, and the Motivation for the View-based Account

It is almost traditional to open papers about object recognition by noting that we can recognize objects in novel viewpoints, even though different views can project radically different images to the retina. One of the most important mysteries of human object recognition is the question of what makes this possible. One influential class of theories—*structural description* theories—holds that this capacity reflects the way our visual systems represent object shape (Biederman, 1987; Hummel & Biederman, 1992; Marr, 1980; Marr & Nishihara, 1978). The general idea is that we decompose an object's image into regions corresponding to volumetric parts (such as *geons*; Biederman, 1987), and then explicitly represent those parts in terms of their relations to one another. For example, a coffee mug might be represented as a curved cylinder (the handle) side-attached to a straight vertical cylinder (the body). The relations are critical: If the curved cylinder were attached to the top of the straight cylinder, then the object would be a bucket rather than a mug (Biederman, 1987).

Structural descriptions based on categorical parts and relations—for example, specifying the handle as simply "curved," rather than specifying its exact degree of curvature, and specifying the relation as "side-attached," rather than specifying the metric details of the attachment—permit recognition in novel viewpoints as a natural consequence (Biederman, 1987). Note that our description of a mug will remain the same as the mug is translated across the visual field, moved closer to or farther from the viewer, or rotated in depth (provided the handle does not disappear behind the body, an provided the mug does not appear in a view—such as directly end-on—that makes it impossible to perceive the shapes of the parts; see Biederman, 1987; Hummel & Biederman, 1992; Marr, 1982). Recognition based on this kind of description would likewise be unaffected by these changes in the mug's image. But rotating the image 90˚ about the line of sight, so that the body is horizontal and the handle is on top, will change the description, making recognition slower and/or more error-prone. Like human object recognition, our categorical structural description is sensitive to rotations about the line of sight, but insensitive to translation, scale, left-right reflection, and some rotations in depth (see Hummel & Biederman, 1992). Categorical structural descriptions are also useful because they permit generalization across members of a class (Biederman, 1987; Marr, 1980): "Curved cylinder side-attached to straight vertical cylinder" describes many different mugs, so representing mugs in this fashion makes it easy to recognize new ones.

Structural description theories have attracted a great deal of attention, both as accounts of human object recognition, and as approaches to object recognition in the machine vision literature. On both scores they have also been vigorously criticized. As researchers in computer vision have discovered, it is not easy to develop algorithms to generate structural descriptions from object images (but see Dickenson, Pentland & Rosenfeld, 1993, Hummel & Biederman, 1992, Hummel & Stankiewicz, 1996a, and Zerroug, 1994, for substantial progress in this direction). Deriving structural descriptions from images is difficult in part because the resulting descriptions can be exquisitely sensitive to the manner in which the image is segmented into parts: The same image, segmented in different ways, may give rise to very different descriptions (Ullman, 1989). This problem—although probably not insurmountable (see Hummel & Stankiewicz, 1996a)—has led

many researchers to doubt the plausibility of structural descriptions as an approach to shape representation and object recognition. (However, it is important to point out that, although we have yet to develop machine vision systems that can reliably segment gray-level images into objects and their parts, people have absolutely no difficulty doing so: It's not that segmentation is impossible; it's just that we do not yet fully understand how it is done.)

Behavioral data from experiments with human subjects have also been cited as evidence against structural descriptions in human object recognition. It is common (and mistaken) to assume that structural descriptions are completely object-centered and view-invariant (see, e.g., Bülthoff, Edelman, & Tarr, 1995; Diwadkar & McNamara, 1997; Schacter, Cooper, & Delaney, 1990). A number of researchers have shown that, under a variety of conditions and with a variety of stimuli, human object recognition is sensitive to rotation in depth (for reviews, see Tarr, 1995; Tarr & Bülthoff, 1995): We are slower to recognize some objects in novel views than in familiar views. Such findings demonstrate that object recognition is not completely viewpoint-invariant, at least not in any strong sense.

In response to considerations such as these, some researchers have rejected structural descriptions in favor of the idea that we recognize objects on the basis of stored *views.* The claims of such view-based theorists are often very strong: The claim is not simply that object recognition is sensitive to variations in viewpoint; it is that we recognize objects by matching images to literal templates ("views") stored in memory. A complete list of references would be too long to include in this chapter, but some of the more influential and/or vocal versions of this argument appear in Edelman (in press), Edelman and Weinshall (1991), Lawson and Humphreys (1996), Olshausen, Anderson and Van Essen (1993), Poggio and Edelman (1990), Tarr (1995), Tarr and Bülthoff (1995), Tarr, Bülthoff, Zabinski, and Blanz (1997), Tarr and Gauthier (1997), Ullman (1989, 1996), and Ullman and Basri (1991).

According to this view-based account, the human object recognition system does not decompose an object's image into parts and specify the parts' interrelations. Instead, the idea is that we recognize objects *holistically*, matching whole images directly to image-like views stored in memory. (Images are not matched in their "raw" retinal form. Rather, some preprocessing is typically assumed to normalize for illumination, absolute location and size on the retina, etc.; see Edelman & Poggio, 1991.) In most view-based models, the stored views are 2-dimensional (2D) and highly viewpoint-specific. Objects are recognized in novel viewpoints by means of operations (detailed shortly) that bring viewed images into register with stored views.

View-based models have recently gained wide popularity in the object recognition community. Indeed, the view-based approach is arguably the dominant current approach to theorizing about shape perception and object recognition: In addition to the list of influential view-based papers above, there is a much longer list of papers that, citing these (and other) papers, simply take view-based theory as a given and proceed from there (see, e.g., Diwadkar & McNamara, 1997). Part of the appeal of view-based models is their apparent consistency with the effects of depth rotations on human object recognition (see Bülthoff & Edelman, 1992; Tarr & Bülthoff, 1995; but see Biederman & Gerhardstein, 1995). Another, perhaps more important part of their appeal is their simplicity. In contrast to structural description theories, which postulate complex parts-based representations of shape, view-based theories postulate only that we store and match individual views. There is substantial parsimony in this assumption, and substantial elegance in the routines that some view-based models use to recognize 3D objects. However, this same simplicity renders these models inadequate as a general account of human object recognition. Although view-matching is remarkably simple, human object recognition is not.

As elaborated in the remainder of this chapter, there is more to object recognition than recognizing objects in novel viewpoints, and most of it demands explanation in terms of structured rather than holistic representations of shape. The properties of human object recognition that I will cite as evidence against the view-based approach are not subtle. They are intuitive and completely obvious to anyone with a working visual system. The only thing subtle about them are the reasons why they are fundamentally incompatible with view-based theories. This is not to say that holistic "view-like" representations play no role in human shape perception. There is evidence that face recognition is based on holistic representations (e.g., Cooper & Wojan, 1997; Tanaka & Farah,

1993), and my colleagues and I have found evidence for both structural descriptions and view-like representations in common object recognition (specifically, view-like representations seem to mediate the recognition of unattended objects; Stankiewicz, Hummel & Cooper, in press; see also Hummel & Stankiewicz, 1996a). However, very little of the interesting action in shape perception and common (non-face) object recognition is attributable to such representations.

It is important to note that the arguments I will present against the view-based approach are not particularly new (except that they are perhaps a bit more formal than most previous arguments): Similar arguments have been around since 1967 (see Clowes, 1967; see also Palmer, 1978), and probably even earlier; and the old arguments are still right. The view-based approach is thus not only wrong, it is regressive.

Algorithms for Object Recognition

In order to understand the core difference between structural descriptions and views, it is important to step back and consider the problem of object recognition, broadly defined. Object recognition is the process of matching a representation of an object's image to a representation in long-term memory. The properties of any given object recognition algorithm (including the algorithm the human visual system uses, whatever that turns out to be) are determined jointly by the representations on which it is based and by the operations that match those representations to memory. (More broadly, *any* algorithm is defined jointly by its representations and the processes that act on them.) Nonetheless, structural description and view-based theories differ in their emphasis on the role of representation vs. process in object recognition. In general, structural descriptions emphasize the role of representation, postulating complex representations, and matching those representations to memory on the basis of relatively simple operations (see, e.g., Hummel & Biederman, 1992). According to these theories, object recognition is robust to variations in viewpoint to the extent that the parts and relations on which it is based are robust to variations in viewpoint. By contrast, view-based theories postulate very simple representations (image-like views), and match them to memory by means of relatively complex operations (such as alignment and view interpolation). According to these theories, object recognition is robust to viewpoint to the extent that differing views can be brought into register with one another.

