Hummel, J. E. (2013). Object recognition. In D. Reisburg (Ed.) *Oxford Handbook of Cognitive Psychology*. 32-46, Oxford, England: Oxford University Press.

Object Recognition

John E. Hummel

Department of Psychology
University of Illinois

Abstract

The dominant approaches to theorizing about and modeling human object recognition are the *view-based* approach, which holds that we mentally represent objects in terms of the (typically 2-dimensional; 2-D) coordinates of their visible 2-D features, and the *structural description* approach, which holds that we represent objects in terms of the (typically categorical) spatial relations among their (typically volumetric) parts. This chapter reviews the history and nature of these (and other) models of object recognition, as well as some of the empirical evidence for and against each of them. I will argue that neither account is adequate to explain the full range of empirical data on human object recognition and conclude by suggesting that the visual system uses an intelligent combination of structure- and view-based approaches.

Keywords: object recognition, view-based model, structural description, visual attention, viewpoint invariance

Object recognition is a fundamental process that serves as a gateway from vision to cognitive processes such as categorization, language and reasoning. The visual representations that allow us to recognize objects do more than merely tell us what we are looking at. They also serve as a basis for visual reasoning and inference: We may recognize a hammer, not only as an object called "hammer" and an object for pounding nails into wood, but also as an object of about the right weight to balance a beam of a certain length on a fulcrum, an object with which to prop open a door, or an object to tie to the end of a rope for the purposes of throwing it over a high branch in the service of making a swing.

One of the most remarkable properties of the human capacity for object recognition is our ability to recognize objects from a variety of viewpoints despite the fact that different viewpoints can projects radically different—even non-overlapping—images to the retina. And given the retinotopic mapping of early cortical visual processing (e.g., in V1, V2 and to a lesser extent, V4), non-overlapping retinal images give rise to non-overlapping cortical representations, even moderately "late" in the ventral processing stream (i.e., the visual pathway responsible for our ability to know what we are looking at, as opposed to knowing how to interact with it motorically; see, e.g., Goodale et al., 1991). But despite the varying retinal and cortical representations resulting from different object views, we somehow manage to recognize all these different images as arising from the same object—an accomplishment that, on its face, is so remarkable that it largely dominated the study of object recognition for three decades or more (see Palmer, 1999, for a review). However, as important and impressive as this capacity is, it is just one aspect of our remarkable capacity for visual object recognition.

The process of recognizing an object is a process of matching a representation of a viewed object to a representation stored in long-term memory (LTM). These stored and matched representations consist largely (albeit not exclusively; Rossion & Pourtois, 2004) of information about an object's shape (Biederman & Ju, 1988; Mapelli & Behrmann, 1997; Op de Beeck et al., 2000). Accordingly, the study of object recognition consist largely (although not exclusively) of the study of the mental representation of object shape, and the vast majority of theories of object recognition are, effectively, theories of the mental representation of shape.

This chapter is organized as follows. In aid of understanding the major theories of object recognition I will begin by reviewing the formal properties of representations of shape. I will next describe the major theories of human shape perception and object recognition. The majority of the chapter will be spent reviewing the empirical literature on object recognition with an eye toward the implications of these findings for the two dominant theories of object recognition—the view-based account and the structural description account. I will argue that neither account is adequate to explain the full range of empirical data on human object recognition and conclude by suggesting that the visual system uses an intelligent combination of structure- and view-based approaches.

Representations of Object Shape

A representation of shape is defined by four properties (Hummel, 1994; Palmer, 1978). The first is a set of primitive elements. These may be as simple as individual pixels (e.g., as found in some ideal observer models of object recognition; Liu, Knill & Kersten, 1995), as complex as volumetric parts (e.g., Marr & Nishihara, 1978) or assertions about the categorical properties of object parts (e.g., Biederman, 1987; Hummel & Biederman, 1992; Hummel, 2001), or more commonly, features of intermediate complexity, such as image edges and vertices (e.g.,

Lowe, 1987; Poggio & Edelman, 1990) or collections of edges and vertices (e.g., Fukushima & Miyake, 1982; Riesenhuber & Poggio, 2002).

The second defining property of a representation of shape is a reference frame within which the arrangement of an object's features or parts is specified. These may be object-centered (e.g., Marr & Nishihara, 1978), viewer-centered (e.g., Hummel & Biederman, 1992; Poggio & Edelman, 1990; Riesenhuber & Poggio, 2002; Ullman & Basri, 1991) or a mixture of the two (e.g., Hummel, 2001; Hummel & Stankiewicz, 1996a; Lowe, 1987; Ullman, 1989).

Mixtures of object- and viewer-centered reference frames come in two varieties. Three-dimensional (3-D) model-based models, such as those proposed by Lowe (1987) and Ullman (1989), hold that we mentally represent objects as detailed 3-D models, which we match to 2-D object images by a kind of "back projection" akin to the manner in which a three-dimensional model is mapped to a specific object view in computer graphics[1]. The viewed two-dimensional object image is then matched point-by-point against the 2-D view produced by the back projection. These models represent a mixture of object- and viewer-centered reference frames in the sense that the stored 3-D model is fully object-centered, but the 2-D back-projected view is matched to the object image in a viewer-centered reference frame.

The other way to mix reference frames is to specify some dimensions in viewer-centered terms and others in object-centered terms. This kind of mixture is possible because a reference frame is defined by three properties, any of which can be either viewer- or object-centered (or, for that matter, environment-centered, although this latter approach is never used): The *origin* is the point relative to which the reference frame is defined (in the viewer-centered case, the origin is defined relative to the viewer, and in the object-centered case, it defined relative to the object). A *scale* maps distances on the object or in the image to distances in the reference frame. And an *orientation* maps directions (in the image or on the object) to directions in the reference frame. A common way to mix object- and viewer-centered components is to define the origin and scale of the reference frame relative to the object but define the orientation relative to the viewer (Hummel, 2001; Hummel & Stankiewicz, 1996a; Olshausen, Anderson & Van Essen, 1993).

