Visual Object Recognition

Michael J. Tarr and Quoc C. Vuong
Department of Cognitive and Linguistic Sciences
Box 1978
Brown University
Providence, RI 02912

The study of object recognition concerns itself with a two-fold problem. First, what is the form of visual object representation? Second, how do observers match object percepts to visual object representations? Unfortunately, the world isn't color coded or conveniently labeled for us. Many objects look similar (think about four-legged mammals, cars, or song birds) and most contain no single feature or mark that uniquely identifies them. Even worse, objects are rarely if ever seen under identical viewing conditions: objects change their size, position, orientation, and relations between parts, viewers move about, and sources of illumination turn on and off or move. Successful object recognition requires generalizing across such changes. Thus, even if an observer has never seen a bear outside of the zoo, on a walk in the woods they can tell that the big brown furry object with teeth 20 ft in front of them is an unfriendly bear and probably best avoided or that the orange-yellow blob hanging from a tree is a tasty papaya.

Consider how walking around an object alters one's viewing direction. Unless the object is rotationally symmetric, for example, a cylinder, the visible shape of the object will change with observer movement – some surfaces will come into view, other surfaces will become occluded and the object's geometry will change both quantitatively and qualitatively (Tarr & Kriegman, 2001). Changes in the image as a consequence of object movement are even more dramatic – not only do the same alterations in shape occur, but the positions of light sources relative to the object also change. This alters both the pattern of shading on the object's surfaces and the shadows cast by some parts of the object on other parts. Transformations in size, position, and mean illumination also alter the image of an object, although somewhat less severely as compared to viewpoint/orientation changes.

Recognizing objects across transformations of the image. Theories of object recognition must provide an account of how observers compensate for a wide variety of changes in the image. Although theories differ in many respects, most attempt to specify how perceptual representations of objects are derived from visual input, what processes are used to recognize these percepts, and the representational format used to encode objects in visual memory. Broadly speaking, two different approaches to these issues have been adopted. One class of theories assumes that there are specific invariant cues to object identity that may be recovered under almost all viewing conditions. These theories are said to be *viewpoint-invariant* in that these invariants provide sufficient information to recognize the object regardless of how the image of an object changes (within some limits) (Marr & Nishihara, 1978; Biederman, 1987). A second class of theories argues that no such general invariants exist[1] and that object features are represented much as they appeared when originally viewed, thereby preserving *viewpoint-dependent* shape information and surface appearance. The features visible in the input image are compared to features in object representations, either by normalizing the input image to approximately the same viewing position as represented in visual memory (Bülthoff & Edelman, 1992; Tarr, 1995) or by computing a statistical estimate of the quality of match between the input image and candidate representations (Perrett, Oram, & Ashbridge, 1998; Riesenhuber & Poggio, 1999).

Viewpoint-invariant and viewpoint-dependent approaches make very different predictions regarding how invariance is achieved (these labels are somewhat misleading in that the goal of *all* theories of recognition is to achieve invariance, that is, the successful recognition of objects across varying viewing conditions). Viewpoint-invariant theories propose that recognition is itself invariant across transformations. That is,

[1] Of course invariants can be found under certain contexts. For example, if there are only three objects to be distinguished and these objects are red, green, and blue, object color becomes an invariant in this context (Tarr & Bülthoff, 1995).

across changes in viewpoint, illumination, etc. there is no change in recognition performance – so long as the appropriate invariants are recoverable, the response of the system remains constant. In comparison, viewpoint-dependent theories hypothesize that recognition is dependent on specific viewing parameters. That is, across changes in viewpoint, illumination, etc. there may be changes in recognition performance – because objects are represented according to how they appeared when originally learned.

Recognizing objects across different instances of a class. Generalization across object views is only one of several demands placed on the visual recognition system. A second factor to consider is the ability to generalize across different instances of a visual object class. Because such instances are typically treated as members of the same category and should elicit the same response, an observer must be able to use one or more instances of a class to recognize new instances of the same class – the bear we are staring at is unlikely to be the same one we saw on our last trip to Alaska or "Animal Planet", but it would be fatal to not realize that this too is a bear. At the same time an observer must not confuse somewhat similar objects that should be recognized as different individuals – it is important that one does not confuse poisonous Ugli fruit with edible papayas. Marr and Nishihara (1978) termed these two goals of object recognition *stability* and *sensitivity*, respectively. These two goals seem to trade-off against one another: as recognition abilities become more stable, that is, the more one can generalize across objects, the less an observer may be able to distinguish between those objects. Conversely, as recognition abilities become more sensitive, that is, the better one is at telling objects apart, the worse an observer may be at deciding that two objects are really the same thing. How to deal with these two competing goals is at the heart of most theories of object recognition and central to the debate that has ensued about the "correct" theory of human recognition abilities.

Recognizing objects at different levels of specificity. A third factor that is important to developing such theories is the nature of recognition itself. That is, what exactly is the recognition task? Consider that one might spy a bear in the distance and want to know a variety of different things: Recognition – is that a bear? Discrimination – is that a bear or a person wearing a fur hat? Is that a grizzly or a brown bear? Identification – is that the friendly Gentle Ben? The key point highlighted by these different recognition tasks is that objects may be visually recognized in different ways and, critically, at different *categorical levels*. In the cognitive categorization literature, there is a distinction made between the superordinate, basic, subordinate, and individual levels (Rosch et al., 1976). A bear can be classified as an animal, as a bear, as a grizzly bear, and as Gentle Ben – each of these labels corresponding respectively to a different categorization of the same object.

Visual recognition can occur roughly at these same levels, although visual categorization is not necessarily isomorphic with the categorization process as studied by many cognitive psychologists. For example, some properties of objects are not strictly visual, but may be relevant to categorization – chairs are used to sit on, but there is no specific visual property that defines "sitability." A second distinction between visual and cognitive categorization is the default level of access. Jolicoeur, Gluck, and Kosslyn (1984) point out that many objects are not recognized at their basic level. For example, the basic level for pelicans is "bird," but most people seeing a pelican would label it "pelican" by default. This level, referred to as the "entry level," places a much greater emphasis on the similarities and differences of an object's visual features relative to other known objects in the same object class (Murphy & Brownell, 1985). The features of pelicans are fairly distinct from those of typical birds, hence, pelicans are labeled first as "pelicans"; in contrast, the features of sparrows are very typical and sparrows are much more likely to be labeled as "birds."

Why are these distinctions between levels of access important to object recognition? As reviewed in the next section, there are several controversies in the field that center on issues related to categorical level. Indeed, the stability/sensitivity tradeoff discussed above is essentially a distinction about whether object recognition should veer more towards the subordinate level (emphasizing sensitivity) or the entry level (emphasizing stability). This issue forms the core of a debate about the appropriate domain of explanation (Tarr & Bülthoff, 1995; Biederman & Gerhardstein, 1995). That is, what is the default (and most typical) level of recognition? Furthermore, the particular recognition mechanisms applied by default may vary with experience, that is, perceptual experts may recognize objects in their domain of expertise at a more specific level than novices (Gauthier & Tarr, 1997a). Some theorists – most notably Biederman (1987) – presuppose that recognition *typically* occurs at the entry level and that any theory of recognition should concentrate on accounting for how the visual system accomplishes this particular task. In contrast, other theorists – Bülthoff, Edelman, Tarr, and others (see, Tarr & Bülthoff, 1998; Hayward & Williams, 2000) – argue that that the hallmark of human recognition abilities is *flexibility* and that any theory of recognition should account for how the visual system can recognize objects at the entry, subordinate, and individual levels (and anything in between). This distinction is almost isomorphic with the viewpoint-invariant/viewpoint-dependent distinction raised earlier. Specifically, viewpoint-invariant theories tend to assume the entry level as the default and concentrate on accounting for how visual recognition at this level may be achieved. In contrast, viewpoint-dependent theories tend to assume that object recognition functions at many different categorical levels, varying with context and task demands.

A second, somewhat related, debate focuses on the scope of posited recognition mechanisms. Some theorists argue that there are at least two distinct mechanisms available for recognition – generally breaking down along the lines of whether or not the recognition discrimination is at the entry or the subordinate level (Jolicoeur, 1990; Farah, 1992). Some researchers have suggested that there may be several "special purpose devices" devoted to the task of recognizing specific object classes, for example, a neural module for face recognition, another for place recognition, and one for common object recognition (Kanwisher, 2000). Alternatively, it has been argued that recognition at many levels and for all object categories can be accomplished by a single, highly plastic system that adapts according to task constraints and experience (Tarr & Gauthier, 2000; Tarr, in press). This and the aforementioned debates have produced an extensive research literature addressing the nature of visual object recognition. In order to better understand these controversies, we next review the particular dimensions typically used both to characterize object representations and to constrain potential mechanisms of recognition.