The fact that structural description theories postulate simple operations on "smart" representations, whereas view-based theories postulate "smart" operations on simple representations is important because, within limits, simplicity in one domain (representation or process) can be compensated by intelligence in the other. The algorithm as a whole depends on both. This state of affairs can make things difficult for the vision scientist trying to understand human object recognition on the basis of behavioral data: Does our ability to recognize objects in novel views reflect the way we represent shape or the processes we use to match those representations to memory? More to the point for our current discussion, Do findings of view-sensitivity imply that object recognition is view-based? As evidenced by recent debates in the literature (see, e.g., Biederman & Gerhardstein, 1995, and Tarr & Bülthoff, 1995), this question can prove very difficult to answer. In fact, the question is worse than difficult. It is impossible, in principle, to decide between view- and structure-based theories based only on patterns of the view-sensitivity or view-invariance in human object recognition.

The reason is that view-sensitivity vs. view-invariance is not the core difference between view- and structure-based algorithms for object recognition. Models based on either approach can be modified to accommodate findings of greater or less view-dependence in human object recognition. For example, the view-based model of Ullman (1989) can in principle predict complete view-invariance in object recognition (see also Lowe, 1987). This model uses alignment to match 2D images to 3D models in memory. Although matching to memory is view-based in the sense of modern view-based models (as elaborated shortly), the object models themselves are 3D and object-centered. Similarly, the structural description model of Hummel and Stankiewicz (in press; Stankiewicz & Hummel, 1996) is based on parts and relations that are not strictly categorical (a departure from the [almost strictly categorical] models of Biederman, 1987, and Hummel & Biederman, 1992). This model predicts systematic effects of viewpoint at brief presentations and

for unattended objects (see also Hummel & Stankiewicz, 1996a; Stankiewicz, Hummel & Cooper, in press)[1]. Even Biederman's original (1987) theory—which is often misconstrued as predicting complete or nearly complete view-invariance—predicts systematic effects of viewpoint in recognition (Biederman & Gerhardstein, 1993, 1995). Thus, the role of viewpoint in object recognition is not the deciding issue in the view- vs. structure-based debate. To understand what the deciding issue is, it is necessary to consider the nature of views and structural descriptions in greater detail.

What is a "view"?

The terms "view" and "view-based" encompass a variety of specific theories and computational models. However, all view-based models share a common, defining assumption: They all assume that objects are represented and matched to memory in terms of their features' *coordinates* in a spatial reference frame (Edelman, in press; Edelman & Weinshall, 1991; Poggio & Edelman, 1990; Ullman, 1989; Ullman & Basri, 1991). The central tenet of the view-based approach is that we represent objects in long-term memory as views, and that—by means of operations on the coordinates of the features in those views—we bring new views into register with stored views (or, in the case of Ullman, 1989, and Lowe, 1987, into register with stored 3D models). The nature of these operations is the primary focus of most view-based theories, and the coordinate-based nature of a view plays an essential role in these operations.

Formally, a view is a vector of spatial coordinates. For example, in a 2D coordinate space (x, y), a view containing five features would be the 10-dimensional vector [x_1, y_1, x_2, y_2,..x_5, y_5]. Views—that is, vectors of coordinates—have four properties that form the foundation of the view-based account: (1) All feature coordinates are expressed relative to a single reference point (namely, the origin of a coordinate system). (Edelman & Poggio, 1991, describe a view-based model that codes each feature in terms of its angle and distance from one other feature. This model uses multiple reference points, but the coordinates are linear (as defined shortly), so it is formally equivalent to a single-reference point linear coordinate system. The reason for this equivalence is discussed in detail in the Appendix. See also Edelman & Poggio, 1991; Hummel & Stankiewicz, 1996b.) (2) The coordinate system is spatial in the sense that vector elements code coordinates in a spatial reference frame. (Although any vector can be described as a point is a high-dimensional space, a vector is spatial in the sense described here only if the vector elements represent spatial coordinates.) (3) The value of each coordinate (vector element) varies linearly with the location of the corresponding feature in the reference frame: If moving a feature distance *d* has effect *c* on a given coordinate, then moving it 2*d* will have effect 2*c*. And (4) views are *holistic*, in the sense that the various features in a view are not represented independently of their locations (list positions) in the vector. In contrast to a symbolic representation, in which symbols are free to "move around" without changing their meaning, features within a view are defined by their locations within the vector (cf. Holyoak & Hummel, this volume; Hummel & Holyoak, 1997; Hummel & Stankiewicz, 1996b). For example, consider an object with five features, and assume that feature A is located at location 1,1, B at 2,2, and so forth, up to E at 5,5. If features are placed into the vector in alphabetical order, then the resulting vector (i.e., view) would be [1,1,2,2,3,3,4,4,5,5]. However, if we

---

[1]The Hummel and Stankiewicz (1996a) model is based on a hybrid representation of shape, in which view-like representations are integrated into (i.e., serve among the components of) the structural descriptions that mediate recognition. The view-like components of these representations are both more view-sensitive than the other components and faster to generate from an object's image. As a result, this model predicts (correctly; see, e.g., Ellis & Allport, 1986; Ellis Allport, Humphreys & Collis, 1989) that recogtnition will be more view-sensitive early in processing (i.e., immediately after an image is presented for recognition) than it is later in processing. The Hummel and Stankiewicz (in press) model extends the models of Hummel and Biederman (1992) and Hummel and Stankiewicz (1996a) with a more detailed account of how categorical (i.e., non-view-like) visual properties are represented. In brief, due to noise (in the stimulus and in the processing system), this model requires both attention and processing time to generate a reliable (i.e., low noise) representation of an object's shape. As a result, the model predicts systematic effects of processing time and attention on shape representation.

reverse the placement of A and B in the vector, then the same object would be represented by the new vector [2,2,1,1,3,3,4,4,5,5]: In a vector representation, feature identity is bound inexorably to list position. A related point is that, because coordinates are dependent on their list positions in the vector, a given value in one part of the vector bears no relation to that same value in another part of the vector: A 1 in the first position is a different thing entirely from a 1 in the second.

To anyone familiar with the properties of coordinate spaces and vectors, (1) - (4) are so intuitive that they hardly seem to deserve mention. However, these properties are worth considering in detail because they constitute the foundation of the view-based approach to representing shape. These properties are also important because they distinguish view-based models from structural description models: Structural descriptions differ from views on each of these properties (Hummel & Biederman, 1992; Hummel & Stankiewicz, 1996a).

What is a "structural description"?

The alternative to a holistic representation, such as a view, is a structured representation, such as a structural description. In a structured representation, complex entities (such as objects) are represented as collections of simpler elements (such as parts or part attributes) in specific relations (cf. Fodor & Pylyshyn, 1988; Holyoak & Hummel, this volume). In contrast to a holistic representation, a structured representation codes elements independently of one another and of their interrelations (Hummel & Biederman, 1992). As a consequence, it is necessary to actively (i.e., dynamically) bind elements to their relations in a structured representation (Holyoak & Hummel, this volume; Hummel & Biederman, 1992; Hummel & Holyoak, 1997; Hummel & Stankiewicz, 1996a). For example, the simple shapes in Figure 1 might be described by the elements *circle* and *square* and the relations *above()* and *larger()*. A structured representation of the shape in Figure 1a would bind *circle* to the agent role of *above()* and the patient role of *larger()*, and bind *square* to the patient role of *above()* and the agent role of *larger()*, forming the structure *above(circle, square)* and *larger(circle, square)*. The form in Figure 1b would be represented by rebinding the *very same elements* to form *above(square, circle)* and *larger(square, circle)*. The key difference between a holistic representation and the structured one is that the holistic representation binds elements to roles by means of list position in the vector, whereas the structured representation binds elements to relational roles dynamically; as a result, the elements of a structured representation are independent of their relations, whereas the elements of a holistic representation are not (see Holyoak & Hummel, this volume).