The third dimension of the representation of shape is a vocabulary of relations specifying the arrangement of the primitives within the reference frame. The most straightforward approach is simply to specify the locations of primitives in terms of their coordinates, i.e., their numerical relations to the origin of the reference frame. This approach characterizes "view-" or "appearance-based" models (Edelman & Intrator, 2001; Poggio & Edelman, 1990; Olshausen et al, 1993; Riesenhuber & Poggio, 2002; Ullman & Basri, 1991) as well as 3-D model-matching models (Lowe, 1987; Ullman, 1989). Alternatively, more complex primitives (such as convex object parts [Marr & Nishihara, 1978; Biederman, 1987] or surfaces [Leek Reppa & Arguin, 1995]) may be represented in terms of their relations, not to the origin of the reference frame, but to one another. Made explicit in this way, inter-part relations may be represented metrically or categorically (e.g., "above" vs. "below" and "larger" vs. "smaller"; Biederman, 1987; Hummel & Biederman, 1992) or both (e.g., Hummel, 2001; Hummel & Stankiewicz, 1996a, 1998). In

---

[1] Given a specification an object's coordinates in 3-D space (e.g., the coordinates of various features on its external surfaces) and given a specification of the angle and distance from which the object is viewed, the mathematics governing the projection of points on the object's visible surfaces to points in the resulting image of the object—i.e., the mathematics of projective geometry—are well constrained and have been well understood for a very long time. It is these mathematics that allow Pixar to create its movie magic.

general, representing inter-part relations explicitly provides tremendous representational flexibility, both in the vocabulary of relations that can be represented, and in the manner in which those relations can be specified.

The final dimension of a representation of shape, closely related to the issue of coordinates vs. relations, is the distinction between *holistic* vs. *analytic* representations. An analytic representation is one in which the components of the representation are made explicit and expressed relative to one another and/or to the representation as a whole. In the case of shape perception, a structural description specifying an object's shape in terms the spatial relations among its component parts would be an analytic representation. Other examples of analytic representations include propositions (e.g., *loves* (John, Mary)) and sentences (e.g., "John loves Mary"). In all three cases, the component parts (e.g., the object parts, in the case of the structural description, or the relation *loves* and the actors John and Mary in the proposition and sentence) are represented explicitly in terms of their relations to one another (e.g., a structural description might express which object parts are connected to one another, which are larger than others, etc.; the proposition and sentence both express that John is the lover and Mary the beloved). In turn, representing these components explicitly implies representing them independently of one another. The independence of John with Mary and *loves* is what makes it possible for us to understand what *loves* (John, Mary) has in common with *hates* (John, Mary) or *loves* (John, Susan), and how they differ (Hummel & Biederman, 1992; Hummel & Holyoak, 1997, 2003).

A holistic representation is simply the opposite: It is one in which the components of the representation are not represented independently of one another or of their place in the representation as a whole. Quintessential examples of holistic representations in human vision include face perception (see Rhodes & Peterson, 2003) and color perception. Face perception is holistic in the sense that we match faces "all of a piece" to representations stored in memory, with little or no explicit knowledge of a face's details. As a result, we are generally capable of saying who looks like whom, but barring obvious categorical features such as scars or facial hair, we are generally at a loss to say why. Similarly, color perception tends to be quite holistic as evidenced by the fact that we have difficulty categorizing colors in terms of their underlying physical dimensions of hue, saturation and brightness. Any color is some combination of these dimensions, but because we perceive them holistically we are at a loss to respond to them individually. (Contrast this situation with the case of "John loves Mary" or the case of a structural description specifying something like "cone on top of brick": Although we cannot easily say how one shade of red differs from another, we can easily say how "cone on top of brick" differs from "cylinder on top of brick".)

There is no necessary logical linkage between relations vs. coordinates on the one hand and analytic vs. holistic representations on the other: In principle, it is possible to imagine either analytic or holistic representations of either relations or coordinates. But as a matter of psychological fact, relational representations are analytic (i.e., when one thinks about relations, one is thinking explicitly; see e.g., Holyoak & Thagard, 1995; Hummel & Holyoak, 1997; Stankiewicz, Hummel & Cooper, 1998; Thoma, Hummel & Davidoff, 2004) whereas coordiate-based representations lend themselves naturally to holistic coding (e.g., as in the case of face perception, for which there is evidence of coordinate-based coding of a face's features; Cooper & Wojan, 2000). As such, the issue of explicit relations vs. coordinate-based coding in shape perception is de facto bound up with the issue of analytic vs. holistic representations of shape: Structural description models, which propose that we represent the relations among an object's parts explicitly, therefore propose that the representation of shape is analytic, whereas view-

based models, which represent objects in terms of their features coordinates, are naturally much more holistic. (At the same time, it is possible to reason explicitly, i.e., analytically, about coordinates, as when we solve problems in analytic geometry.)

Models of Object Recognition

Any model of object recognition is a collection of commitments to the four dimensions of shape representation summarized above. Although in principle the dimensions are independent of one another, in practice commitments to particular values on them tend to cluster.

One group of models, collectively referred to as *view*- or *appearance-based*, are based on simple primitives such as vertices (Edelman & Intrator, 2001; Olshausen et al., 1993; Poggio & Edelman, 1990; Ullman & Basri, 1991) or collections of contour elements (Riesenhuber & Poggio, 1999, 2002), represented in terms of their numerical coordinates in a viewer-centered reference frame (or at least partly viewer-centered; many view-based models assume that images are normalized for translation and scale prior to recognition). In some of these models (e.g., Edelman & Intrator, 2001, 2003; Poggio & Edelman, 1990; Ullman & Basri, 1991), the representation of object shape is taken to be the vector of the 2-D retinal coordinates of the vertices present in an object view; that is, the primitives themselves are not part of the representation at all except inasmuch as they determine which coordinates get into the vector. In these models, the primitives are simple local image elements (lines and/or vertices), the reference frame is viewer-centered, the relations are numerical coordinates and the representation is holistic.

As noted previously, a related group of models, *3-D model-matching* models (e.g., Lowe, 1987; Ullman, 1989), assumes that objects are represented in LTM as 3-D models but are matched to object images by back-projecting the models onto the image. These models assume an object-centered reference frame for the 3-D model and a viewer-centered reference frame for the 2-D matching process. Like the view-based models, matching in these models is based on local image features specified in terms of their retinotopic coordinates.