The Nature of Object Representations

There is an overwhelming body of evidence addressing the nature of representation in object recognition and visual cognition. Despite this, there is an alarming absence of a comprehensive account of object recognition. Rather, as outlined above, most theorists have more or less tried to develop a framework along a particular subset of issues in order to frame a particular theory (Biederman, 1987; Edelman, 1997; Marr & Nishihara, 1978; Pinker, 1984; Poggio & Edelman, 1990; Tarr, 1995). Moreover, while there have been some attempts to integrate low- and mid-level visual processing into theories of object perception and recognition (Marr, 1982), most researchers have restricted themselves to a narrower problem that isolates recognition mechanisms from the processing that precedes it.

The range of behavioral and neural data indicates that the representations of objects and the mechanisms used to recognize objects are highly flexible (Tarr & Black, 1994). This presents a challenge to any theory of object recognition (including those that argue for flexibility, but cannot explain it). Selecting a representational format typically depends on how a particular theory addresses a broad set of interdependent issues, including the factors reviewed above. The critical issues include: (1) the features of the representation, (2) the degree to which 3D structure, if any, is encoded, (3) the spatial relationships among features within the representation, (4) the frame of reference used to specify the locations of features, and (5) the normalization mechanisms, if any, used to operate on the input image or on the representation. Together, these issues are crucial to understanding the problem of object recognition; they also provide metrics by which the strengths and weaknesses of theories can be identified (Bülthoff & Edelman, 1993; Hummel, 1994).

The Nature of Object Features

What are features? A loose definition is that features are the elementary units used in the representation of objects (Marr & Nishihara, 1978). This definition, however, leaves open a wide range of possible feature types, from local features that measure metric properties of objects at specific locations to global features that only represent qualitative characteristics of objects. Examples of local features include receptive field responses measuring brightness or color, oriented lines, T-junctions, corners, etc. (Tanaka, 1996). Examples of global features include 3D component parts realized as simple volumes that roughly capture the actual shape of an object (Marr & Nishihara, 1978; Biederman, 1987).

Immediately, significant differences between these two approaches are apparent. On the one hand, an appealing aspect of local features is that they are readily derivable from retinal input and the natural result of earlier visual processing as discussed in prior chapters in this volume. In contrast, 3D parts must be recovered from 2D images in a manner that is not entirely obvious given what is currently known about visual processing. On the other hand, it is hard to imagine how stability is achieved using only local features – the set of features visible in one viewpoint of an object is likely to be very different from the feature sets that are visible in other viewpoints of the same object or other similar objects. Even slight variations in viewpoint, illumination, or configuration may change the value of local responses and, hence, the object representation. Furthermore, 3D parts yield stability – so long as the same invariants are visible, the same set of 3D parts may be recovered from many different viewpoints and across many different instances of an object class. Thus, variations in viewpoint, illumination, or configuration are likely to have little impact on the qualitative representation of the object.

Dimensionality

The range of features that may form the representation is quite wide, but cutting across all possible formats is their degree of dimensionality, that is, how many spatial dimensions are encoded. The physical world is three-dimensional, yet the optic array sampled by the retinae is two-dimensional. As discussed in earlier chapters, one goal of vision is to recover properties of the 3D world from this 2D input (Marr, 1982). Indeed, 3D perception seems critical for grasping things, walking around them, playing ping-pong, etc. However, recovery of 3D shape may not be critical to the process of remembering and recognizing objects. Thus, one can ask whether object representations are faithful to the full 3D structure of objects or to the 2D optic array, or to something in between. As discussed above, some theories argue that complete, 3D models of objects are recovered (Marr & Nishihara, 1978) or that object representations are 3D, but can vary depending on the features visible from different viewpoints (Biederman, 1987). Others have argued that object representations are strictly 2D; that is, preserving the appearance of the object in the image with no reference to 3D shape or relations (Edelman, 1993). An intermediate stance is that object representations are not strictly 2D or 3D, but rather represent objects in terms of visible surfaces, including local depth and orientation information. Such a representation is sometimes termed "two-and-one-half-dimensional" (2.5D; Marr, 1982). Critically, both 2D and 2.5D representations only depict surfaces visible in the original image – there is no recovery or reconstruction or extrapolation about the 3D structure of unseen surfaces or parts; 3D information instead arises from more local processes such as shape-from-shading, stereo, and structure-from-motion. In contrast, 3D representations include not only surface features visible in the input (the output of local 3D recovery mechanisms) but also additional globally recovered information about an object's 3D structure (e.g., the 3D shape of an object part). Such 3D representations are appealing because they encode objects with a structure that is isomorphic with their instantiation in the physical world. However, deriving 3D representations is computationally difficult because 3D information must be recovered and integrated (Bülthoff & Edelman, 1993).

How are Features Related to One Another?

Features are the building blocks of object representations. But by themselves, they are not sufficient to characterize, either quantitatively or qualitatively, the appearance of objects. A face, for example, is not a random arrangement of eyes, ears, nose, and mouth, but rather a particular set of features in a particular spatial arrangement. Object representations must therefore express the spatial relations between features. One aspect of how this is accomplished is whether the spatial relations between features are represented at a single level, in which all features share the same status, or whether there is a hierarchy of relations. For instance, Marr and Nishihara (1978) hypothesized that a small number of parts at the top of a hierarchy are progressively decomposed into constituent parts and their spatial relationships at finer and finer scales – for example, an arm can be decomposed into an upper arm, forearm, and hand. The hand, in turn, can be further decomposed at an even finer scale into a palm and fingers. Local image features can similarly be structured: elaborate features can be decomposed into simpler ones (Riesenhuber & Poggio, 1999). Thus, entry-level categories might be captured by a higher level of the hierarchy (in that the common features that define an entry-level category, for example, “bear,” are typically more global) and subordinate-level categories might only be captured by finer levels (e.g., the subtle features that distinguish a grizzly bear from a brown bear). Alternatively, object representations might only encode the coarsest level of structure: two to three parts (and their relations) at most with no hierarchy specifying structure at finer scales (Biederman, 1987; Hummel & Biederman, 1992).

Since visual input is inherently spatial, early and intermediate visual representations necessarily encode the quantitative positions of features – in essence, representations within the visual system are inherently *spatial* early on (Marr, 1982). At issue, however, is the degree to which metric information is preserved in higher-level, long-term object representations. Building on the examples cited in the previous section, there is dichotomy between a local, quantitative approach and a global, qualitative approach. At one extreme, the metric relations between features present in the image are kept more or less intact in higher-level object representations. The resulting *template* would be highly sensitive to changes in the image, such that any variation in the spatial relations between features, no matter how slight, would require a new representation (although this variation may be compensated for using a variety of strategies – see the discussion on normalization mechanisms below). Typically, however, even if one were to posit local, quantitative features, the relations between them are assumed to be somewhat more flexible. For example, many local feature theories posit that the relative positions of features are probabilistic (Edelman 1995; Tarr & Bülthoff, 1998; Riesenhuber & Poggio, 1999). The resulting object representation would still be sensitive to metric variations between a known version of an object and a new image of that object, but would not fail to find correspondences between like features in slightly different locations – a *deformable template*, so to speak (Ullman & Basri, 1991). Models that assume local, quantitative measures as the molar features along with relatively precise localization of these features in the image have been dubbed *image-based models* (although this term still encompasses a wide variety of approaches).

Global, qualitative models tend to assume a much coarser coding of the spatial relations between features. Biederman (1987; Hummel & Biederman, 1992) argues that spatial relations are encoded in a qualitative manner that discards metric relations between object features yet preserves their critical structural relations. On this view, for example, the representation would code that one part is above another part, but not how far above or how directly above (a "top-of" relation). The resulting concatenation of features (in Biederman's model, 3D parts) and qualitative structural relations is often referred to as a *structural description*.

One other possibility should be noted. *All* spatial relations between features might be discarded and only the features themselves represented: so a face, for instance, might just be a jumble of features! Such a scheme can be conceptualized as an array of non-localized *feature detectors* uniformly distributed across the retinal array (dubbed "Pandemonium" by Selfridge, 1959). The resulting representation might be more stable, but only so long as the same features or a subset of these features are present somewhere in the image and their presence uniquely specifies the appropriate object. Although there are obvious problems with this approach, it may have more merit than it is often given credit for, particularly if one assumes an extremely rich feature vocabulary and a large number of features per object (see Tarr, 1999).

Frames of Reference

As pointed out in the previous section, most theorists agree that intermediate stages of visual processing preserve at least the rough geometry of retinal inputs. Thus, there is implicitly, from the perspective of the observer, a specification of the spatial relations between features for intermediate-level image representations. Ultimately, however, the spatial relations between features are typically assumed to be explicit in high-level object representations. This explicit coding is generally thought to place features in locations specified relative to one or more anchor points or frames of reference (Marr, 1982).