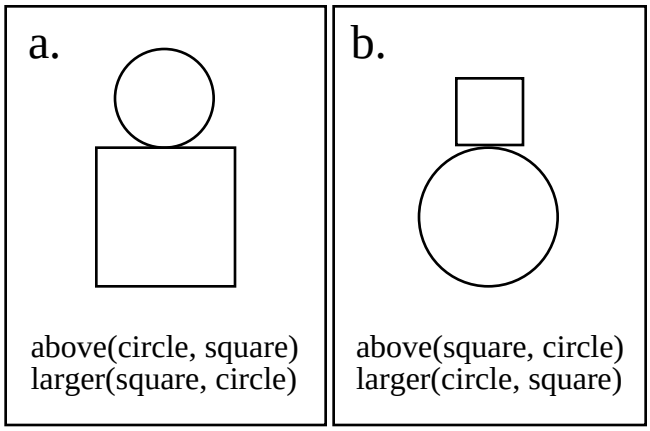

Figure 1. Two simple shapes along with hypothetical structural descriptions.

Although the previous example used discrete symbols to denote elements and relations, a structural description can also be represented as a collection of vectors (Holyoak & Hummel, this volume; Hummel & Biederman, 1992; Hummel & Holyoak, 1997; Hummel & Stankiewicz, 1996a, in press; Shastri & Ajjanagadde, 1993). For example, each vector might code a separate set of bindings, with one vector binding *circle* to *above-agent* and *larger-patient*, and another binding *square* to *above-patient* and *larger-agent.* (Note that although this is a vector, it is not a linear coordinate vector.) Augmented with a means for dynamically binding element to relations (e.g., through synchrony of firing; Hummel & Biederman, 1992), a single vector can be part of a larger compositional structure (e.g., in the way that *circle* is part of *above(circle, square)*). But because relational structures are composed from collections of element-relation bindings (e.g., collections of vectors), no single vector can, by itself, be a complete relational structure (in the way that *circle* is not a complete relational structure). To represent relational structures—that is, explicit relations—a representational system must be able to dynamically bind relational roles to their fillers. "Flat" vector representations lack any means for dynamic binding, and therefore cannot represent relations explicitly (cf. Holyoak & Hummel, this volume; Hummel & Biederman, 1992; Hummel &

Holyoak, 1997; Marcus, 1997). The capacity to bind simple elements into relational structures is the hallmark of a structured, *symbolic* representation (cf. Fodor & Pylyshyn, 1988; Gallistel, 1990; Halford, Bain, & Maybery, in press; Holyoak & Hummel, this volume), and is the single most important difference between a structural description and a view.[2]

The strengths of the view-based approach

View-based models are concerned primarily with understanding how the coordinates of an object's features (in an image) change as the object is rotated in depth, and with exploiting these regularities to recognize objects in unfamiliar viewpoints. By this criterion, view-based models have been very successful. For example, Ullman and Basri (1991) note that the coordinates of the features in any 2D view of an object can be expressed as a linear combination of their coordinates in a finite number of other views (provided the same features are visible in all views). Their view-based model exploits this fact to recognize objects in novel views. By discovering which linear combination characterizes a given new view, this model recognizes objects at novel orientations in depth, even though it stores only 2D views in memory. The models of Poggio, Edelman and their colleagues (Edelman, in press; Edelman, Cutzu, & Duvdevani-Bar, 1996; Edelman & Poggio, 1991; Edelman, & Weinshall, 1991; Poggio & Edelman, 1990; Poggio & Vetter, 1992) exploit similar relationships between the coordinates of features in different views. Like Ullman and Basri's model, these models store only 2D views in memory, but can recognize 3D objects at novel orientations in depth. Because of the way they exploit the properties of linear coordinate vectors, view-based models have met with substantial success, both as accounts of how objects could in principle be recognized in novel viewpoints, and as accounts of some effects of viewpoint in human object recognition (Bülthoff & Edelman, 1992; Edelman & Weinshall, 1991; see also Tarr, 1995; Tarr & Pinker, 1989, 1990).

The successes of these (and other) view-based models have made it clear that it is not necessary to postulate complex structural descriptions to explain how we might recognize objects in unfamiliar viewpoints: Simple template-like representations can explain more than we have traditionally supposed. Some researchers have taken this idea even further, suggesting that it is not necessary to postulate visual representations, per se, at all. For example, Edelman (in press) notes "Indeed, the idea of second-order isomorphisms places the responsibility of representation where it belongs--in the world" (p. 18): Effectively, the claim is that visual systems need not bother to represent shape (see also Edelman, et al., 1996). Not all view-based theories take such a hard anti-representationalist line (see, e.g., Tarr, 1995; Tarr, et al., 1997), but they all assume that we can explain all the interesting action in object recognition in terms of the way the coordinates of an object's features behave in a holistic linear spatial coordinate system. In making the leap from "view-based models can recognize objects at novel orientations in depth," to "we don't need structural descriptions to understand human object recognition," view-based theorists have thrown the baby out with the bath water. They have gone from "object recognition is not completely view-invariant" to "object representations are unstructured." As I shall argue in the remainder of this chapter, the representation of shape is decidedly structured, and this fact makes the view-based approach fundamentally and irreparably flawed as an account of human shape perception and object recognition.

---

[2]Most modern structural description models also happen to use complex parts as primitive elements of representation, whereas view-based models use simpler features. Although this difference is often cited as the major difference between the approaches, in fact it is incidental: It is just as easy to define a structural description on simple features as on complex parts; and it is perfectly possible to define a holistic view on parts rather than features (see Hummel & Stankiewicz, 1996a).

Limitations of the View-Based Approach

Biederman and Gerhardstein (1995) discuss many limitations of the view-based approach, so I shall not repeat that discussion here. Instead, I shall focus on the core limitation of the approach, which derives directly from its foundational assumption: that object shape is represented in a linear spatial coordinate system. The mathematics of view-matching depend on the mathematics of linear spatial coordinates, so there is a deep computational reason why views are —and must be—holistic vectors of linear spatial coordinates (see Edelman, in press; Hummel & Stankiewicz, 1996b).

The independence of elements and relations in a structured representation makes two things possible that are not possible in a holistic representation. First, in a structured representation it is possible to evaluate (i.e., respond to) entities and their relations independently (Halford, et al., in press). For example, our structured representation of Figure 1 makes it possible to appreciate what Figures 1a and 1b have in common and the ways in which they differ. By contrast, in a holistic representation, the shapes in Figure 1a and 1b are simply different, and the specific way(s) in which they differ are not recoverable (see Edelman, in press; Poggio & Edelman, 1990; Tanaka & Farah, 1993). The second property that derives from the independence of elements and relations in a structured representation is the capacity to recombine a finite number of elements and relations into a large (potentially infinite) number of specific structures. It is in this sense that structured representations are symbolic (see Fodor & Pylyshyn, 1988). In a holistic representation, a separate element (e.g., view) is required for each new entity, and recombination is not possible.

The holistic/structured distinction described here maps directly onto the more familiar integral/separable distinction discussed is the categorization literature (see Garner, 1974). Like the elements of a holistic representation, integral dimensions (such as the hue and brightness of a color), are not represented independently of one another, and therefore cannot be responded to independently; and like the elements of a structured representation, separable dimensions are represented independently, and can be responded to independently. According to view-based theory, parts and relations should behave as integral dimensions. But in human shape perception, parts and their relations are decidedly separable. A brief glance at Figure 1 reveals that we have no difficulty appreciating the similarity of the circle in (a) to the circle in (b) (even though the circles are not identical, and would therefore activate different views). Similarly, Goldstone, Medin, and Gentner (1991) showed that people evaluate shapes and their relations independently in judging the similarity of different figures, and Saiki and Hummel (in press) showed that shapes and relations are perceptually separable in terms of the "classic" measures of perceptual integrality/separability. Although this may seem a minor point, its importance is difficult to overemphasize: According to the view-based approach, it is a *complete mystery* how we appreciate the similarity of the two circles in Figure 1. (The question of how a view-based theorist would respond to this fact is addressed in detail later.)

A related point concerns the role of relational structures in similarity judgments. Not only *can* we explicitly evaluate relations for the purposes of computing similarity, in important respects we are *compelled* to: Markman and Gentner (1993) showed that subjects make similarity judgments by aligning structured representations of to-be-compared items. This phenomenon, too, is apparent in Figure 1. Looking at Figure 1, we not only appreciate that both shapes contain a circle and a square; we also appreciate that the relations in (a) are reversed in (b). That is, on the basis of the parts' shapes, the circle in (a) corresponds to the circle in (b); on the basis of the relations, the cirle corresponds to the square. The alternative correspondences are easy for us to appreciate, but they require us to represent the relations between the simple shapes both explicitly and independently of the shapes, themselves. As such, they are impossible for a model based on holistic views to appreciate.