A third group of models assert that objects are represented as *structural descriptions* (e.g., Biederman, 1987; Clowes, 1967; Dickenson, Pentland & Rosenfeld, 1992; Hummel & Biederman, 1992; Marr & Nishihara, 1978; Sutherland, 1968), which specify an object's volumetric parts in terms of their relations to one another. In these models, the primitives are volumetric parts or part attributes (i.e., generalized cylinders [Dickenson et al., 1992; Marr & Nishihara, 1978] or geons, which are classes of generalized cylinders that can be discriminated from one another based on non-accidental properties of image edges [Biederman, 1987; Hummel, 2001; Hummel & Biederman, 1992; Hummel & Stankiewicz, 1996a]), the reference frame may either be object-centered (Marr & Nishihara, 1978), viewer-centered (Biederman, 1987; Hummel & Biederman, 1992) or mixed (Hummel, 2001; Hummel & Stankiewicz, 1996a), the relations between an object's parts are represented explicitly and the resulting representations are analytic. For example, a coffee mug might be represented as a curved cylinder (the handle) end-connected to the side of a straight vertical cylinder (the body of the mug).

Still a fourth class of models assumes a combination of the view- and structural-description-based approaches (Hummel, 2001; Hummel & Stankiewicz, 1996a). These models assume that the visual system generates part-based structural descriptions of attended objects and simultaneously generates view-like holistic representations (which are nonetheless invariant with translation and scale) of unattended objects.

Evaluating View- and Structure-based Models of Object Recognition

The view- and structural description-based approaches are the dominant approaches to modeling/theorizing about human object recognition and have motivated the majority of the theory-driven empirical research. Accordingly, I shall structure my review of the empirical literature around an evaluation of these two general theoretical approaches.

The structural description approach to object recognition was first proposed by Clowes (1967) and Sutherland (1968) and it was first proposed as a solution to the problem of view-invariant object recognition by Marr (1982; Marr & Nishihara, 1978). (Indeed, the roots of the structural description account, with its emphasis on relations rather than metric coordinates, date back at least as far as the Gestaltists; e.g., Köhler, 1940; Wertheimer, 1924/1950.) It was first popularized as a serious theory of human object recognition by Biederman (1987), who proposed that people recognize objects as collections of *geons*—classes of volumetric primitives that can be distinguished from one another based on non-accidental 2-D properties of image edges—in particular categorical relations to one another. The impact of Biederman's work was that it showed how the information necessary for view-invariant object recognition could be recovered from the information available in an object's 2-D image. Hummel and Biederman (1992) later showed how the fundamental operations proposed by Biederman (1987) could be accomplished in a neural computing architecture, and demonstrated that the resulting model successfully simulated the strengths and limitations of the human ability to recognize objects despite variations in viewpoint.

The view- and structure-based approaches to object recognition grew from very different starting points. With Marr & Nishihara's (1978) work, the structural description approach started with an analysis at Marr's (1982) computational theory level of how the problem of view-invariant recognition could be accomplished, in the abstract. Biederman (1987) provided a psychologically plausible algorithmic account of this general approach, and Hummel and Biederman (1992) demonstrated how the resulting algorithm could be realized in neural hardware. Central to (especially more recent) structural description accounts is the assumption that objects are represented in terms of their parts' spatial relations to one another.

The view-based approach was arguably inspired much more from the opposite end: From the very beginning (e.g., Poggio & Edelman, 1990), the goal was to understand how neural networks could accomplish pattern recognition, and neural plausibility remains a primary motivation behind much view-based modeling (see, e.g., Riesenhuber & Poggio, 1999, 2002; Edelman & Intrator, 2003). Fundamental insights came from Poggio and Girosi (1990) who showed that a Gaussian radial basis function, which is easy to instantiate in a neural computing architecture, is in some senses an optimal classifier, and from Ullman and Basri (1991), who showed that, provided some basic assumptions could be satisfied (e.g., that all features of an object be visible in all possible views of that object), an object could be recognized in any 3-D view as a linear combination of stored 2-D views of the object. In other words, view-invariant recognition of 3-D objects could be accomplished by simple vector operations (which are easy to implement in neural networks) on stored 2-D views without having to store, compute or even

estimate any of the object's 3-D properties. Central to all these modeling efforts is the assumption that objects are represented as vectors of features and/or feature coordinates.[2]

The effects of viewpoint on object recognition

Throughout most of the 1990s and into the early 2000s, an often acrimonious debate raged—both in the literature and at scientific meetings—between the proponents of view-based models of object recognition and the proponents of structural description models. This debate centered on the question of just how view-specific vs. view-invariant the representation of object shape really is. The proponents of structural description models argued (and marshaled evidence demonstrating) that the visual representation of shape is largely invariant with (i.e., unaffected by changes in) the location of an object's image in the visual field (Biederman & Cooper, 1991a), the size of the image (up to the limits of visual acuity; Biederman & Cooper, 1992), left-right reflection (Biederman & Cooper, 1991a) and rotation in depth up to parts occlusion (Biederman & Bar, 1999, 2000; Biederman & Gerhardstein, 1993, 1995; Biederman & Subramaniam, 1997; Hayworth & Biederman, 2005; Kayert, Biederman, Op De Beeck, & Vogels, 2005; Kayert, Biederman & Vogels, 2003; Vogels, Biederman, Bar & Lorincz, 2001).

By contrast, the proponents of view-based models of object recognition argued (and marshaled evidence demonstrating) that object recognition is largely sensitive to rotation in the picture plane (Tarr & Pinker, 1989; 1990) and rotation in depth (Edelman & Bülthoff, 1992; Hayward & Tarr, 1997; Logothetis, Pauls, Bülthoff, & Poggio, 1994; Poggio & Edelman, 1990; Tarr & Bülthoff, 1995; Tarr, Bülthoff, Zabinski & Blanz, 1997; Tarr, Williams, Hayward & Gauthier, 1998).