The most common distinction between reference frames is whether they are *viewpoint-independent* or *viewpoint-dependent*. Embedded in these two types of approaches are several different kinds of frames, each relying on different anchor points. For example, viewpoint-independent models encompass both object-centered and viewpoint-invariant representations. Consider what happens if the features of an object are defined relative to the object itself: although changes in viewpoint alter the appearance of the object, they do not change the position of a given feature relative to other features in the object (so long as the object remains rigid). Thus, the representation of the object does not change with many changes in the image. The best known instantiation of an object-centered theory was proposed by Marr and Nishihara (1978). They suggest that an object's features are specified relative to its axis of elongation, although other axes, such as the axis of symmetry, are also possible (McMullen & Farah, 1991). So long as an observer can recover the elongation axis for a

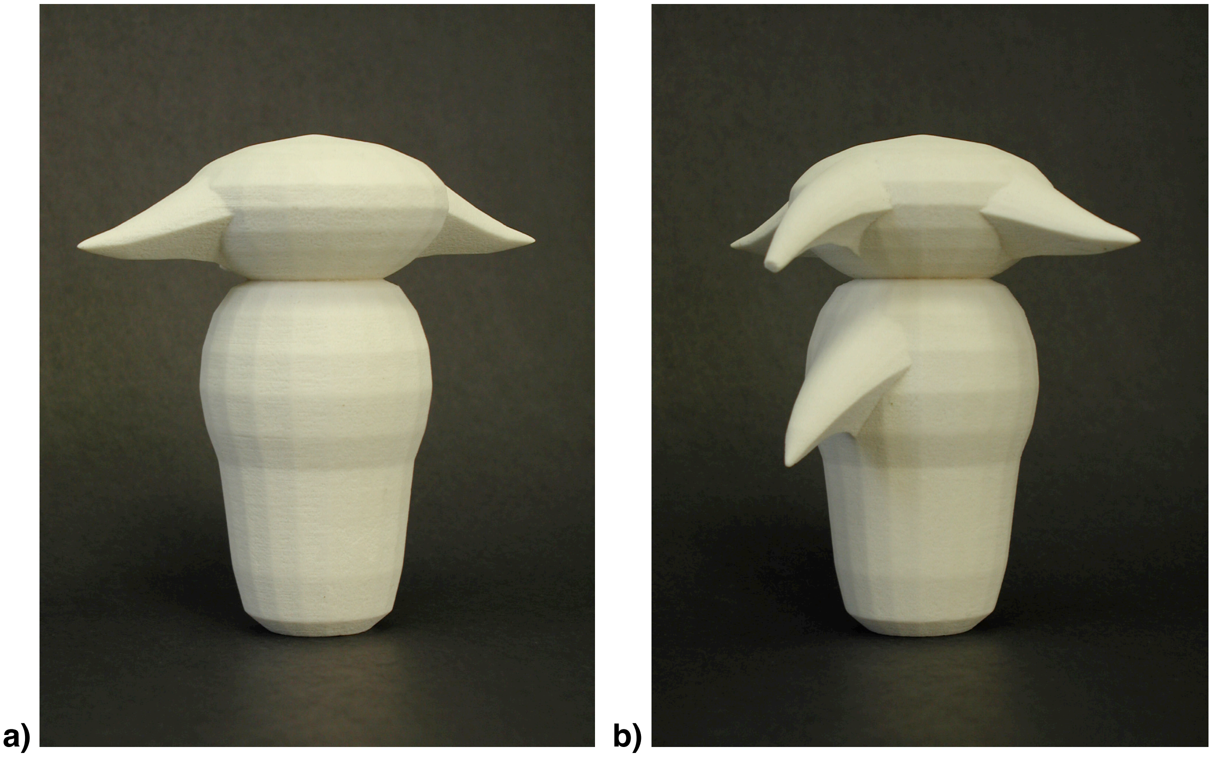

Figure 1. Different viewpoints of objects often reveal (or occlude) different features. Thus, it seems unlikely that a complete 3D object representation could be derived from any single viewpoint.

given object, a canonical description of that object can be constructed for any image of it. Ideally, only a single viewpoint-independent representation of each object would be encoded in visual memory in order for that object to be recognized from all viewpoints. Although this approach has some advantages, in practice it has proven rather difficult to develop methods for the reliable derivation of canonical axes of elongation for most objects without recourse to at least some aspects of the object's identity. Moreover, it seems unlikely that a single representation of an object could suffice when there exist many viewpoints from which some significant features of the object are completely occluded (Figure 1).

Biederman's (1987) theory of recognition attempts to address some of these shortcomings. Like Marr and Nishihara, he assumes a structural description comprised of 3D parts, but his theory posits viewpoint-invariant (not object-centered) object representations. What precisely is the distinction? Rather than attempt to encode the position of features in any specific coordinate system, Biederman side-stepped the problem by proposing that particular collections of viewpoint-invariant features (sometimes referred to as "non-accidental" properties, i.e., local configurations of edges that are so unlikely to have occurred by accident that they must reflect meaningful 3D structure) map onto a given 3D volume. For example, if several Y-vertices and arrow-vertices appear in an image along with three parallel lines, they might specify a brick. In Biederman's model the brick 3D primitive would entirely replace the image features that specified the part. Because the parts themselves and their relations are represented only at a non-metric, qualitative level, for example, brick on-top-of cylinder, the representation does not use a strictly object-centered frame of reference (although the spatial relations are still object relative). The trick here is that the particular features specifying the 3D parts are invariants with respect to viewpoint (and possibly illumination). Thus, many changes in viewpoint would not change which particular 3D primitives were recovered. Biederman, however, acknowledges that it is impossible to recover parts that are not visible from a given viewpoint. Thus, he allows for multiple representations for each object – each depicting different collections of visible parts for the same object under different viewing conditions (Biederman & Gehardstein, 1993).

In contrast to viewpoint-invariant models, viewpoint-dependent models inherently encompass retinotopic, viewer-centered (egocentric), and environment-centered (spatiotopic or allocentric) frames, anchored to the retinal image, the observer, or the environment, respectively. That is, objects are represented from a particular viewpoint, which entails multiple representations. Put another way, object representations that use a viewer-centered reference frame are tied more or less directly to the object as it appears to the viewer or, in the case of allocentric frames, relative to the environment. As such, they are typically assumed to be less abstract and more visually-rich than viewpoint-independent representations (although this is simply a particular choice of the field; viewpoint-dependence/independence and the richness of the representation are technically separable issues). It is often thought that viewpoint-dependent representations may be more readily computed from retinal images as compared to viewpoint-independent representations. However, there is an associated cost in that viewpoint-dependent representations are less stable across changes in viewpoint in that they necessarily encode distinct viewpoints of the same object as distinct object representations. Thus, theories adopting this approach require a large number of viewpoints for each known object. Although this approach places higher demands on memory capacity, it does potentially reduce the degree of computation necessary for deriving high-level object representations for recognition.

Normalization Procedures

Regardless of the molar features of the representation – local features, 3D parts, or something in between – if some degree of viewpoint dependency is assumed, then the representation for a single object or class will consist of a set of distinct feature collections, each depicting the appearance of the object from a different vantage point. This leads to a significant theoretical problem: different viewpoints of the same object must somehow be linked to form a coherent representation of the 3D object. One solution might be to find a rough correspondence between the features present in different viewpoints. For example, the head of the bear is visible from both the front and the side, so this might be a clue that the two images arose from the same object. Unfortunately, simple geometric correspondence seems unlikely to solve this problem –

if such correspondences were available (i.e., if it were possible to map one viewpoint of an object into another viewpoint of that same object), then recognition might proceed without the need to learn the new viewpoint in the first place! So it would seem that viewpoints are either distinct or they aren't (Jolicoeur, 1985).

The conundrum of how an observer might recognize a novel viewpoint of a familiar object was addressed by Tarr and Pinker (1989). They built on the finding that human perceivers have available a "mental rotation" process (Shepard & Metzler, 1971) by which they can transform a mental image of a 3D object from one viewpoint to another. Shepard and others had reasoned that although the mental rotation process was useful for mental problem solving, it was not appropriate for object recognition. The argument was that in order to know the direction and "target" of a given rotation, an observer must already know the identity of object in question; therefore executing a mental rotation would be moot. Put another way, how would the recognition system determine the correct direction and magnitude of the transformation prior to recognition? Ullman (1989) pointed out that an "alignment" between the input shape and known object representations could be carried out on the basis of partial information. That is, a small portion of the input could be used to compute both the most likely matches for the current input, as well as the transformation necessary to align this input with its putative matches. In practice, this means that a subset of local features in the input image are compared, in parallel, to features encoded in stored object representations. Each comparison returns a goodness-of-fit measure and the transformation necessary to align the image with the particular candidate representation (Ullman, 1989). The transformation actually executed is based on the best match among these. Thus, observers could learn one or more viewpoints of an object and then use these known viewpoints plus normalization procedures to map from unfamiliar to familiar viewpoints during recognition.

Jolicoeur (1985) provided some of the first data suggesting that such a process exists by demonstrating that the time it takes to name a familiar object increases as that object is rotated further and further away from its familiar, upright orientation. However, this result was problematic in that upright viewpoints of mono-oriented objects are “canonical” in that they are the most frequently seen and most preferred views (Palmer, Rosch, & Chase, 1981). Thus, the pattern of naming times obtained by Jolicoeur might speak more to the “goodness” of different object views than to the mechanisms used in recognition.