The role of relational structures in visual similarity is a fundamental issue with important implications for the question of how we generalize from familiar shapes to new ones. Our ability to generalize over variations in shape lies at the heart of our ability to recognize objects as members of a class, and thus to generalize our knowledge about one object to others. The question of how we do this is at least as important as the question of how we recognize familiar objects in novel

viewpoints. As a theoretical problem, it is even more difficult. The laws of projective geometry constrain the possible appearances of any single object in different views—a fact that lies at the heart of view-based theory. But no such laws constrain the possible shapes of the various members of a class. Consider, for example, the vast array of shapes in the class *chairs*, and how the physical constraints on what makes an object a suitable chair differ from the constraints on, say, what makes an object a suitable knife. Although view-based models have met with some success accounting for our ability to generalize from one object view to other views of that *same* object, they fail utterly to account for our ability to generalize from one shape to another (the claims of Edelman, in press, notwithstanding).

Visual generalization is tied closely to visual similarity: We usually recognize a novel chair as a chair because it *looks like* a chair. (We can also categorize objects on the basis of more "theory-based" considerations [see, e.g., Rips, 1989], but the present discussion is concerned only with generalization based on visual similarity. But as an aside, it is interesting to note that the kinds of knowledge that make theory-based categorization possible—e.g., the fact that an object has a horizontal surface of a size and height suitable for sitting—cannot be explained without appeal to explicit structural descriptions.) A model's predictions about visual similarity thus underlie its predictions about visual generalization. View-based models, as a class, make very general predictions about visual similarity: Two objects will appear similar to the extent that they have the same features in the same coordinates (cf. Hummel & Stankiewicz, 1996b); the pair-wise relations among their features (e.g., whether one feature is above or below another) do not matter *at all* except inasmuch as they are reflected in the features' coordinates.

Consider the *Basis* object and its variants (*V1* and *V2*) in Figure 2. Hummel and Stankiewicz (1996b) generated V1 from the Basis object by moving one line (the short vertical) six pixels; V2 was generated by moving that same line and the long horizontal to which it is attached the same direction and distance. In terms of their features' coordinates—that is according to any view-based model—the Basis object is more similar to V1 (which differs in the location of one line) than to V2 (which differs in the location of two). Hummel and Stankiewicz (1996b) ran the Poggio and Edelman (1990) model on these and several similar figures, and the model does indeed rate the Basis-V1 match higher than the Basis-V2 match. To human observers, the Basis object looks much more like V2 than like V1. This demonstration holds for a wide variety of stimuli over a wide variety of tasks (Hummel & Stankiewicz, 1996b).

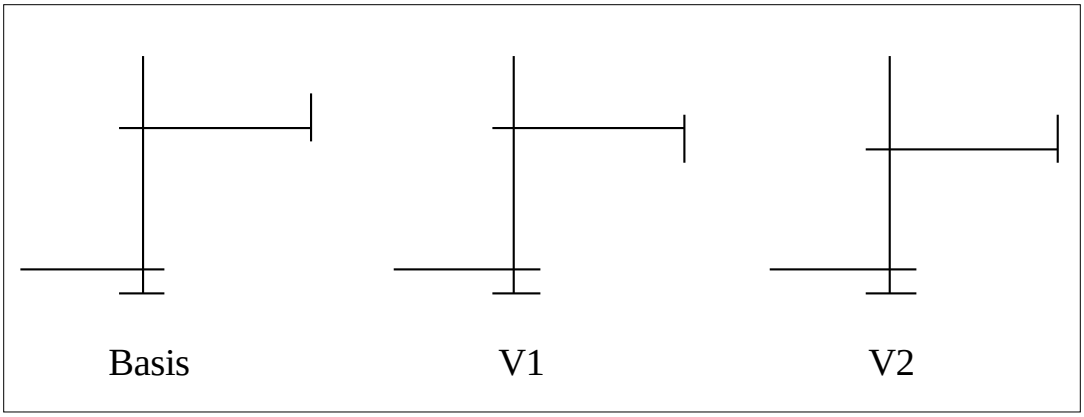

Figure 2. Sample stimuli from the experiments of Hummel and Stankiewicz (1996b, adapted by permission). V1 was made from the Basis object by moving the short vertical line six pixels down. V2 was made from the basis object by moving that same line and the long horizontal to which it is attached the same direction and distance.

Virtually everyone is willing to agree that the Basis object looks more like V2 than V1, but few people are willing to believe that any modern model of object recognition could possibly predict otherwise. The prediction seems absurd. And, indeed, it is completely absurd. But it is a very

general, fundamental prediction of all view-based models. The prediction derives directly from the view-based models' use of linear spatial coordinates, so it holds for *any* view-based model, regardless of whether the model uses 2D or 3D coordinates, Cartesian or polar coordinates, viewer- or object-centered coordinates, etc. (see the Appendix and Hummel & Stankiewicz, 1996b, for a formal discussion of why this is true). It is also independent of the feature set on which the coordinates are defined. A common objection to the example in Figure 2 is that it is misleading because it is based on the wrong set of features: The difference between the Basis object and V2 is a small difference in the location of the T-shaped feature, whereas the difference between the basis object and V1 is that one feature (an upward-pointing T-shape) has been replaced with a qualitatively different feature (a downward-pointing T-shape); a view-based model based on the right set of features (more complex than simple lines) would predict the right similarity relations between the objects.

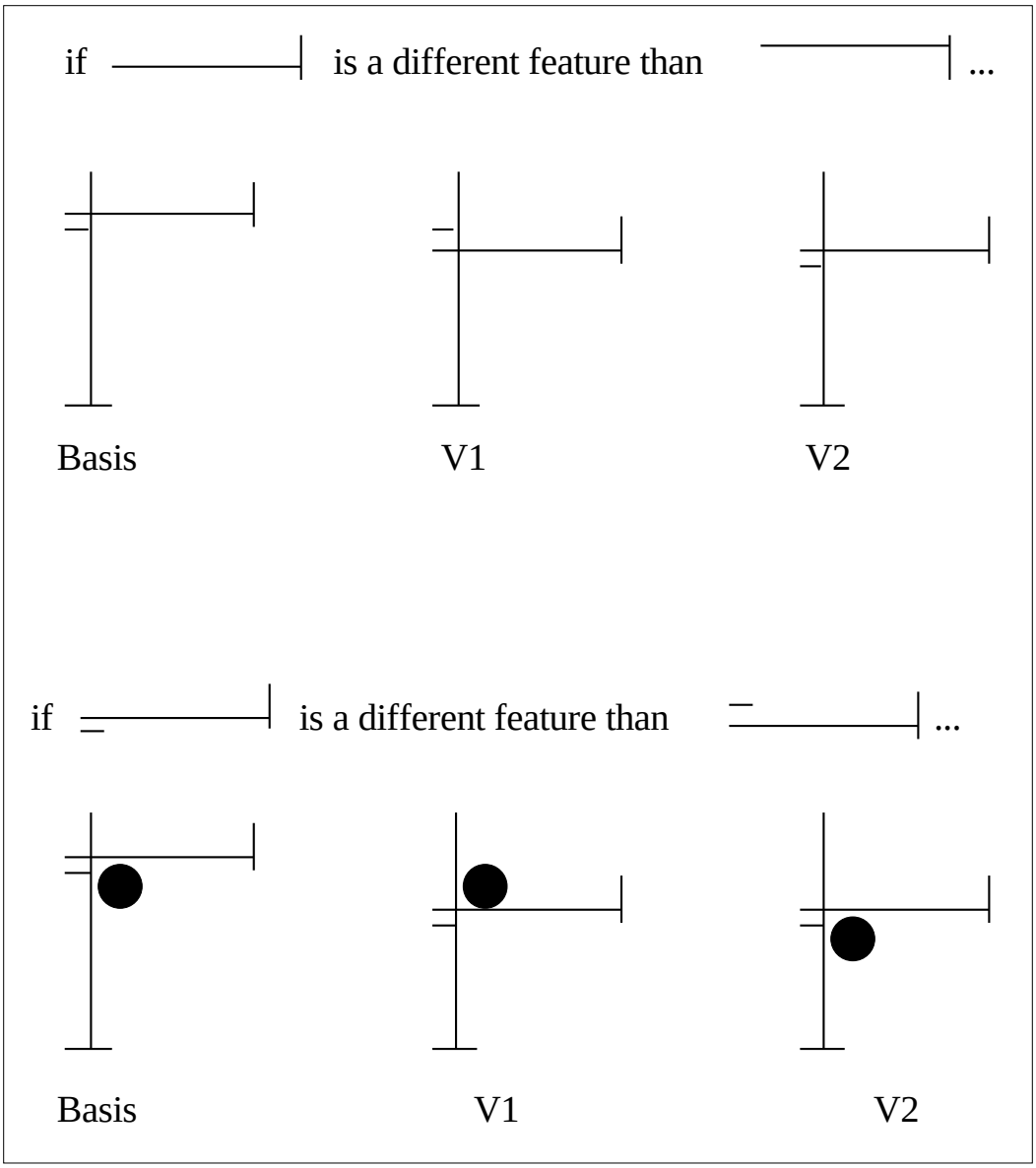

Figure 3. Demonstration that, even with an increasingly complex definition of "features", it is possible to create a more detectable change in a stimulus by moving one feature and crossing a categorical boundary, than is created by moving two features the same distance without crossing a categorical boundary.