It turns out that the question of whether object recognition looks view-invariant or view-sensitive hinges largely on the stimuli and tasks used to ask the question (Biederman & Subramaniam, 1997; Liu, Knill & Kersten, 1995). With objects that can be easily segmented into distinct volumetric parts in distinct spatial relations, the visual representation of shape (as

---

[2] At this point, students of cognitive psychology may be scratching their heads and wondering, "Hey, isn't this *view-based* approach just the same thing as *template matching*, which I read about in my intro to cog psych textbook and which is obviously wrong as an account of human object recognition?" The answer is, "Well, yes and no." Template matching, the whipping-boy presented and dismissed in intro texts, is most often presented as matching, pixel-by-pixel, a viewed image of an object to a literal image stored in the head. Inasmuch as view-based models match feature coordinates rather than individual pixels, the answer is No, they are not the same (an object's image contains fewer "features" than pixels). However, in all other respects they are very much the same, in the sense of directly matching the exact 2-D shape of an object's image to a set of precise 2-D coordinates stored in memory. And in this sense, the view-based approach was basically falsified before Ullman & Basri (1991) or Poggio & Edelman (1990) published their first papers advocating the approach. At the time I, too, scratched my head and asked, "Hey, don't we already know what's wrong with this approach?" (see Hummel & Biederman, 1992; Hummel, 1994, 2000; Hummel & Stankiewicz, 1996b). But for whatever reason, the proponents of view-based models never bothered to cite the Gestaltists and others who had so long before demonstrated the fundamental inadequacy of their approach.

measured, for example, by visual priming) appears quite view-invariant, up to parts occlusion: One view of a volumetric object will visually prime a different view of the same object just as much as it primes itself, so long as all the same parts are visible in both views (Biederman & Gerhardstein, 1993). It is this kind of stimulus that Biederman and colleagues used in the majority of their studies demonstrating view-invariance. But with stimuli that are not easy to segment into distinct parts in distinct relations (e.g., the kind of object that would result from bending a paperclip in several places at various angles) the representation of shape is quite sensitive to the orientation from which the object is viewed. It is this kind of stimulus that Bülthoff, Tarr and colleagues used in the majority of their studies demonstrating view-sensitivity.[3]

The debate over the view-dependence vs. view-invariance of object recognition was fueled largely by the fact that view-based models (as suggested by the name) *tend to be* more view-sensitive than structural description models, whose early development (Marr & Nishihara, 1978; Biederman, 1987) was motivated largely by the question of how the visual system achieves view-invariant object recognition, and which therefore tends to predict substantial view-invariance in the visual representation of object shape. The question of view-dependence vs. view-invariance therefore *seemed to be* the deciding factor between these two classes of models.

But it is not (Hummel, 1994, 2000, 2003a; Hummel & Stankiewicz, 1996b, 1998). Models in either class can be modified to act more or less view-invariant as demanded by the data. That is, neither are view-based models fundamentally view-sensitive nor are structural description models fundamentally view-invariant. For example, the view-based model of Ullman and Basri (1991) stores only 2-D representations of individual object views in memory, but predicts *complete* view-invariance in object recognition provided enough such views are stored (where "enough" can be a surprisingly small number). 3-D model-matching models, which are very much like view-based models in how they match object images to models in memory, also predict complete invariance in object recognition. And it is perfectly possible to render a structural description model more view-sensitive simply by including, in an object's description, information about the angle in which it is viewed (see Hummel, 2001b; Stankiewicz, 2002). As

---

[3] In response to this situation, it is natural to ask "So: Who's right? Which kind of stimulus is 'better'?" By way of answering this question, it is worth noting that the use of bent paperclip-like stimuli was motivated by the fact that the view-based model of Ullman & Basri (1991) only works if all of an object's features are visible in all views. If one part of an object can occlude another part of that object (e.g., in the way that you cannot see the backs of most objects when you are looking at their fronts), then the mathematics that form the foundation of the Ullman & Basri model become undefined (i.e., the system breaks, resulting in a runtime error if you happen to be running the model on a computer). Bent paperclips have the desirable property that all their features (i.e., the places where they bend) are visible in (virtually) all possible views. In other words, bent paperclips were chosen as stimuli because they are the only kind of stimulus most view-based models of the time were capable of recognizing. They are a "good" choice of experimental materials to the extent that they are representative of other objects in our visual world. Rather than passing judgment myself, I invite the reader to list as many objects as he or she can that, like bent paperclips, have all their features visible in all views.

such, all the fighting about the view-sensitivity vs. view-invariance of object recognition that dominated the 1990s and early 2000s turned out to be much ado about not very much.

Although view-sensitivity is not the (or even a) fundamental difference between the view- and structure-based approaches to object recognition, there are four differences between these competing accounts that do turn out to be fundamental: (1) They differ in their commitment to strictly 2-D information vs. more 3-D inferences about the representation of object shape. (2) They differ in their approach to representing the configuration of an object's features or parts, with view-based models strongly committed to representing an object's features in terms of their numerical coordinates (see Hummel, 2000, for an explanation of why; in brief, it is because feature vectors are mathematically well-behaved whereas categorical relations are not) and structural description theories committed to representing an object's features or parts in terms of their (typically categorical) spatial relations to one another. (3) View-based models are committed to a holistic coding of object shape, in which all an object's features are represented as coordinates in the same, unitary feature vector, whereas structural descriptions are part-based analytic representations, which represent an object's part attributes independently of one another and of their interrelations (see Hummel, 2001, 2003b). And (4) because of (3), structural description models predict that the representation of an object's shape will differ qualitatively depending on whether it is attended or ignored; view-based models predict no qualitative effects of visual attention on the representation of object shape. On all four of these fundamental commitments, the view-based approach runs into difficulty as an account of human shape perception and object recognition. And on the fourth, so does the structural description approach.

The role of 2-D and 3-D representations of shape in object recognition

Liu, Knill & Kersten (1995) did a groundbreaking study demonstrating that no model based strictly on 2-D representations of object shape can provide an adequate account of human shape perception. They trained human observers to classify two kinds of objects: "bent paperclip"-like objects of the kind that have been used to demonstrate view-sensitivity in object recognition and novel volumetric objects of the kind that have been used to demonstrate view-invariance. They also trained two ideal observers (i.e., mathematical models that perform a task as well as it is logically possible to perform it with the information they are given) to classify the same objects. One of these ideal observers was a 2-D ideal: An ideal observer that, like a view-based model, stored only 2-D object views in memory and matched them against 2-D images for classification. Being an ideal observer, this model performed as well as it is logically possible to perform the object classification task using only stored 2-D views. The other ideal observer was a 3-D ideal. It made inferences about an object's 3-D shape based on the information in a 2-D view of an object and used these inferences as the basis for classifying the viewed images. Since ideal observers are ideal, human observers rarely perform as well they do, with efficiency (i.e., human accuracy divided by ideal accuracy) typically well below 0.5. The only way a human can outperform an ideal observer is by having access to information to which the ideal does not.