Tarr and Pinker’s (1989) innovation was to use novel 2D objects shown to subjects in multiple viewpoints. Subjects learned the names of four of the objects and then practiced naming these objects plus three distractors (for which the correct response was “none-

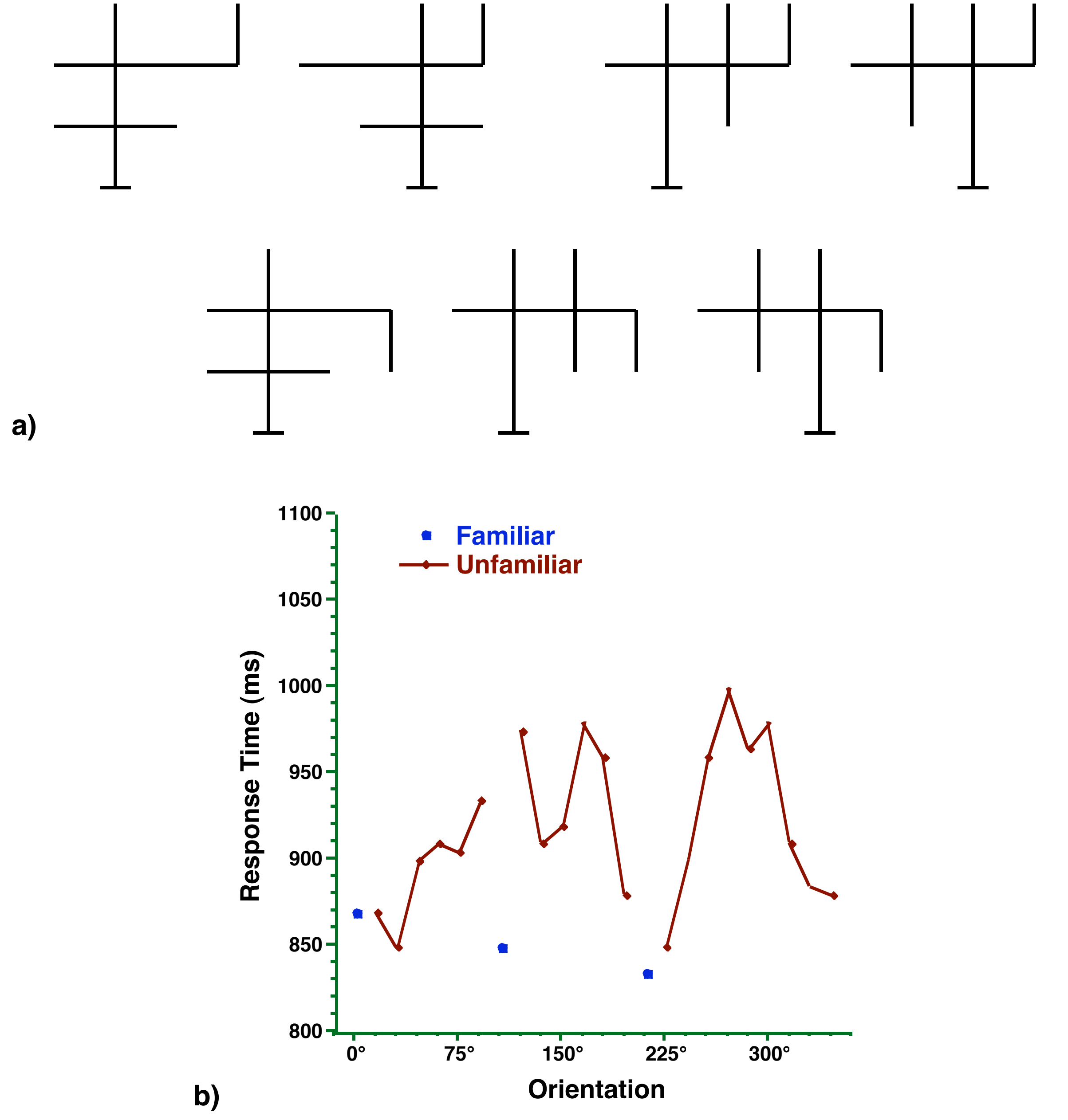

Figure 2. a) The novel 2D objects used as stimuli in Tarr and Pinker (1989). **b)** Average response times for subjects to name these objects in trained, familiar orientations and unfamiliar orientations. Naming times increased systematically with distance from the nearest familiar orientation. Adapted from Tarr and Pinker (1989) Figures 1 (p. 243) and 6 (p. 255).

of-the-above") in several orientations generated by rotations in the picture-plane (Figure 2a). Tarr and Pinker's subjects rapidly became equally fast at naming these objects from all trained orientations. Subjects' naming times changed when new picture-plane orientations were then introduced: they remained equally fast at familiar orientations, but naming times were progressively slower as the objects were rotated further and further away from a familiar orientation (Figure 2b). Thus, subjects were learning to encode and use those orientations that were seen most frequently (and not merely geometrically "good" views). This suggests that observers were able to invoke normalization procedures to map unfamiliar orientations of known objects to familiar orientations of those same objects. Tarr and Pinker (1989) hypothesized that these normalization procedures were based on the same mental rotation process discovered by Shepard and Metzler (1971). Tarr (1995) reported corroborating evidence using novel 3D versions of Tarr and Pinker's 2D objects rotated in depth.

Some researchers have pointed out that mental rotation is an ill-defined process. What does it really mean to "rotate" a mental image? Several researchers have offered computational mechanisms for implementing normalization procedures with behavioral signatures similar to that predicted for mental rotation. These include linear combinations of views (Ullman & Basri, 1991), view interpolation (Poggio & Edelman, 1990), and statistical evidence accumulation (Perrett, Oram, & Ashbridge, 1998). Some of these normalization mechanisms make further predictions regarding recognition behavior for new viewpoints of known objects. For example, Bülthoff and Edelman (1992) obtained some evidence consistent with the view-interpolation models of normalization and Perrett, Oram, and Ashbridge (1998) found that responses of populations of neurons in monkey visual cortex were consistent with the evidence accumulation account of normalization.

Note that viewpoint-invariant theories of recognition do not require normalization as a way for recognizing unfamiliar viewpoints of familiar objects. In particular, both Marr and Nishihara (1978) and Biederman (1987) assume that viewpoint-invariant recovery mechanisms are sufficient to recognize an object from any viewpoint, known or unknown, or that viewpoint-invariant mechanisms are viewpoint limited, but span extremely wide viewpoint regions, effectively making recognition behavior viewpoint-independent (Biederman & Gerhardstein, 1993; see Tarr & Bülthoff, 1995, for a critique of these claims and Biederman & Gerhardstein, 1995 for their reply).

Although these normalization mechanisms show how disparate viewpoints of objects may be mapped onto one another, they still leave open one of our original questions: When two viewpoints of a single object are so distinct as to require separate representations, how does the recognition system ever map these onto the same object name or category (or are observers destined to never know that it was the same thing coming as going)? One intriguing possibility is that statistical mechanisms similar to those that seem to function in learning distinct viewpoints are used to connect disparate visual information over time. Consider that when viewing an object the single image one is most likely to see next is another viewpoint of that same object (Tarr & Bülthoff, 1998). Thus, a mechanism that associates what one sees at one instant with what one sees at the next instant would produce the necessary relationships. Recent neural (Miyashita, 1988) and behavioral (Wallis, 1996) evidence indicates that primate visual memory does learn to associate distinct images if they co-occur over time. Thus, whether views are defined by 3D parts or by local features, temporal associations may provide the "glue" for building coherent mental representations of objects (Tarr & Bülthoff, 1998).

Theories of Object Recognition

Structural-Description Models

We have now set the stage for an enumeration of two primary approaches to visual object recognition: *structural-description* and *image-based* theories. For historical reasons we will begin by reviewing the structural-description approach. One consequence of the computing revolution of the late 1960's and early 1970's was the development of sophisticated tools for generating realistic computer graphics (Foley & Van Dam, 1982). To generate images of synthetic 3D objects, graphic programmers employed volumetric primitives – that is, 3D volumes that depicted the approximate shape of a part of a real 3D object. Thus, a teakettle might be synthesized by a sphere flattened on the top and bottom, a very flat cylinder for the lid, and two bent thin cylinders for the handle and spout. Based on such techniques, some researchers suggested that similar representations might be used by both biological and machine vision systems for object recognition (Binford, 1971). Specifically, objects would be learned by decomposing them into a collection of 3D parts and then remembering that part configuration. Recognition would proceed by recovering 3D parts from an image and then matching this new configuration to those stored in object memory. One appealing element of this approach was the representational power of the primitives – called “generalized cones” (or “generalized cylinders”) by Binford. A generalized cone represents 3D shape as three sets of parameters: (1) an arbitrarily shaped cross-section that (2) can scale arbitrarily as (3) it is swept across an arbitrarily shaped axis. These three parameter sets are typically defined by algebraic functions that together capture the shape of the object part.