This objection does not work, however, because for any set of features—no matter how complex—it is always possible to create examples in which moving one feature results in a more detectable change than moving two: It is only necessary to change a categorical relation in the former case but not the latter. Figure 3 replicates the demonstration in Figure 2 with a series of increasingly complex "features". Note that in each case, V2 looks more like the basis object than V1 does. Note also that it would be a simple matter to generalize this demonstration to 3D object-centered coordinates (e.g., by moving parts closer to or farther from the viewer), so it is not simply a function of the 2D viewer-centered coordinates on which the example is nominally based.

The reason the Basis object looks more like V2 than V1 is that we explicitly perceive the categorical spatial relations among the objects' parts. A categorical above-below relation differs between V1 and the Basis object; in V2, twice as many coordinates differ, but no categorical relations differ. The Basis looks more like V2 than V1 because categorical relations—not coordinates—drive visual similarity, generalization, and object recognition (at least for attended objects; see Stankiewicz et al., in press). As noted previously, explicit relations require non-holistic *structured* representations, i.e., structural descriptions (Fodor & Pylyshyn, 1988; Halford et al., in press; Hummel & Biederman, 1992; Marr, 1982).

Because the view-based approach fails to capture the appropriate similarity relations among different objects, it fails to account for our ability to generalize from one object shape to another: To a view-based model, the wrong things look alike. A related limitation is that, although view-matching can tell you *which* things look alike (as long as you do not expect its intuitions to conform to your own), it cannot tell you *why* they look alike. To a holistic representation, things are simply similar or not-so-similar (Edelman, in press). This property of holistic view-matching is evident in our own intuitions about the similarity of different faces (which we represent holistically). We can often say who looks like whom, but barring obvious features such as scars, facial hair, etc., it is typically difficult to say why (Tanaka & Farah, 1993). The output of holistic view-matching is thus not very useful. To decide whether to sit on an object, it is necessary to know *why* it looks like a chair (i.e., does it have a surface suitable for sitting?), not simply *how much* it looks like a chair. A view-based model can tell you that some new object looks, say, 80% like a chair (i.e., by activating a "chair" view to 80% of its maximum activation); but this statement is as likely to apply to a trash compactor or a wheat thresher as to a footstool, and only one of these objects would make a suitable chair. (An electric chair also looks a lot like a chair, but most of us hope to avoid ever sitting in one.) This failing of the view-based approach is fundamental because it reflects the foundational assumptions of the approach. The same properties that allow view-based models to generalize from one object view to another—namely, the properties of linear spatial coordinate systems—preclude their generalizing in a human-like fashion from one object shape to another. There is no way to modify the view-based approach to overcome this limitation, short of discarding coordinates in favor of explicit relations—that is, discarding views in favor of structural descriptions.

Replies in Support of the View-Based Approach

I am arguing that view-based theories are hopelessly flawed as a general account of human shape perception and object recognition, so no view-based theorist would take my arguments lying down. In the following, I shall try to anticipate the most likely objections to my arguments, and respond to each.

Objection 1: You have missed the point. View-based theory is not about coordinates, it is about the role of information-rich, viewpoint-dependent features in object recognition. For example, Tarr et al. (1987), emphasize the role of features in views, stating that "View-based theories propose that recognition relies on features tied to the input image or images..." (p. 282). This version of view-based theory makes no explicit reference to the role of coordinates in views.

Reply: The problems with this objection begin with the fact that the proposed features, and the operations that match them to memory, are undefined. (All *working* view-based models are of the

coordinate-based type; see Hummel & Stankiewicz, 1996b.) As a result, this version of view-based theory, although it does not explicitly endorse the coordinate-based approach to coding an object's features, also does not escape it. Let us grant the view-based theorist any population of features he/she desires, and allow him/her to add or delete features from the theory at will. The question will still remain How do these features combine to form representations of whole objects? Unless the features alone somehow uniquely specify object identity (i.e., in the way that a linear combination of global fourier components uniquely specifies an image), it will be necessary to know the features' configuration (i.e., where they are located, either in the image or relative to one another).[3] The view-based theorist must therefore choose some basis for representing the features' configuration. If he/she chooses coordinates, then the theory is subject to the limitations detailed above; if he/she chooses explicit relations (the right choice), then he/she has abandoned views in favor of structural descriptions.

Objection 2: The view-based claim is not about features or about coordinates vs. relations, but about the nature of the reference frame. Whereas structural description models assume an object-centered reference frame, view-based theories assume that "Interobject relations are mentally represented in a viewpoint-dependent manner" (Diwadkar & McNamara, 1997; p. 307).

Reply: This objection is simply factually incorrect. Not all structural descriptions are object-centered. In fact, the question of structural descriptions vs. views is *unrelated* to the question of object- vs. viewer-centered reference frames (Hummel, 1994). The structural description models of Hummel and Biederman (1992) and Hummel and Stankiewicz (1996a, in press) are all based on relations defined in a *viewer-centered* reference frame. Following Biederman's (1987) theory, the models' capacity to discount variations in viewpoint derives from their use of categorical part attributes (e.g., straight vs. curved major axis) and categorical relations (e.g., above vs. below). Based on these (and other) categorical attributes and relations, the models account both for the invariances of human object recognition (e.g., invariance with translation, scale, and left-right reflection) and for the sensitivity of human object recognition to rotation in the picture plane (see Hummel & Biederman, 1992).

The models' ability to simulate the effects of picture-plane rotation is especially interesting. When human observers recognize pictures of common objects, response times and error rates increase monotonically as the image is rotated to about 135 degrees, and then decrease between 135 and 180 degrees (see, e.g., Jolicoeur, 1985). Although this function is usually interpreted as evidence for "mental rotation" in object recognition, the Hummel and Biederman model simulates it without performing rotations of any kind. Rather, this function simply reflects the way an object's categorical attributes and relations change as its image is rotated in the picture plane. (For example, if part A is above part B in an upright image, then A will be beside B after a 90 degree rotation, and below B after a 180 degree rotation.) If the model's structural descriptions were based on an object-centered reference frame, as is often incorrectly assumed, then the model would not be able to simulate this aspect of the human data.

Objection 3: The Hummel and Stankiewicz (1996b) experiments falsify models based on the use of literal retinotopic (x, y) coordinates, but we already knew that you cannot explain object recognition using those kinds of coordinates. Not all view-based models use literal x, y coordinates, so the Hummel and Stankiewicz experiments do not falsify all view-based models.

Reply: In fact, the Hummel and Stankiewicz experiments falsify models based on *any* set of linear coordinates, not just retinotopic (x, y) coordinates. As long as the coordinates are linear (as they are in all view-based models), the prediction is the same, regardless of the reference frame (as elaborated in the Appendix): Moving two features a given distance will always affect similarity more than moving only one of those features that distance (Hummel & Stankiewicz, 1996b).

---

[3]Recognition by simple feature listing is inadequate as a theory of human object recognition (see Hummel & Biederman, 1992). Unless the features serving object recognition are like global fourier components (which they are not [see, e.g., Tanaka, 1993]), then simply listing an object's features (e.g., "14 horizontal lines, 16 vertical lines...") no more specifies its shape than listing colors specifies the appearance of the Mona Lisa.

Objection 4: Your argument, most notably your claim that a view-based model could not appreciate the equivalence of the two circles in Figure 1, attacks a straw man. It is perfectly plausible to postulate a hierarchy of views for any given object. For example, the object in Figure 1a could activate three different views: One for the circle, one for the square, and a third for the entire circle-above-square configuration. Because a "circle" view would be activated by both Figure 1a and Figure 1b, the view-based model would appreciate the equivalence of the two circles.