With the bent paperclip objects, Liu et al. (1995) found that the 2-D ideal observer outperformed the human observers (i.e., with human efficiency less than 1.0), a pattern consistent with the humans storing and matching 2-D views of these objects. However, with the volumetric objects, the human observers *outperformed* the 2-D ideal, indicating that the human observers were using information about 3-D shape to which the 2-D ideal did not have access. (Unsurprisingly, the humans still under-performed the 3-D ideal even with the volumetric

objects.) The human observers' ability to outperform the 2-D ideal with volumetric objects is very strong evidence that people do not recognize volumetric objects simply by storing and matching 2-D views of those objects.

The role of spatial relations in shape perception and object recognition

Another fundamental difference between the view- and structure-based approaches (a difference that also happens to characterize the difference between non-human and human primates; see Penn, Holyoak & Povenelli, 2008) concerns whether relations are represented explicitly or not. Structural descriptions (and humans) represent relations (e.g., relative size, relative location, relative orientation, etc.) explicitly; view-based models (and, apparently, non-human primates) do not. Along with language, the human ability to reason explicitly about relations is arguably the basis of virtually everything uniquely human, including art, science, mathematics and engineering (see Holyoak & Thagard, 1995; Hummel & Holyoak, 1997, 2003).

That we can perceive the spatial relations among an object's parts is undeniable and is evidenced, for example, by our ability to name them and to use them as the basis for making analogies between objects (see Hummel, 2000). And although the perception of spatial relations—and the role of explicit relations in object recognition—have received comparatively little empirical attention, what evidence there is suggests that, as predicted by the structural description approach, explicit representations of an object's inter-part spatial relations play an important role in the representation and encoding of object shape.

Hummel and Stankiewicz (1996b) directly tested categorical spatial relations against coordinate-based representations using a variety of object identification tasks (including basic recognition, same/different judgments and similarity judgments). Specifically, we trained human subjects to recognize a collection of stick-arrangement objects like those used by Tarr & Pinker (1989, 1990) in their demonstrations of the sensitivity of object recognition to orientation in the picture plane. We deliberately chose these objects because they do not have distinguishable parts in distinguishable relations (all their "parts" are simply straight lines of various lengths connected at right angles); that is, we designed the objects to be similar to the kinds of objects typically used to demonstrate view-based effects in studies of object recognition.

After training, we tested our subjects for their ability to distinguish the trained objects from two kinds of distractors (see Figure 1): One kind of distractor (*coordinate* distractors) was designed to maximize a distractor's similarity to a trained object in terms of the literal coordinates of its critical features (the line endpoints) while changing one categorical relation present in the trained object (e.g., one line might be moved so that it went from being *centered-above* the line to which it was connected to being *centered-below* the line to which it was connected, a change that affected exactly two sets of coordinates, namely, those of the endpoints of the moved line). The other kind of distractor (*categorical* distractors) were designed to distort more coordinates (four rather than two) but leave the categorical (e.g., *above/below*) relations between the parts intact.

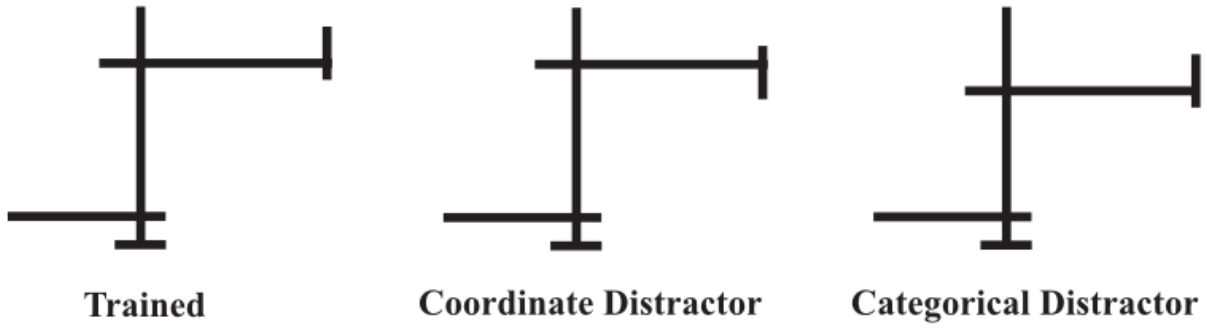

Figure 1. Illustration of the kind of stimuli used by Hummel & Stankiewicz (1996b). For each trained stimulus (Trained in the Figure) there were two distractorsi, one designed to match the trained stimulus in terms of its literal feature coordinates but mismatch in one categorical spatial relation (Coordinate Distractor) and one designed to mismatch the trained stimulus in twice as many stimulus coordinates but to match the trained stimulus in terms of the categorical relations among its parts (Categorical Distractor).

To the extent that people represent the trained objects in terms of the coordinates of their critical points, they should find the coordinate distractors more confusable with the trained targets than the categorical distractors. But to the extent that they represent the trained targets in terms of the categorical relations among their parts, they should find the categorical distractors more confusable with the corresponding trained targets than the coordinate distractors. The results were not subtle. Across a variety of tasks, our subjects found the categorical distractors much more confusable with the corresponding targets than the coordinate distractors.

Cooper and Wojan (2000) conducted a similar study with faces as stimuli. Their interest was in whether face identification (as in "Who's face is this?") is based on a coordinate-based representation whereas basic-level face recognition (as in "Is this a face or not?") is based on an explicit representation of the categorical relations among the face's parts. Their logic was similar to that of Hummel and Stankiewicz (1996b): They showed subjects either undistorted images of the faces of famous people, distortions in which one eye was moved up from its normal location on the face (changing its categorical spatial relation to the other eye but disrupting the coordinates of only one eye), or distortions in which both eyes had been moved up from their normal locations (preserving the categorical spatial relation between the eyes, but distorting twice as many feature coordinates as in the one-eye-moved case). To the extent that people use a coordinate-based representation, the two-eyes-moved images should be harder to recognize than the one-eye-moved images; but to the extent that they use a representation based on the categorical relations between the features of a face, the one-eye-moved images should be harder to recognize. Given a face identification task ("who is this?"), Cooper and Wojan found that subjects had more difficulty with the two-eyes-moved than the one-eye-moved stimuli, suggesting that face identification is based on a holistic, coordinate-based representation of the stimulus face (consistent with the vast literature on face identification; see, e.g., Rhodes & Peterson, 2003). But given a face recognition task ("is this a face or not?"), the one-eye-moved stimuli were harder to classify than the two-eyes-moved stimuli, consistent with a representation based on explicit categorical relations between the faces' parts.