Marr and Nishihara built on this concept in their seminal 1978 theory of recognition. In many ways they proposed the first viable account of human object recognition, presenting a model that seemed to address the factors of invariance, stability, and level

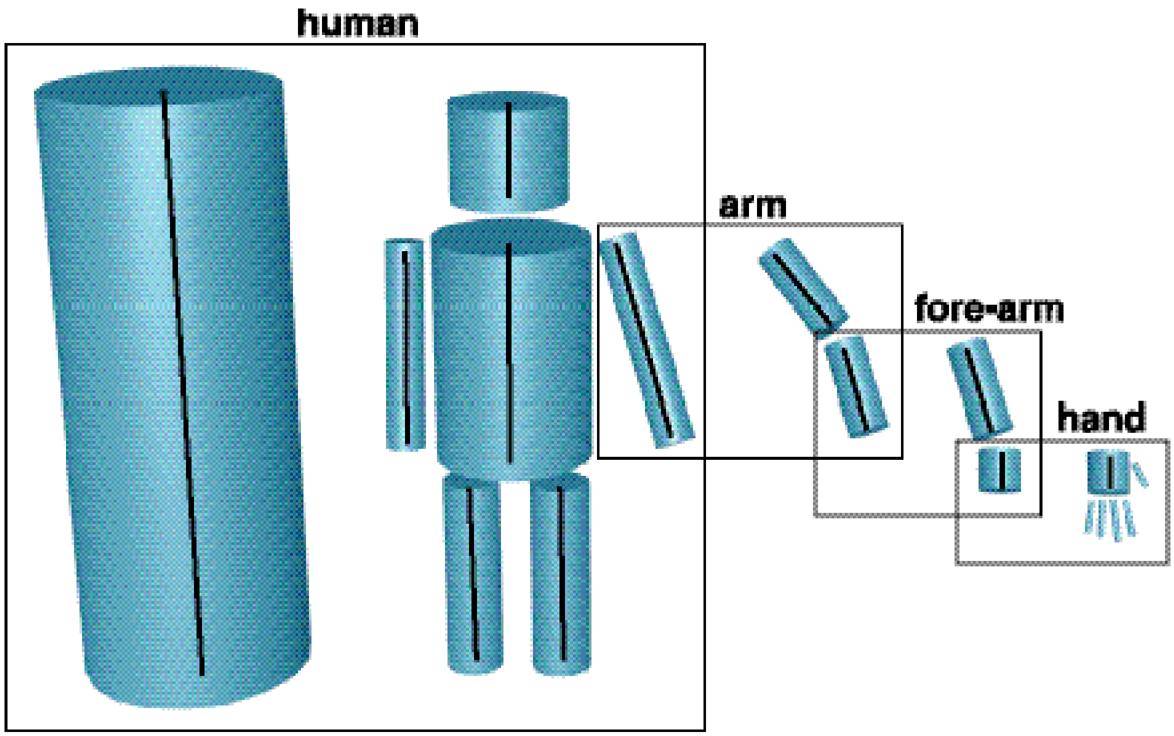

Figure 3. Schematic diagram of the multi-scale 3D representation of a human figure using object-centered generalized cones. Although cylinders are used for illustrative

of access. As mentioned previously, they placed a significant computational burden on the reconstruction of the 3D scene. In particular, in their model a necessary step in the recognition process is recovering 3D parts from the input image. Recognizing the power of generalized cones, Marr and Nishihara suggested that observers use information about an object's bounding contour to locate its major axis of elongation. This axis can then be used as the sweeping axis for the creation of a generalized cone structural description relating individual 3D parts to one another at multiple scales (Figure 3). Invariance was accomplished by virtue of the object-centered coordinate system in which these 3D parts were parameterized. Thus, regardless of viewpoint and across most viewing conditions, the same viewpoint-invariant, 3D structural description would be recovered by identifying the appropriate image features, (e.g., the bounding contour and major axes of the object) recovering a canonical set of 3D parts, and matching the resultant 3D representation to like representations in visual memory.

Biederman's (1987; Hummel & Biederman, 1992) model – "Recognition-By-Components" (RBC) – is quite similar to Marr and Nishihara's theory. However, two innovations made the RBC model more plausible in the eyes of many researchers. First, RBC assumes a restricted set of volumetric primitives, dubbed "geons" (Figure 4). Second, RBC assumes that geons are recovered on the basis of highly stable non-accidental image properties, that is, shape configurations that are unlikely to have occurred purely by chance (Lowe, 1987). One example of a non-accidental property is three edges meeting at a single point (an arrow or Y junction) – it is far more likely that this image configuration is the result of an inside or outside corner of a rectangular 3D object than the chance meeting of some random disconnected edges. Biederman considered the set of 36 or so 3D volumes specified by the combinations of non-accidental properties in the image: the presence of particular edge junctions or vertices, the shape of the major axes, symmetry of the cross section around these axes, and the scaling of the cross section. For example, a cylinder is specified as a curved cross section (i.e., a circle) with rotational and reflection symmetry, constant size, and a straight axis. An important point to keep in mind is that these attributes are defined qualitatively, for instance, a cross section is either straight or curved, there is no in-between or degree of curvedness. Although the reader might be skeptical about how one might represent the huge range of objects that exist using this limited toolkit, consider tinker-toys; there are only a few different kinds of building blocks and connectors in the can, yet one can construct models that approximate almost any type of object.

Not only do the possible combinations of non-accidental properties enumerate a restricted set of 3D volumes – geons – but they also allow a method for recovering these particular volumes. The fact that geons are generated by contrasting qualitative differences in non-accidental properties is crucial to this process. Consider a brick and cylinder. Most non-accidental projections of a brick have three parallel edges and three outer arrow vertices (points of co-termination). In contrast, most non-accidental projections of a cylinder have only two parallel edges and two tangent Y-vertices. These and other contrasts provide the leverage for inferring the presence of specific geons in the image. Hypothetically an observer only needs to examine the image of a 3D object for the presence of the critical non-accidental properties and then replace these properties with the appropriate, qualitatively-specified geon. The complete configuration is referred to as a Geon-Structural-Description. Because the RBC approach restricts the set of possible 3D volumes and, consequently, can rely on a more plausible method of

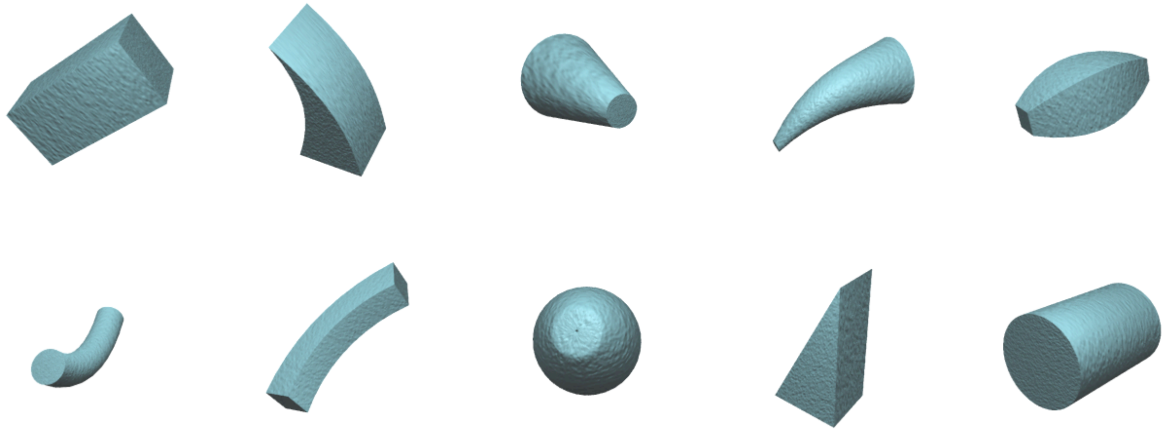

Figure 4. Ten of the 30 or so geons – 3D volumetric primitives – that Biederman's (1987) RBC model posits as the building blocks of object representations.

3D part recovery, RBC is computationally more tractable than Marr and Nishihara's theory.

RBC deviates from Marr and Nishihara's model in three other, related respects: RBC assumes only one level of representation, that is, 2 or 3 geons per object; RBC only attempts to account for object recognition at the entry level; and, as reviewed earlier, RBC posits qualitative spatial relations between parts.

Evaluating Structural-Description Models

What are the strengths and weaknesses of Marr and Nishihara's approach to object recognition? Perhaps the two most appealing aspects of the model are: 1) The invariance across viewpoint and other image transformations obtained by adopting a viewpoint-independent object-centered frame of reference for the description of parts; and, 2) the balance between stability and sensitivity achieved by representing objects at multiple scales in a hierarchical fashion. However, the model is in some ways almost too powerful. That is, generalized cones are represented by an arbitrary axis and cross section, therefore the mathematical description of these elements may be quite complex. Moreover, even at coarser scales, it is not clear that different instances of an object class will give rise to the same generalized cone descriptions and, thus, class generalization may be difficult. Another serious issue is that the method for recovering generalized cones is not well-specified and has never been implemented successfully. Finally, empirical studies of human object recognition have not obtained much evidence in support of the sort of in invariances predicted by Marr and Nishihara. All in all, while promising in some respects, these and other shortcomings of Marr and Nishihara's model rendered it a less than plausible account of human object recognition. Indeed, many of the extensions to Marr and Nishihara's model proposed in RBC seem explicitly designed to address these shortcomings.