Reply: If the view-based model could ignore the fact that the circles are different sizes, then said model would indeed make the equivalence of the two circles explicit. However, even this model would fail to make explicit either the part-whole relation between the circle and the circle-square configuration, or the spatial relation between the circle and the square. To the hierarchy-of-views representation, the "circle" view is a completely different thing than the "circle-above-square" view. In terms of this representation, they have no more in common than the "circle" view has in common with a view of the Eiffel Tower. (This is a consequence of the views' being holistic: By definition, only a compositional representation makes explicit the relations between the parts and the whole; see, e.g., Tanaka & Farah, 1993.) By contrast, the human visual system appreciates not only that both Figures 1a and 1b contain circles, but also that the circle in Figure 1a is a subset of the figure as a whole. Making this relation explicit in the view-based model would allow the model to appreciate the same thing, but it would turn the view-based model into a structural description.

A related limitation of the hierarchy-of-views proposal is that it still assumes that each part is represented as an indivisible whole. This assumption also falls short of human perception, because there is good reason to believe that even individual parts (e.g., geons) are themselves composed of properties that can be shared by other parts. Note that you can appreciate that a cylinder has one property in common with a cone (both have round cross sections) and has another property in common with a brick (both have parallel sides). If "cylinder," "cone," and "brick" were represented as separate holistic views, then we would not be able to appreciate their shared properties. Recall that the only thing a view makes explicit (i.e., the only thing it represents) is the coordinates of an object's features. That is, coordinates are the one and only currency for expressing similarity in a view-based representation. But the property *round cross section* cannot even be expressed in the language of feature coordinates. "Round cross section" is an *abstract invariant* that is both more and less than any finite set of feature coordinates (see Hummel & Kellman, 1997). It is more because a new set of coordinates might arise from a new instance of a volume with a round cross section, and it is less because many of the attributes of those new coordinates have nothing to do with their forming a round crosss section. Views are thus inadequate—in principle—to capture the fact that cylinders and cones share the property *round cross section*

By contrast, the structural description models of Hummel & Biederman (1992) and Hummel and Stankiewicz (1996a, in press) can and do capture this shared property. These models represent geons, not as "geon views," but as collections of abstract invariants (such as *round cross section* and *non-parallel sides*; see Hummel & Biederman, 1992). Specifically, geons are represented as patterns of activity distributed over units representing geon attributes (abstract invariants) and relations. Geons that share attributes share the units representing those attributes, thereby making explicit what the geons have in common. Units are bound dynamically into geons (and geons are bound to their inter-relations) by synchrony of firing. For example, a cone is represented by firing *round cross section* in synchrony with *non-parallel sides*; a cylinder is represented by firing *round cross section* in synchrony with *parallel sides*. This capacity for dynamic binding makes these models structured and, effectively, symbolic (see Holyoak & Hummel, this volume).

Objection 5: Your arguments about the similarity relations among the Basis objects and their variants (Figures 2 and 3) fail to take into consideration that a view-based model might differentially weight an object's features (e.g., as proposed by Edelman and Poggio, 1991): The greater a feature's weight, the more important that features' coordinates are for recognition. With the right feature

weightings, a view-based model can account for the demonstration in Figure 2 and the findings of Hummel and Stankiewicz (1996b).
Reply: No, it cannot. The only difference between V1 and V2 is the location of the long horizontal line, which is in its original (Basis object) location in V1 but not in V2. Therefore, the only way to make a view-based model rate the Basis-V1 match lower than the Basis-V2 match is to assign a *negative* weight to the long horizontal. The negative weight would "penalize" the Basis-V1 match for having the long horizontal in the same location in both objects. As a result, the Basis-V2 match (which would not be penalized, since the long horizontal changes location from Basis to V2) would be greater than the Basis-V1 match. (Note that giving the long horizontal a weight of zero will not do the trick, as it will simply cause the Basis-V1 match to equal the Basis-V2 match. To make the Basis-V1 match be lower than the Basis-V2 match, the weight on the horizontal *must* be negative.) Although this feature weighting will make the Basis-V1 match lower than the Basis-V2 match (consistent with the human data), it makes the Basis object is less similar to *itself* than to a variant in which the long horizontal has moved.

Objection 6: Your arguments (and the experiments of Hummel & Stankiewicz, 1996b) are based on coordinates expressed relative to a single reference point (i.e., the origin of the coordinate system). However, Edelman and Poggio (1991) showed that their model can also be implemented using multiple reference points, where each feature location is expressed, not relative to the origin of the coordinate system, but relative to the location of some other feature.
Reply: Even this version of the Edelman and Poggio model assumes coordinates that vary linearly with the location of the feature in the reference frame. As detailed in the Appendix, a linear coordinate system with $N > 1$ reference points is formally equivalent to a linear coordinate system based on a single reference point (Hummel & Stankiewicz, 1996b). As a result, the feature-to-feature version of the Edelman and Poggio model makes exactly the same predictions as the coordinate-based version (as noted by Edelman and Poggio).

Objection 7: Ullman (1989, 1996) explicitly acknowledges that a complete view-based theory will require "deformable templates" to recognize non-rigid objects. For example, recognizing an object as a person requires a model to tolerate variations in the location of the arms relative to the body. A view-based model augmented with routines to tolerate such deformations would likely also be able to account for the findings of Hummel and Stankiewicz (1996b) and related findings. The idea of a transformable template is an example of the point you made earlier about an algorithm being defined both by its representations and by the processes that act on them: With the right processes, even simple templates can account for the properties of human object recognition.
Reply: This objection is directly analogous to Objection 4 above, and its limitations are likewise analogous. A model based on deformable templates must know which parts of the template are free to move, and in what directions. That is, the model must have knowledge of the parts and their legal relations. For example, the deformable template would have to know which parts of the template correspond to arms, which parts correspond to the torso, and what kinds of relations between arms and torso are legal. Adding this knowledge to the view-based model turns it into a structural description model. And even if the model knows these legal relations, then as long as the parts are represented as individual sub-templates (rather than as collections of abstract invariants), the deformable template model will still be unable to appreciate what different parts have in common (as detailed in the reply to Objection 4).

Objection 8: Structural descriptions cannot account for the metric precision of shape perception or the view-dependence of object recognition.
Reply: Yes they can. Hummel and Stankiewicz (in press; Stankiewicz & Hummel, 1996) describe a structural description model that accounts both for our ability to perceive metric properties (e.g., the degree of non-parallelism of a pair of lines) and the importance of categorical properties in shape perception and object recognition (such as the difference between parallel and non-parallel; see, e.g., Cooper & Biederman, 1993; Hummel & Stankiewicz, 1996b). In addition, this model makes novel predictions about the relationship between visual attention, viewpoint-sensitivity, and

metric precision is shape perception. Structural description models can account for findings of viewpoint-dependence in shape perception. Contrary to popular belief, structural descriptions are not strictly object-centered, view-invariant representations (Hummel, 1994; Hummel & Biederman, 1992; Hummel & Stankiewicz, in press; Marr, 1980).

One (apparent) limitation of the structural description approach is that, as a general approach, it can account for anything: Structural descriptions are symbolic representations, so as a general paradigm, the structural description approach is Turing complete. That is, for any (computable) pattern of behavioral data, there exists some structural description model that can account for it. The same cannot be said for the view-based approach. View-based models are holistic, and therefore non-symbolic, so the view-based approach is not Turing complete. As a general approach, view-matching is more falsifiable than the structural description approach. And, indeed, it has been falsified. The important question is not "Do structural descriptions mediate human object recognition?", but rather, "*What kind* of structural description mediates human object recognition?"

Summary and Conclusions

It is currently popular to argue that evidence for view-sensitivity in human object recognition indicates that object recognition is view-based. In fact, the evidence indicates no such thing. There is some evidence for view-sensitivity in human object recognition, but such findings do not imply that object recognition is literally view-based. If shape perception and object recognition really were view-based, then the would would look very different to us than it does. The core limitations of view-based theory derive directly from its foundational assumption that objects are represented as holistic views based on linear spatial coordinates. This assumption is indispensable to view-based theory because it is central to its account of our ability to recognize objects in novel viewpoints. The problem is that this assumption is deeply flawed. For the purposes of common object recognition and classification, the representation of shape is neither holistic nor based on linear spatial coordinates.