Turning to object category learning, Saiki and Hummel (1996) showed that object categories defined by conjunctions of parts in specific spatial relations are much easier to learn

than categories defined by conjunctions of parts in specific colors or colors in specific relations, even when the learnability of categories defined by colors, parts and relations, individually, is equated. In other words, the human object category learning system is biased to learn objects consisting of parts in specific spatial relations. In related research, Saiki and Hummel (1998a) showed that people perceive the spatial relations among connected objects parts differently than the relations among non-connected parts (e.g., among other things, people are more sensitive to part-relation bindings for connected than non-connected parts).

In summary, the comparatively few data there are on the representation of spatial relations for object recognition suggest that people do explicitly represent the spatial relations among an object's parts in the service of object recognition and object category learning.

Independence of part attributes in the representation of shape

One of the most basic predictions of Biederman's (1987) theory of object recognition is that people will activate a representation of an object's convex parts on the way to recognizing the object. In a now classic collection of experiments, Biederman and Cooper (1991b) used a visual priming paradigm to directly test this prediction. In Experiment 1, Biederman and Cooper formed complementary images of objects by starting with line drawings and removing half the contours and vertices from each of an object's parts. The removed contours and vertices were then used to make a second line drawing of the same object. The two line drawings were complements of one another in the sense that, although they depicted all the same volumetric parts, they had no local features (contours and vertices) in common. Biederman and Cooper showed that one member of an image-complement pair would visually prime the other member of the pair just as much as it primed itself: From the perspective of the visual priming, there was no difference between the two images, even though they had no local features in common.

In Experiment 2, Biederman and Cooper (1991b) formed complementary images by removing half the volumetric parts from one image in order to form the other. As with the complementary images from Experiment 1, together the two complements formed a complete line drawing of the object. But in Experiment 2, in contrast to Experiment 1, the two images depicted none of the same parts. This time, Biederman and Cooper found that complementary images did not visually prime one another at all. Taken together, the results of Experiments 1 and 2 show that two line drawings of the same object will visually prime one another if and only if they depict the same volumetric parts—a pattern that constitutes strong evidence for some kind of parts-based representation of shape being activated in the service of object recognition. Based on Biederman and Cooper's data it is impossible to know whether the observed priming resides in a representation of volumetric parts, per se, vs. whether it resides in a representation of the surfaces composing those volumes (Leek, Reppa & Arguin, 2005). But regardless, it is clear that the priming resides in a representation of parts (surfaces or volumes) substantially more abstract than simple local image features.

Part of representing the relations among an object's parts explicitly is representing them independently of the parts so related (Arguin & Saumier, 2005; Hummel, 1994, 2000; 2001; Hummel & Biederman, 1992; Hummel & Holyoak, 1997, 2003). A corollary of this principle is that the visual system is well advised to represent high-level shape attributes, such as the attributes describing the shape of a geon, independently as well. In Hummel & Biederman's (1992) structural description model, JIM, as well as in JIM's successors, JIM2 (Hummel & Stankiewicz, 1996) and JIM3 (Hummel, 2001), this independence takes the form of separate

units representing the various shape attributes of an object part (e.g., one set of neuron-like units represents whether a geon's cross section is straight or curved, a separate set represents whether its major axis is straight or curved, and still other units represent its other shape attributes, its location in the visual field, etc.) and the relations between parts (e.g., separate units represent relations such as *above*, *below*, *larger*, *smaller*, etc.). This independence makes it possible to appreciate what different geons have in common and how they differ. Representing a cone-shaped part simply as the atom "cone" and a cylinder simply as "cylinder" fails to specify what cones have in common with cylinders. But representing a cone as "curved cross-section, straight major axis, non-parallel sides" and a cylinder as "curved cross section, straight major axis, parallel sides" makes explicit how a cone is similar to a cylinder and how they differ.

Behaviorally, independence predicts that part shape attributes ought to be perceptually separable from one another and from the relations in which a part is engaged. Using Garner's (1974) criteria, Saiki and Hummel (1998b, Experiment 3) showed that object parts are perceived independently of their spatial relations, and using a visual search paradigm, Arguin and Saumier (2005) also found evidence for independent coding of object parts and their spatial relations. Using a noise-masking paradigm, Stankiewicz (2002) showed that part attributes such as aspect ratio and axis curvature are represented independently of one another. Stankiewicz also showed that part shape is perceptually independent of viewpoint. As predicted by the structure-based account, these findings indicate that part attributes and their relations are represented explicitly and independently of one another. These findings are inconsistent with the view-based account's prediction that all aspects of object shape should be perceptually integral with (i.e., non-separable from) one another and with the viewpoint in which the object is depicted.

Neurally, the independence implied by the structure-based account predicts that there ought to be neurons in visual cortex that respond to individual attributes of geons, independently of the geon's other attributes. Kayaert, Biederman and Vogels (2005) report evidence for just such neurons in macaque inferotemporal cortex (specifically, in area TE). The same researchers also found that these neurons are more sensitive to changes in a geon's non-accidental properties (e.g., a change from a straight to a curved major axis) than to otherwise equivalent (e.g., in terms of their impact on early visual representations in V1) metric changes in a geon's shape (e.g., in the degree of curvature of the major axis). Similar findings of independence were reported by Tanaka (1993), Tsunoda et al. (2001) and Baker, Behrmann and Olson (2002) using different stimulus materials.

The role of attention in shape perception and object recognition

Representing visual properties independently of one another (i.e., analytically, rather than holistically) offers numerous advantages from a computational perspective (Hummel & Biederman, 1992), but it comes with a cost: If part attributes and relations are represented independently, then visual attention is necessary to ensure that they are bound together correctly (Hummel & Stankiewicz, 1996a; see also Treisman & Gelade, 1980; Treisman & Schmidt, 1982). It is here that the structure-based account runs into difficulty. This account correctly predicts that attention ought to be necessary to generate a structural description of object shape (Stankiewicz, Hummel & Cooper, 1998; Stankiewicz & Hummel, 2002; Thoma, Davidoff & Hummel, 2007; Thoma, Hummel & Davidoff, 2004). However, it incorrectly predicts that, without the ability to generate a structural description, recognition of unattended objects should fail.