What are the implications of these additional assumptions in RBC? On the positive side, RBC provides a more specific account of how 3D primitives might be derived from 2D images. Moreover, by severely restricting the vocabulary of 3D primitives to only 36 or so geons, RBC removes the potential computational complexity of describing complicated 3D parts as a series of mathematical functions. Further efficiency is derived from the fact that RBC limits the number of parts in each object description. Finally, the fact that geons and the relations between geons are qualitatively defined provides invariance across a wide range of viewpoints, as well as position, size, and other variations in viewing conditions. Moreover, different instances of an object category often contain similar parts, thus the use of geons facilitates a many-to-one mapping between individual exemplars and their visual object class.

At the same time, some researchers have argued that, much as with Marr and Nishihara's model, RBC specifies a theory that does not account for many aspects of human visual recognition behavior. First, consider the recovery process. RBC relies heavily on particular configurations of edges. Yet there is little evidence to suggest that early and mid-level vision provide an edge map that looks anything like a clean line drawing (Sanocki, Bowyer, Heath, & Sarkar, 1998). For instance, depending on the direction of the lighting, shadows may produce spurious edges on many of an object's surfaces. Thus, although using edge-based non-accidental properties may seem appealing at first blush, the reliability of recovery mechanisms based on such features remains suspect. Moreover, by insisting on a singular approach that is edge-based, RBC relegates surface characteristics, including color, texture, and shading, to a secondary role in recognition (Biederman & Ju, 1988). Yet there is a growing body of data indicating that surface properties are critical to the recognition process and more or less integrated into object representations (see Price & Humphreys, 1989; Tarr, Kersten, & Bülthoff, 1998; Tanaka, Weiskopf, & Williams, in press).

Second, consider the nature of the representation. RBC assumes a single level of two to three qualitatively-defined geons. This is a major reason why the theory only attempts to explain entry-level recognition. Such an impoverished object representation would be useless for the recognition of objects at the subordinate level or of specific individuals, could easily confuse objects that are actually distinct at the entry level (but visually-similar), or distinguish between objects that are actually members of the same entry-level class (Tarr & Bülthoff, 1995).

Third, as discussed earlier, RBC (as well as Marr & Nishihara's theory) predicts that over a wide range of changes in viewpoint, recognition performance should be viewpoint invariant. Although there are some limited circumstances in which observers are able to invariantly recognize known objects when shown in never-before-seen viewpoints, by and large the common finding has been some cost in both accuracy and response time when objects must be recognized across changes in viewpoint (Hayward & Tarr, 1997; Tarr, Bülthoff, Zabinski, & Blanz, 1997; Tarr, Williams, Hayward, & Gauthier, 1998).

In sum, there are limitations to RBC and a need for alternative mechanisms. Indeed, Hummel and Biederman (1992) and Biederman and Gerhardstein (1995) acknowledge that RBC captures some, but not all, recognition phenomena. The question is whether RBC forms the bulk of the explanation or, as claimed by Tarr and Bülthoff (1995), constitutes a restricted model that is inconsistent with much of the psychophysical and neural data. Thus, although some aspects of the structural-description approach are appealing and have been incorporated into recent models of recognition, it is still unclear whether the overall concept of a single-level, 3D qualitative part-based representation can account for human object recognition abilities.

Image-Based Models

The most common alternative to a structural-description account is an image-based representation. Although the term “image-based” (or “view-based”; within the computer vision community such models are also sometimes called “appearance-based”) has been criticized as too vague, there are assumptions that constrain the theories. Perhaps the most critical element of any image-based model is that the features of the representation preserve aspects of an object as it *originally appeared* in the image. Note that this statement does not restrict the representation only to shape features, but allows for measures of almost any object property, including color, texture, shading, local depth, and spatial frequency, as well as shape (e.g., Edelman, 1993). The inclusion of non-shape properties is a significant difference between the image-based and structural-description approaches. A second important difference is how image-based theories encode 3D structure: Instead of a single 3D object model or a set of 3D models, image-based theories represent 3D objects as a collection of *views*, each view depicting the appearance of the object under specific viewing conditions (Tarr, 1995). This “multiple-views" representation supports the invariant recognition of 3D objects in much the same manner as structural descriptions.

In light of the framework outlined earlier, the majority of image-based models posit local features that are generally thought of as visual information processed over restricted regions – the output of the filtering that is implemented in early and mid-level vision. Often this information is characterized as part of the *surface* representation, and, indeed, there are many reasons why surfaces are attractive as the molar unit of high-level visual representations. At the same time, how individual surfaces are spatially related to one another is an issue of some debate. At one extreme, some researchers have argued that the overall representation simply preserves the quantitative 2D spatial relations visible in the image (Poggio & Edelman, 1990), for example, a rigid template. At another extreme, others have argued for a completely unordered collection of features that preserve nothing about their spatial relationship (Mel, 1997). Both approaches have some merit, but break under fairly obvious conditions. Thus, the trend has been to implement hybrid models in which the features are image-based, but are related to one another in a hierarchy that captures multiple levels of object structure (Hummel & Stankiewicz, 1996; Riesenhuber & Poggio, 1999; Lowe, 2000; Ullman & Sali, 2000). In some sense, such models are structural descriptions in that they relate the positions of object features to one another in a multi-scale hierarchy, but they are different from either Marr and Nishihara's or Biederman's models in that they assume the features are viewpoint-dependent local features rather viewpoint-independent 3D parts.

Consider two other aspects that are used to characterize object representations. In terms of dimensionality, image-based models rarely assume a unitary 3D representation. Rather, they posit either features that remain 2D or, through the recovery of local depth, are 2.5D. Thus, at most, image-based features represent local depth, that is, surface slant and orientation from the perspective of the viewer. Critically, this representation of depth does not impart viewpoint invariance (Bülthoff & Edelman, 1992). This brings us to the last characteristic of image-based models, the use of a viewpoint-dependent frame of reference – something implied by the nature of the features themselves and the way in which depth information is represented (if at all). Indeed, a major characteristic of image-based theories is that they are viewpoint dependent, that is, as the input image deviates from the originally learned viewing conditions, there are systematic costs in recognition accuracy and speed. This prediction is central to many studies that have examined the object recognition process.

Image Normalization

One other aspect of image-based models is important to elucidate. Because an object's representation is tied to specific viewing conditions, even minimal changes in viewing conditions may produce a mismatch between a viewpoint-based representation and a new viewpoint of the same object. Consequently, one must either posit a large number of views to represent a single object – with the potential to exceed the limits of human memory capacity – or a mechanism to *normalize* the processed input image to object viewpoints encoded in visual memory. As discussed earlier, there are several different proposals for how normalization might be accomplished. These can be divided roughly into four classes: mental transformation models (Shepard & Metzler, 1971; Tarr & Pinker, 1989); interpolation models (Poggio & Edelman, 1990; Bülthoff & Edelman, 1992; Ullman & Basri, 1991); alignment models (Ullman, 1989); and evidence accumulation models (Perrett, Oram, & Ashbridge, 1998).

What do these different approaches to normalization have in common and how do they differ from one another? First and foremost, they almost all predict that the normalization process is capacity limited so that the magnitude of the normalization impacts recognition performance. That is, the larger the difference between the input image and stored representations, the larger the cost in recognition time and accuracy. At the same time, different view-based models make different predictions regarding how these costs will manifest themselves, particularly across changes in viewpoint.

Mental transformation models (Tarr & Pinker, 1989; Tarr, 1995) predict that the costs will vary as a direct function of the angular difference between familiar and unfamiliar viewpoints of an object. Put another way, recognition performance will be straightforwardly determined by how far away a new viewpoint is from a known viewpoint. This prediction is based on the idea that mental transformations are analogs to physical transformations, following a continuous path that traverses the same space that the object would if it were actually being rotated. Thus, larger transformations take more time and incur more errors as compared to smaller transformations and the magnitude of both response times and error rates is proportional to the magnitude of the transformation (Shepard & Cooper, 1982).

View-interpolation models (Poggio & Edelman, 1990; Bülthoff & Edelman, 1992; see also Ullman & Basri, 1991) predict that costs will vary depending on how the view space is spanned by familiar viewpoints. That is, a better estimate, and consequently smaller performance costs, of how the image of an object is likely to change with rotation in depth can be made if the rotation is bounded by two or more known viewpoints. On the other hand, if a rotation in depth is not bounded by known viewpoints, approximations of the appearance of an object from a new view will be poorer and the costs associated with normalization will be larger. Remember interpolating between points on a graph in high-school math: you could do a better job predicting the shape of the curve if you interpolated between two points and the points were closer together; if there were few points and they were far apart, your ability to predict was diminished. The same applies to view interpolation – simply think of each known viewpoint as a point on a 3D graph and a new viewpoint as a point between these actually plotted points. Unfamiliar viewpoints between familiar viewpoints are likely to fall on the “line” connecting the two known points and hence are better recognized than unfamiliar viewpoints that do not fall on this “line”.