View-based theory is currently very popular, but when one takes the claims of view-based theory seriously, it becomes clear that the Emperor is naked. What is missing from the view-based account is any serious theory of representation. View-based theorists are typically quite open about the fact that they are more concerned with the operations that bring viewed images into register with stored views than with the vagaries of exactly how those views are represented. But this is a mistake. By ignoring the question of representation, view-based models answer the question *How do we bring images into register with views n memory?* without asking the prior question <u>Do</u> *we bring images into register with views in memory?* In doing so, they provide some very elegant answers to a non-question. If visual representations were unstructured views, then such operations might be useful. But visual representations are not unstructured views. They are highly structured, and it is impossible to understand object recognition (and other aspects of visual perception and cognition) without understanding *how* they are structured.

View-based theorists are not alone in their assumption that one can understand the mind without also understanding the structure of mental representations. It is currently popular to use recurrent back-propagation and its relatives to model everything from Hippocampal function (e.g., McClelland, et al., 1995), to category learning (e.g., Noelle & Cotrell, 1996), to language (e.g., Elman, et al., 1996). These models, like view-based models of object recognition, represent all knowledge holistically, i.e., as simple feature vectors. They lack—indeed, explicitly eschew—any basis for binding those vectors into compositional (i.e., symbolic) structures. Also like the view-based models, these models have an impressive list of accomplishments: It is not immediately apparent that they have lost anything by rejecting structured representations. But they have. For example, Marcus (in press) showed that such networks cannot generalize outside their training space. In essence, this means that, regardless of the number of examples on which such a network is trained, there will always be some inferences that, although trivial for a person, are impossible for the network (see also Holyoak & Hummel, this volume; Marcus, this volume). Like the limitations

of view-based models, the limitations of these models become most apparent when they are required to generalize to new cases.

The reason for this limitation is that such networks cannot bind values to variables (in the same way that view-based models cannot bind elements to relations). As a result, they cannot execute a function in any truly general way (Marcus, in press; see also Holyoak & Hummel, this volume). Note that even the trivial identity function, $f(x) = x$, requires a capacity to represent the variable x and bind arbitrary values to it. Models without the capacity for binding cannot execute even this trivial function.

Cheng and Park (1997) have shown that even causal reasoning—a task that is most often modeled with simple structure-free models (see Cheng, 1997, for a review)—depends on the capacity to explicitly bind values to variables. Fodor and Pylyshyn (1988) have noted similar fundamental problems stemming from the inability of traditional connectionist networks to represent and manipulate structured knowledge (see also Hummel & Biederman, 1992; Hummel & Holyoak, 1997; Shastri & Ajjanagadde, 1993).

Like human object recognition, human thinking is based on highly structured representations. Neither can be understood in terms of "smart" procedures (such as view interpolation or recurrent back propagation) operating on simple holistic representations.

References

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94* (2), 115-147.

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for 3-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 1162-1182.

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

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

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

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367-405.

Clowes, M. B. (1967). Perception, picture processing and computers. In N.L. Collins & D. Michie (Eds.), *Machine Intelligence*, (Vol 1, pp. 181-197). Edinburgh, Scotland: Oliver & Boyd.

Cooper, E. E., & Biederman, I. (1993). Geon differences during object recognition are more salient than metric differences. Poster presented at the annual meeting of the Psychonomic Society.

Cooper, E. E & Wojan, T. J. (1997). Differences in the coding of spatial relations in face and object identification. *Submitted to Journal of Experimental Psychology: Human Perception and Performance*.

Dickinson, S. J., Pentland, A. P., & Rosenfeld, A. (1992). 3-D shape recovery using distributed aspect matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 14*, 174-198.

Diwadkar, V. A., & McNamara, T. P. (1997). Viewpoint dependence in scene recognition. *Psychological Science, 8*, 302-307.

Edelman, S. (in press). Representation is representation of similarities. *Behavioral and Brain Sciences.*

Edelman, S., Cutzu, F., & Duvdevani-Bar, S. (1996). Similarity to reference shapes as a basis for shape representation. *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, pp. 260-265.

Edelman, S., & Poggio, T. (1991). Bringing the grandmother back into the picture: A memory-based view of object recognition. MIT A.I. Memo No. 1181. April.

Edelman, S., & Weinshall, D. (1991). A self-organizing multiple-view representation of 3-D objects. *Biological Cybernetics, 64*, 209-219.

Elman, J. L., Bates, E., Johnson, M. H., Karmaloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Inateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Fodor, J.A. & Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: A critical analysis. In Pinker, S., & Mehler, J. (eds.) *Connections and Symbols*. MIT Press, Cambridge, MA.

Gallistel, C. R. (1990). *The Organization of Learning*. Cambridge, MA: MIT Press.

Garner, W. R. (1974). *The processing of information and structure*. Hillsdale, NJ: Erlbaum.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relations, attributes, and the non-independence of features in similarity judgments. *Cognitive Psychology, 23, 222-262*.

Halford, G., Bain, J., and Maybery, M. (in press). "Induction of relational schemas: Common processes in reasoning and learning set acquisition" *Cognitive Psychology*.

Hummel, J. E. (1994). Reference frames and relations in computational models of object recognition. *Current Directions in Psychological Science, 3*, 111-116.

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

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104*, 427-466.

Hummel, J. E., & Kellman, P. K. (1997). Finding the pope in the pizza: Abstract invariants and cognitive constraints on perceptual learning. Commentary on P. Schyns, R. Goldstone, & J. Thibaut, The development of features in object concepts. Submitted to *Behavioral and Brain Sciences.*

Hummel, J. E., & Stankiewicz, B. J. (1996a). An architecture for rapid, hierarchical structural description. In T. Inui and J. McClelland (Eds.). *Attention and Performance XVI: Information Integration in Perception and Communication* (pp. 93-121). Cambridge, MA: MIT Press.

Hummel, J. E., & Stankiewicz, B. J. (1996b). Categorical relations in shape perception. *Spatial Vision, 10*, 201-236.

Hummel, J. E., & Stankiewicz, B. J. (in press). Two roles for attention in shape perception: A structural description model of visual scrutiny. *Visual Cognition*.

Jolicoeur, P. (1985). The time to name disoriented natural objects. *Memory & Cognition*, *13*, 289-303.
Lawson, R., & Humphreys, G. W. (1996). View specificity in object processing: Evidence from picture matching. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 395-416.
Marcus, G. F. (in press). Rethinking eliminative connectionism. *Behavioral and Brain Sciences.*
Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431-467.
Marr, D. (1980). *Vision*. Freeman: San Francisco.
Marr, D., & Nishihara, H. K. (1978). Representation and recognition of three dimensional shapes. *Proceedings of the Royal Society of London*, Series B. 200, 269-294.
McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419-437.
Marcus, G. F. (in press). Rethinking eliminative connectionism. *Behavioral and Brain Sciences.*
Noelle, D. C. & Cottrell, G. W. (1996). Modeling interference effects in instructed category learning. In G. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Spciety*: Mahwah, NJ: Lawrence Erlbaum (475-480).
Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, *13*, 4700-4719.
Park, J., & Cheng, P. W. (1997). Boundary conditions for "overexpectation" in causal learning with discrete trials: A test of the Power PC theory. Manuscript submitted for publication.
Poggio, T. & Edelman, S. (1990). A neural network that learns to recognize three-dimensional objects. *Nature*, *343*, 263-266.
Poggio, T. & Vetter, T. (1992). Recognition and structure and from one 2D model view: observations on prototypes, object classes, and symmetries. MIT AI Memo 1347, Massachussetts Institute of Technology.
Quinlan, P. T. (1991). Differing approaches to two-dimensional shape recognition. *Psychological Bulletin*, *109* (2), 224-241.
Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Voisniadou and A. Ortony (Eds.). *Similarity, Analogy, and Thought*. New York: Cambridge University Press.
Saiki, J., & Hummel, J. E. (1997). Connectedness and structural representations in object category learning. *Memory and Cognition*., accepted pending revision.
Schacter, D. L., Cooper, L. A., & Delaney, S. M. (1990). Implicit memory for unfamiliar objects depends on access to structural descriptions. *Journal of Experimental Psychology: General*, *119*, 5-24.
Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings. *Behavioral and Brain Sciences*, *16*, 417-494.
Stankiewicz, B. J., Hummel J. E., & Cooper, E. E. (in press). The role of attention in priming for left-right reflections of object images: Evidence for a dual representation of object shape. *Journal of Experimental Psychology: Human Perception and Performance*.
Stankiewicz, B. J., & Hummel, J. E. (1996). MetriCat: A representation for basic and subordinate-level classification. *Proceedings of the 18th Annual Conference of the Cognitive Science Society* , pp. 254-259.
Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science, 262*, 685-688.
Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *146A*, 225-245.
Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & 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? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 1494-1505.
Tarr, M. J., Bülthoff, H. H., Zabinski, M., and Blanz, V. (1997). To what extent do unique parts influence recognition across changes in viewpoint? *Psychological Science*, *8*, 282-289.
Tarr, M. J., & Gauthier, I. (1997). Do viewpoint-dependent mechanisms generalize across members of a class? Submitted to *Cognition*.
Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, *21*, 233-282.
Tarr, M. J. & Pinker. S. (1990). When does human object recognition use a viewer-centered reference frame? *Psychological Science, 1* (4), 253-256.
Ullman, S. (1989). Aligning pictoral descriptions: An approach to object recognition. *Cognition*, *32*, 193-254.
Ullman, S. (1996). *High-level Vision: Object Recognition and Visual Cognition*. Cambridge MA: MIT Press.