Evidence for the recognition of unattended objects comes in the form of both negative (Tipper, 1985; Treisman & DeSchepper, 1996) and positive (Stankiewicz et al., 1998; Stankiewicz & Hummel, 2002; Thoma et al., 2007; Thoma et al., 2004) priming for unattended shapes. (Negative priming is an increase in response times and/or error rates to recognize previously seen items; positive priming [typically called simply "priming"] is a decrease in response times and/or error rates to recognize previously seen items. Whether priming for ignored objects is negative or positive depends on the difficulty of the selection task, with difficult selection [e.g., naming the green object in a display depicting a green line drawing of one object overlaid on a red line drawing of another] leading to negative priming and easy selection [e.g., naming one of two spatially separated line drawings] leading to positive priming; see Stankiewicz et al., 2008.) Importantly, priming for ignored objects is not localized in low-level retinotopic features (e.g., as represented in V1). Tipper (1985) showed that negative priming for ignored objects extends all the way to the level of meaning, and Stankiewicz and Hummel (2002) showed that, in contrast to the visual features represented in V1 and V2, positive priming for unattended objects in invariant with translation and scale.

The view-based approach correctly predicts priming for unattended objects (since views are holistic, they do not depend on attention for binding; see Hummel & Stankiewicz, 1996a), but incorrectly predicts that the representation of an unattended shape should be the same as the representation of an attended shape. That is, whereas the structure-based account predicts too great a role for attention in object recognition (i.e., that recognition will fail completely in the absence of attention), the view-based account predicts too little.

Stankiewicz, Thoma and their colleagues have shown that the visual representation of an attended object differs qualitatively from that of an unattended object. Whereas visual priming for both attended and unattended object images is invariant with translation across the visual field and changes in scale (Stankiewicz & Hummel, 2002), attended images visually prime their left-right reflections whereas ignored images do not (Stankiewicz et al., 1998). Similarly, visual priming for attended images is robust to configural distortions (as when an image is cut with a vertical line down the center and the left half is moved to the right and the right half to the left) whereas priming for ignored images is not (Thoma et al., 2004). Attended images also prime their inverted (i.e., upside-down) counterparts, whereas ignored images do not (Thoma et al., 2007).

The findings of Stankiewicz, Thoma and colleagues can be accommodated neither by the structural account nor by the view-based account. Indeed, they were predicted (and the experiments themselves motivated) by a hybrid model (Hummel & Stankiewicz's, 1996, JIM2 and Hummel's, 2001, refinement, JIM3), whose development was a response to the strikingly complementary strengths and limitations of the structure- and view-based accounts. Hummel and Stankiewicz were motivated to develop JIM2 by the observation that structure-based models account for a large number of findings in the literature on object recognition and shape perception (as reviewed above), while at the same time being inconsistent with the automaticity and speed of object recognition. (Although not summarized above, object recognition is not only too automatic to be consistent with the structure-based approach, it is also too fast: At least for overlearned stimuli, object-sensitive neurons in macaque inferotemporal cortex will respond to their preferred stimulus in a feed-forward fashion [Tovee et al, 1993]—processing that is much too fast to depend on the dynamic binding necessary to generate a structural description [Hummel & Stankiewicz, 1996a].)

The resulting model is a hybrid that generates geon-and-relation-based structural descriptions of attended object images but which uses a holistic "surface map" (akin to a view-based representation in the sense that features are represented at each of several locations in a reference frame, but at much lower spatial resolution than a view-based model and with more complex primitives than simple image features) to recognize familiar objects in familiar views rapidly and without the need for visual attention (see Hummel, 2001, for the most recent version of the model). In general, the model predicts that visual priming for attended images will reflect the properties of both the structured representation of shape and the holistic representation (both of which become active in response to an attended image), whereas priming for ignored images will reflect the properties of the holistic surface map only.

More specifically, the model makes a complex set of predictions about the effects of attention and images changes on visual priming in an object recognition task (predictions that, at the time of its first publication in 1996, were untested). As reviewed above, the model predicts that: (a) priming for both attended and ignored images will be invariant with translation across the visual field and changes in scale (confirmed by Stankiewicz & Hummel, 2002); (b) attended images, but not ignored images, will visually prime their left-right reflections (confirmed by Stankiewicz et al., 1998); (c) attended images will prime configural distortions of themselves, but ignored images will not (confirmed by Thoma et al., 2004); and (d) attended images will prime inverted versions of themselves but ignored images will not (confirmed by Thoma et al., 2007).

Summary and Conclusions

Human beings have large brains and roughly half of the human brain (or a bit more) is involved in one way or another in processing visual information. Given this, it is perhaps unsurprising that we are extremely good at visually recognizing objects. When we attend to an object we can visually segment it into its parts, perceive the attributes of those parts independently of one another and independently of the parts' relations to one another, and perceive and reason about the relations among the parts. Armed with these relational representations, we can recognize familiar objects in novel viewpoints and recognize novel instances of known object classes. We can also reason about the attributes of an object, its parts and the relations among them, in the service of making judgments about whether the object would make, say, an appropriate doorstop, a useful weapon or a charming element in a modern sculpture to serve as a gift to our spouse on the occasion of our anniversary. All these abilities rely upon and reflect our capacity to represent objects relationally. Our relational representations of objects serve as a perceptual gateway to our ability to reason relationally about the world, which, in turn, sets us apart from all other primate species (and quite probably *all* other species).

At the same time, if all we had were these relational representations, object recognition would be attention-demanding, labor-intensive, calorie-consuming and frustratingly slow, as we would have to attend to every object in our environment in order to recognize it. So we appear to have evolved a shortcut (or, in all likelihood, the "shortcut" evolved before the relational ability, so it is probably more accurate to say that we simply did not evolve away our shortcut): Familiar objects in familiar views can be recognized rapidly and automatically, freeing up our attentional resources to focus on more interesting matters.

Together, the automatic holistic route to recognition and the effortful relational one give us the best of both worlds: We can be as fast and automatic as a view-based model most of the time, and as smart as a structural description when we need to be.

References

Arguin, M. & Saumier, D. (2005). Independent processing of parts and of their spatial organisation in complex visual objects. *Psychological Science, 15*, 629-633.

Baker, C., Behrmann, M., & Olson, C. (2002). Influence of visual discrimination training on the representation of parts and wholes in monkey inferotemporal cortex. *Nature Neuroscience*, *5*, 1210-1216.

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

Biederman, I., & Bar, M. (1999). One-shot viewpoint invariance in matching novel objects. *Vision Research*, *39*, 2885-2899.