Alignment models (Ullman, 1989) actually do not predict performance costs *per se*, although they do posit that 3D objects are represented as multiple viewpoint-specific feature sets. The reason is that they assume that once the alignment transformation has been determined between an input image and the correct view in memory, the alignment may be executed in a single step. There are several reasons why, in practice, this single-shot approach is unlikely to work. First, the appropriate alignment transformation is determined by comparing a small subset of the input with object representations in visual memory (the same sort of pre-processing used to determine the rotation in mental transformation models). Although this process might appear to be time consuming, there are methods for executing the comparison in parallel, thereby simultaneously comparing a given feature subset from the input with all candidate object representations. However, the reliability of this process will decrease as the alignment features increase in their dissimilarity from those in visual memory. Thus, larger viewpoint differences will, in all probability, lead to less reliable recognition performance. Second, the idea that an alignment can be performed by a single-step transformation, regardless of magnitude, works only if object features are represented in a form that allows rigid, 3D rotations to be applied to all features simultaneously (e.g., a 3D matrix of shape coordinates). If the features of the representation are not in this format – for example, their 3D positions are unknown – then other, most likely incremental, processes must be used to align input images with views in object memory. Indeed, most researchers now consider alignment models to be mathematical approximations of the normalization process, rather than actual models of how normalization is implemented in biological systems (Ullman, 1996).

Finally, evidence accumulation models (Perrett, Oram, & Ashbridge, 1998) attempt to account for normalization-like behaviors without actually positing any transformation of the input image. Recall that image-based representations are often characterized by a large number of local viewpoint-specific measures of image properties: color, shape, texture, etc. What happens to these measures if viewing conditions change? Here intuition may fail us: it might seem that even a small object rotation or shift in lighting direction would change every local feature in a manner that would lead to an entirely different image description. In truth, this is not the case. The appearance of some features will not change at all or only marginally, others will change more dramatically, but generally in a systematic fashion; and only a fraction of the visible features will appear entirely different or disappear. Thus, for a large collection of local image features, the overall difference between the original image of an object and a new image of the same object under different viewing conditions will be related to how much viewing conditions have changed. That is, given a large number of features in an object representation, the cumulative response of these features is related almost monotonically to the degree of rotation or the magnitude of lighting change from the known to unknown image. The implication is that objects may be represented as image-specific views, but that rather than using normalization mechanisms, image similarity across large numbers of features provides a means for generalizing from known to unknown viewpoints.

Why is this approach referred to as an “evidence accumulation” model? Consider that each measure of an object image can be thought of as a statistical feature detector – for a given location the better an actual feature matches the preferred feature of the detector, the stronger the response. Thus, each detector “accumulates evidence” about the presence of particular features, while the overall response of the collection provides an estimate of the likelihood of a particular image of a particular object given the current image. What is perhaps most interesting about such models is that they make predictions quite similar to those of mental transformation and interpolation models. That is, as the magnitude of the change increases so will the magnitude of response times and error rates. Consequently, a recognition system that relies on evidence accumulation to generalize from known to unknown viewpoints will produce a response pattern that *appears* as if the image is being normalized.

Evaluating Image-Based Models

How do the assumptions of image-based models stand up to the data? Let’s consider the most critical characteristic of the approach, that object representations are view-based. The implication is that observers should better remember what objects looked liked under familiar viewing conditions, that is, when seeing viewpoints that they have seen before. A corollary of this prediction is that recognition performance should decrease as the object image is transformed further and further from a known viewpoint. The precise pattern of how performance will change in relation to changes in the image will depend on the particular version of normalization that one adopts, but the basic point of diminished performance with larger changes remains constant in most image-based models.

A considerable body of behavioral and neurophysiological results is consistent with this basic prediction of image-based models. As mentioned earlier, several studies observed that it took longer for subjects to name familiar common objects as those objects were rotated in the picture-plane away from their upright, canonical orientation

(Jolicoeur, 1985) or for subjects to name novel objects in new, unfamiliar viewpoints as those objects were rotated away from trained viewpoints (Tarr & Pinker, 1989, Tarr, 1995). These and other results provide evidence for viewpoint-specific image-based object representations that are generalized to new viewing conditions using a mental transformation process.

An analogous result was obtained by Bülthoff and Edelman (1992; Edelman & Bülthoff, 1992). They found that recognition performance diminished as novel viewpoints were located further and further away from familiar viewpoints, but they also observed differences in the magnitude of the cost depending on whether new viewpoints were located between or beyond familiar viewpoints (a pattern also obtained in Tarr, 1995). Thus, not only did Bülthoff and Edelman provide further evidence for object representations based on multiple image-based views, but also their data suggest that the normalization mechanism used for the recognition of novel viewpoints is view interpolation, not mental transformation. Over the past decade researchers have added to these results, generating a sizeable collection of data that clearly support image-based models (Humphrey & Khan, 1992; Lawson & Humphreys, 1996; Srinivas, 1995; Tarr et al., 1998; for a review see, Tarr & Bülthoff, 1998).

Similar conclusions can be made based on results from single-unit recording studies in the inferior temporal cortex (IT) of monkeys. For example, Logothetis, Pauls, and Poggio (1995) found that if monkeys were trained to recognize novel 3D objects (the same as those used in Bülthoff and Edelman, 1992) from specific viewpoints, neurons in their inferior temporal cortex became “view-tuned.” That is, individual neurons were found to be selective for specific objects, but only for familiar viewpoints that the monkey had actually seen (Figure 5). At the same time, the monkey’s recognition performance was invariant across these highly-familiar viewpoints – similar to the finding by Tarr and Pinker (1989) that following familiarization with a set of viewpoints, human observers represent 3D objects as multiple image-specific views at those viewpoints. Although it seems unlikely that objects or views of objects are represented by single neurons (Booth & Rolls, 1998), the fact that neurons selective for particular objects – indicating that such neurons at least participate in the object representation – are also selective for familiar viewpoints of those objects provides evidence that the overall object representations are, themselves, image-based (a similar study by Booth & Rolls, 1998, was interpreted as evidence for viewpoint-invariant coding, but in fact they obtained approximately the same proportion of view-tuned cells as observed in Logothetis et al., 1995).

Further evidence for this claim comes from several studies that examined the viewpoint sensitivity of single IT neurons selective for human and monkey heads. That is, not only do researchers often find neurons that are highly active when presented with particular faces, but their strongest response is viewpoint dependent; typically for the frontal or profile viewpoints (Perrett, Rolls, & Caan, 1982; Perrett, Mistlin, & Chitty, 1987). Recently, Perrett, Oram, and Ashbridge (1998) have built on this finding, showing view-tuned single neurons selective for various body parts (head, hands, arms, legs,

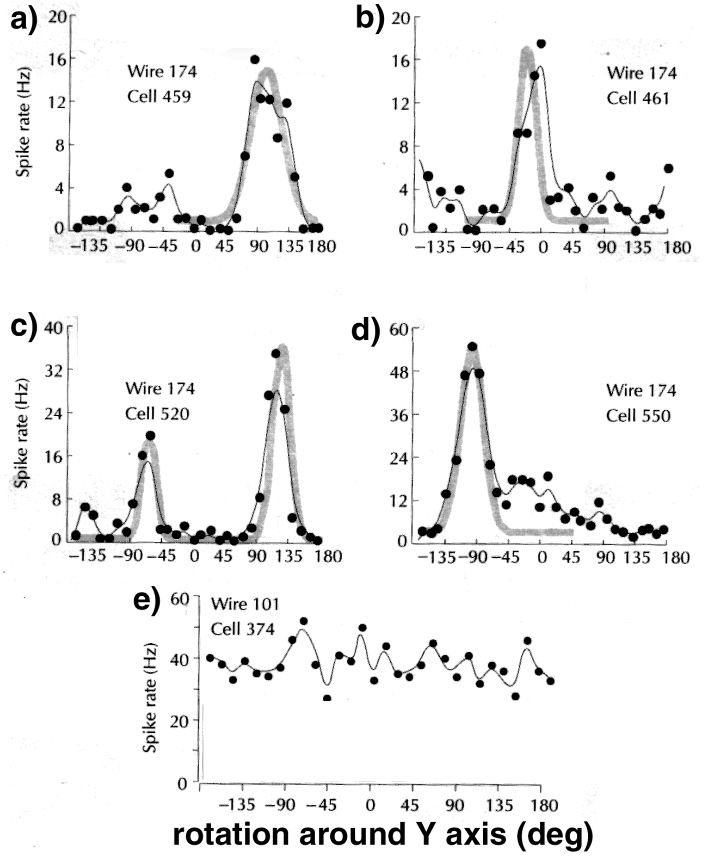

Figure 5. (a-d) Viewpoint-sensitive responses of four separate neurons in the inferior temporal cortex of a monkey trained to recognize novel 3D objects in specific viewpoints. These responses are for a single object seen from multiple viewpoints. **e)** An IT neuron that responds in a viewpoint-independent manner to a trained object. Figure adapted from Logothetis, Pauls, and Poggio (1995) Figure 5 (p. 557)

torso) in their upright orientations. Perrett, Oram, and Ashbridge found that the cumulative response of the neurons coding for different features of the figure predicted the monkey's recognition performance. That is, as the stimulus figure was rotated in the picture-plane, the response of the individual feature detectors diminished – when summed together, their responses decreased in monotonic manner with increasing misorientation – much as would be predicted by a mental transformation or interpolation model. Note, however, the response of the system was actually determined by a set of summed local responses – exactly what the evidence accumulation model predicts (see also Gauthier & Tarr, 1997b).