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

Zerroug, M. (1994). Segmentation and inference of 3-D descriptions from intensity images. Technical report IRIS-94-327, University of Southern California.

Appendix

A linear coordinate system is a coordinate system in which the mapping from locations in the world (e.g., in a retinal image) to values in the coordinate system is linear. In a linear coordinate system, any transformation, $\mathbf{t}_i$, on coordinates, $\mathbf{c}^w$, in the world (e.g., as a result of moving a feature in an image) has a linear effect on the representation of those coordinates, $\mathbf{c}^r$, in the coordinate system: The greater the effect of the transformation on the world, the greater its effect on a linear coordinate representation of that world. This is true for any linear coordinate system, regardless of the number of reference points on which it is based. The result is that all linear coordinate systems are formally equivalent in the sense that they generate the same ordinal similarity relations: If coordinate vector $\mathbf{c}_1{}^r$ (i.e., the representation of $\mathbf{c}_1{}^w$) is more similar to vector $\mathbf{c}_2{}^r$ (the representation of $\mathbf{c}_2{}^w$) than to $\mathbf{c}_3{}^r$ in one linear coordinate system, then $\mathbf{c}_1{}^r$ will be more similar to $\mathbf{c}_2{}^r$ than to $\mathbf{c}_3{}^r$ in any other linear coordinate system (provided the systems have the same feature weightings, as discussed below).

For example, consider the coordinates of three points, a, b, and c in a retinal image: a is at location 1,1, b is at 2,2, and c is at 3,3. That is, $\mathbf{c}_1{}^w$ is the vector [1 1 2 2 3 3]. Note that we can express any transformation on this vector as another vector, $\mathbf{t}_i$, which we add to $\mathbf{c}^w$. For example, we can move feature c up (i.e., increase its coordinate in the vertical dimension) by one unit by setting $\mathbf{t}_1$ to [0 0 0 0 0 1] and adding it to $\mathbf{c}_1{}^w$. The resulting image, $\mathbf{c}_2{}^w$, is the vector [1 1 2 2 3 4]. To move c up by two units, we add transformation $\mathbf{t}_2$ ([0 0 0 0 0 2]) to $\mathbf{c}_1{}^w$ to get the image $\mathbf{c}_3{}^w$ = [1 1 2 2 3 5]. Let us quantify the *effect*, $e(\mathbf{t}_i, \mathbf{c}_j{}^w)$, of transformation $\mathbf{t}_i$ on the image $\mathbf{c}_j{}^w$ as the sum of the differences between the original vector ($\mathbf{c}_j{}^w$) and the transformed vector ($\mathbf{c}_j{}^w + \mathbf{t}_i$):

$$e(\mathbf{t}_i, \mathbf{c}_j^w) = \sum_j \left| \mathbf{c}_j^w - \mathbf{c}_j^w + \mathbf{t}_i \right|. \quad (1)$$

(Eq. (1) simplifies to the sum of vector entries in $\mathbf{t}_i$.) Likewise, let us quantify the effect of $\mathbf{t}_i$ on the representation, $\mathbf{c}_j{}^r$, of that image ($\mathbf{c}_j{}^w$) as the sum of the differences between the representation ($\mathbf{c}_j{}^r$) of the original image and the representation, $\mathbf{c}_{j(t)}{}^r$, of the transformed image ($\mathbf{c}_j{}^w + \mathbf{t}_i$):

$$e(\mathbf{t}_i, \mathbf{c}_j^e) = \sum_j \left| \mathbf{c}_j^r - \mathbf{c}_{j(t)}^r \right|. \quad (2)$$

It is now straightforward to show that, in any linear coordinate system, if $e(\mathbf{t}_2, \mathbf{c}_j{}^w) > e(\mathbf{t}_1, \mathbf{c}_j{}^w)$, then $e(\mathbf{t}_i, \mathbf{c}_j{}^r) \geq e(\mathbf{t}_i, \mathbf{c}_j{}^r)$. That is, the greater the effect of a transformation of the coordinates in an image, the greater the effect of that transformation on any linear coordinate representation on that image. Consider two linear coordinate representations: **s** uses a single reference point, and in **m** the first feature (a) is related to the origin of the coordinate system, and every other feature is related to the previous feature (i.e., b to a, c to b, etc.). Representation **m** thus uses three reference points (one per feature). The most obvious version of **s** would be simply to set $\mathbf{s} = \mathbf{c}^w$, (i.e., to adopt for the single reference point representation the same coordinates as used in the image) but in such an example, it is trivial that the effects on the representation will scale with the effects on the image (since the two are identical). Therefore, to make the example less trivial, for **s** we shall use the first feature (a) as the origin of the coordinate system. The effects of transformations $\mathbf{t}_1$ and $\mathbf{t}_2$ on **s** and **m** are shown in the table below. As above, $\mathbf{c}_2$ is the result of transformation $\mathbf{t}_1$ on image $\mathbf{c}_1$, and $\mathbf{c}_3$ is the result of transformation $\mathbf{t}_2$ on $\mathbf{c}_1$. An additional transformation, $\mathbf{t}_3$, is included to illustrate the effect of moving a different feature.

Transformation	Image ($\mathbf{c}^w$)	**s** ($\mathbf{c}^r$)	**m** ($\mathbf{c}^r$)
	$\mathbf{c}_1$ = [1 1 2 2 3 3]	[0 0 1 1 2 2]	[1 1 1 1 1 1]
$\mathbf{t}_1$ = [0 0 0 0 0 1]			
	$\mathbf{c}_2$ = [1 1 2 2 3 4]	[0 0 1 1 2 3]	[1 1 1 1 1 2]
$\mathbf{t}_2$ = [0 0 0 0 0 2]			
	$\mathbf{c}_3$ = [1 1 2 2 3 5]	[0 0 1 1 2 4]	[1 1 1 1 1 3]
$\mathbf{t}_3$ = [0 0 1 0 0 0]			
	$\mathbf{c}_4$ = [1 1 3 2 3 3]	[0 0 2 1 2 2]	[1 1 2 1 1 1]

Note that, in both the single reference point (**s**) and multiple reference point (**m**) coordinate representations, the effect of $\mathbf{t}_1$ on $\mathbf{c}_1$ is 1.0, and the effect of $\mathbf{t}_2$ on $\mathbf{c}_1$ is 2.0. In these examples, the effects on the representations are identical to the effects on the original image, because both representations used the same scale as the image (i.e., a distance of 1.0 in the image maps to a distance of 1.0 in both representations). Changing the scales of the representations will change the absolute magnitudes of the effects of $\mathbf{t}_1$ and $\mathbf{t}_2$, but it will not change their ratio. The only way to change the ratio of the effects of various transformations is to assign different weights to different features (coordinates) in the representation (see Edelman & Poggio, 1991). For example, if all coordinates are given equal weight, then the effect of $\mathbf{t}_3$ is equal to the effect of $\mathbf{t}_1$ (namely, 1.0). But if the weight on the x coordinate of feature b (the third vector entry) is reduced to zero, then $\mathbf{t}_3$ will have no effect on either **s** or **m**. But as long as **s** uses the same weights as **m**, the two representations will always show identical effects of identical transformations. That is, **s** and **m** are formally equivalent.

It is important to emphasize that **s** and **m** are formally equivalent only because they are both linear. If **m** were based on non-linear (e.g., categorical) relations (rather than linear) relations, then **s** and **m** would not be formally equivalent (see Hummel & Stankiewicz, 1996b).