Biederman, I., & Bar, M. (2000). Views on views: Response to Hayward & Tarr (2000). *Vision Research*, *40*, 3901-3905.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, *20*, 38-64.

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 recognition. *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 and conditions for 3D 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 reply to Tarr & Bülthoff (1995). *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 1506-1514.

Biederman, I. & Subramaniam, S. (1997). Predicting the shape similarity of objects without distinguishing viewpoint invariant properties (VIPs) or parts. *Investigative Ophthalmology and Visual Science*, *38*, 998.

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., & Wojan, T. J. (2000). Differences in the coding of spatial relations in face identification and basic-level object recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26,* 470-488.

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.

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.

Fukushima, K. & Miyake, S. (1982). Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. Pattern Recognition, 15, 455-469.

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

Goodale, M. A., Milner, D. A., Jakobson, L. S., & Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature*, *349*, 154-156.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.

Hummel, J. E. (2000). Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich & A. Markman (Eds.). *Cognitive dynamics: Conceptual change in humans and machines* (pp. 157 - 185). Mahwah, NJ: Erlbaum.

Hummel, J. E. (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual Cognition*, *8*, 489 - 517.

Hummel, J. E. (2003a). Effective systematicity in, effective systematicity out: A reply to Edelman & Intrator (2003). *Cognitive Science*, *27*, 327-329.

Hummel, J. E. (2003b). The complementary properties of holistic and analytic representations of object shape. In G. Rhodes and M. Peterson (Eds.), *Perception of faces, objects, and scenes: Analytic and holistic processes* (pp. 212-234). Westport, CT: Greenwood.

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., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220-264.

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. (1998). Two roles for attention in shape perception: A structural description model of visual scrutiny. *Visual Cognition*, *5*, 49-79.

Liu Z, Knill D C, Kersten D (1995). Object classification for human & ideal observers. *Vision Research* 35: 549-568.

Lowe, D. G. (1987). The viewpoint consistency constraint. *International Journal of Computer Vision*, *1*, 57-72.

Kayaert, G., Biederman, I., & Vogels, R. (2005). Representation of regular and irregular shapes in macaque inferotemporal cortex. *Cerebral Cortex*, *15*, 1308-1321.

Kayaert, G., Biederman, I., Op De Beeck, H., & Vogels, R. (2005). Tuning for shape dimensions in macaque inferior temporal cortex. *European Journal of Neuroscience*, *22*, 212-224.

Kobatake, E., & Tanaka, K. (1994). Neuronal selection to complex object features in the ventral visual pathway of the macaque visual cortex. *Journal of Neuroscience*, *71*, 856-867.

Köhler, W. (1940). *Dynamics in psychology*. New York: Liveright.

Leek, E.C., Reppa, I. & Arguin, M. (2005). The structure of 3D object shape representations: Evidence from whole-part matching. *Journal of Experimental Psychology: Human Perception and Performance, 31*, 668-684.

Logothetis, N. K., Pauls, J., Bülthoff, H. H., & Poggio, T. (1994). View-dependent object representation by monkeys. *Current Biology*, *4*, 401-414.

Mapelli, D. and Behrmann, M. (1997). The role of color in object recognition: Evidence from visual agnosia. *NeuroCase*, *3*, 237-247.

Marr, D. (1982). *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.

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. *Journal of Neuroscience, 13,* 4700-4719.

Op de Beeck, H., Beatse, E., Wagemans, J., Sunaert, S., & Van Hecke P. (2000). The representation of shape in the context of visual object categorization tasks. *Neuroimage*, *12*, 28 - 40.

Palmer, S. E. (1978). Fundamental aspects of cognitive representation. In E. Rosch and B. B. Lloyd (Eds.) *Cognition and Categorization*. (pp. 259-303). Hillsdale, NJ: Lawrence Erlbaum.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: MIT Press.

Penn, D. C., Holyoak, K. J., & Povinelli, D., J. (2008). Darwin's mistake: Explaining the discontinuity between human and non-human minds. *Behavioral and Brain Sciences*, *31*, 109-178.

Poggio, T. & Girosi, F. (1990). Regularization algorithms that are equivalent to multilayer networks. *Science*, *277*, 978-982.

Rhodes, G. & Peterson, M. (2003). *Perception of faces, objects, and scenes: Analytic and holistic processe*. Westport, CT: Greenwood.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*, 217-236.

Saiki, J., & Hummel, J. E. (1996). Attribute conjunctions and the part configuration advantage in object category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1002-1019.

Saiki, J. & Hummel, J. E. (1998). Connectedness and the integration of parts with relations in shape perception. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 227-251.

Stankiewicz, B.J. & Hummel, J.E. (2002) The role of attention in scale- and translation-invariant object recognition. *Visual Cognition*, *9*, 719-739.

Stankiewicz, B. J., Hummel, J. E., & Cooper, E. E. (1998). 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, 24*, 732-744.

Sutherland, N. S. (1968). Outlines of a theory of visual pattern recognition in animals and man. *Proceedings of the Royal Society of London* (Series B), *171*, 95-103.

Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science*, *262*, 685-688.

Tarr, M. J., Bülthoff, H. H., Zabinski, M., & Blanz, V. (1997). To what extent do unique parts influence recognition across viewpoint? *Psychological Science*, *8*, 282-289.

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience*, *1*, 275-277.

Tipper, S. P. (1985). The negative priming effect: Inhibitory effects of ignored primes. *Quarterly Journal of Experimental Psychology*, *37A*, 571-590.

Tovee, M. J., Rolls, E. T., Treves, A., & Bellis, R. P. (1993). Information encoding and the responses of individual neurons in the primate temporal visual cortex. *Journal of Neurophysiology*, *70*, 640 – 654.

Treisman, A. & DeSchepper, B. (1996). Object tokens, attention, and visual memory. In T. Inui and J. McClelland (Eds.), *Attention and Performance XVI: Information Integration in Perception and Communication*. (pp. 15-46). Cambridge, MA: MIT Press.

Treisman, A. & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, *12,* 97-136.

Treisman, A.M. & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, *14*, 107-141.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, *32*, 193-254.

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

Wertheimer, M. (1924/1950). Gestalt theory. In W. D. Ellis (Ed.), *A sourcebook of Gestalt psychology* (pp. 1-11). New York: The Humanities Press.