Overall, a great deal of behavioral and neurophysiological data may be accommodated in an image-based model. There is, however, one oft-cited criticism that must be addressed. Specifically, it has been argued that while image-based models are quite good at identifying specific instances of objects, they are poor at generalizing across instances, that is, recognizing objects at the entry level. For example, given known images of several bears, within an image-based framework how is a new bear recognized as a bear (Biederman & Gerhardstein, 1993). The answer is surprisingly simple. Most recent image-based models rely on stochastic feature detectors that respond more or less as the input feature deviates from the originally measured feature. It is this sort of response that mediates the overall responses in Perrett, Oram, and Ashbridge's (1998) evidence accumulation model and Riesenhuber and Poggio's (1999) HMAX model. Thus, such models can explain decreases in recognition performance, for instance due to changes in viewpoint, as a consequence of decreases in image similarity (typically computed over image features, not single pixels or undifferentiated images). The *same* principle may be used to account for generalization across specific instances in an image-based framework. Two instances from the same object class are likely to be similar to one another in image-feature space; therefore stochastic feature detectors representing a given object will respond more strongly to that object's cohorts – other instances of the same object class. Not only does this account provide a plausible mechanism for class generalization in image-based models, but two recent psychophysical studies provide evidence for image-based class generalization (Gauthier

& Tarr, 1997b; Tarr & Gauthier, 1998). The critical result in both studies is that observers are able to recognize novel objects that are visually similar to previously-learned novel objects and that their pattern of responses implicates viewpoint-dependent, image-based mechanisms. At the same time, there is no generalization for visually-different objects, suggesting that the observed viewpoint-dependent generalization is used for recognizing new instances of a known object class.

A second criticism of image-based models is that they are memory intensive. That is, there is a potential “combinatorial explosion” in the number of images/views that may be needed to represent completely each known object. Consider even a relatively simple object such as our bear. Depending upon what information is encoded in each familiar view of the bear, even a slight change in its appearance or one’s viewing position would produce a new collection of features in the image of the bear and, hence, lead to the representation of a new separate view. Thus, even a single, individual bear might come to be represented by 1000’s of views – a combinatorial explosion that would tax the capacity of our visual memory. One possible response to this concern is that memory in the human brain is plentiful and 1000’s of views per an object is not actually implausible. This is not an entirely satisfying response in that the same argument may not hold if the numbers increase by an order of magnitude or more. Perhaps a more plausible answer is that only a small number of views are used to describe each object and that images that deviate from known views are recognized using normalization mechanisms. If a new image is sufficiently different from known views or occurs quite frequently, it too will come to be represented. However, this explanation does not provide any hard and fast numbers on exactly how many views per an object are necessary or exactly what features are used to determine the similarity between views.

In summary, image-based models are able to posit a single mechanism that appears to account for behavior in a wide range of recognition tasks, including both entry-level recognition and the recognition of specific individuals within a class (Tarr, in press). At the same time, important aspects of image-based models are still under-specified in many respects, including the specific features of the representation, the number of views sufficient to represent an object or object class, and the exact normalization mechanisms used to match unfamiliar views to familiar ones.

Is A Single Mechanism Really Sufficient?

In the previous section we raised the possibility that a single recognition mechanism might be capable of supporting a wide array of recognition tasks, ranging from the individual level to the entry level. Although some image-based computational implementations appear to be able to handle this range of recognition tasks, there looms the larger question of whether there is any empirical evidence to support such unified approaches. Indeed, a logical argument regarding the nature of object recognition has often been used to argue for dual systems: one system that supports the recognition of objects at the entry level (discriminating between visually-*dissimilar* objects such as dogs and chairs) and that relies on separable object parts, and another system that supports the recognition of objects at the subordinate level (discriminating between visually-*similar* objects such as different faces) and relies on more “holistic” object representations (Humphreys & Riddoch, 1984; Jolicoeur, 1990; Farah, 1992). Different versions of this approach are possible, but generally they are all motivated not only by the processing demands of different levels of recognition, but more specifically by the demands of face recognition as compared to “normal” object recognition (Biederman, 1987; Farah 1992; Kanwisher, 2000).

The argument is as follows: however normal object recognition is accomplished, face recognition (and possibly other forms of subordinate-level recognition) is different in that it requires subtle discriminations between individuals that are members of a highly homogeneous object class. As such, face recognition recruits more “holistic” or “configural” information than is ordinarily called for. Therefore, according to this view, face and object recognition should be seen as separable processes. Supporting this claim are a wide range of behavioral studies that appear to demonstrate “face-specific” processing (e.g., Yin, 1969; Tanaka & Farah, 1993). Reinforcing this conclusion, there is a neuropsychological impairment from brain-injury – prosopagnosia – of object recognition that appears to be restricted exclusively to the recognition of individual faces (Farah, 1990). Finally, there are recent findings from neuroimaging (PET and fMRI) that seem to reveal a small region of inferotemporal cortex – the “fusiform face area” (FFA) – that is selectively more active for faces as compared to other objects (Kanwisher, McDermott, & Chun, 1997).

Evaluating this evidence, however, requires careful consideration of both the *default* level of categorical access for control object class (faces are recognized at the individual level by default) and the degree of perceptual expertise with the object class (all subjects are perceptual experts at face recognition). When these two factors are taken into account, a far different picture of visual object recognition emerges. Specifically, behavioral studies have found that once subjects are perceptual experts with a class of novel objects (“Greebles” – one Greeble is shown in Figure 1), subjects show the same recognition behaviors that all individuals show with faces. Consider what makes someone an expert: it is the fact that they recognize objects in the domain of expertise *automatically* at the individual level (Tanaka & Taylor, 1991). This is true for faces (I recognize my sister first as “Joanna” not as a woman or a Caucasian) and for Greebles once a subject has become a Greeble expert (i.e., they apply the individual-level names for Greebles by default). Given such perceptual expertise, the same “holistic” or “configural” effects obtained for faces and for Greebles: within a domain of expertise moving some parts of an object affects the recognition of other parts (Gauthier & Tarr, 1997a); the same is not true for objects that are not from a domain of expertise. Similarly, when prosopagnosic subjects are forced to make individual-level discriminations between Greebles, snowflakes, or even familiar, common objects, their recognition impairment for faces and objects appears equivalent across all object classes. Therefore these brain-injured subjects do not have the hypothesized face-specific recognition deficit (Gauthier, Behrmann, & Tarr, 1999). Finally, neuroimaging studies that have compared the subordinate-level recognition of familiar, common objects to the recognition of faces have revealed common neural substrates for both recognition tasks (Gauthier et al., 1997, 2000b). At the same time, the putative FFA has been found to activate not only for faces, but also for Greebles once subjects are made to be Greeble experts (Gauthier et al., 1999). Reinforcing the importance of expertise in putatively face-specific effects, bird and car experts also show activation in their FFA for birds and cars, but only in their domain of expertise (Gauthier et al., 2000a). Thus, there is little evidence to support the existence “face-specific” perceptual processes or neural substrates. Rather current results point to a single visual recognition system that can be “tuned” by experience to recognize specific object classes by default at a more subordinate level than is ordinarily the case (Tarr & Gauthier, 2000). The exact nature of this system, however, is still open to debate.

Conclusions: Is Any Model Adequate?

Despite our earlier statement that image-based models do a better job of accounting for extant data, it is unclear whether any model provides an adequate explanation of human object recognition. Although various groups have argued that one model or another offers a comprehensive theory (Tarr & Bülthoff, 1995; Biederman & Gerhardstein, 1995), the truth is, at present, there is no single model that can explain the range of behavioral, neurophysiological, and neuropsychological data that has been obtained under various conditions. Indeed, perhaps the most significant challenge to any theory is that human object recognition is so flexible, supporting accurate recognition across a myriad of tasks, levels of specificity, degrees of expertise, and changing viewing parameters.

Consider the case of viewpoint. Many studies have attempted to assess whether object recognition is viewpoint dependent or viewpoint invariant (Biederman & Gerhardstein, 1993; Bülthoff & Edelman, 1992; Lawson & Humphreys, 1996; Tarr, 1995; Tarr et al., 1998), yet asking the question in a strictly dichotomous manner is futile. The fact is, recognition performance varies almost continuously depending on the similarity of the target object relative to the distractor objects (Hayward & Williams, 2000). There is no canonical “viewpoint dependent” result and there are few cases in which recognition is truly viewpoint invariant (Tarr & Bülthoff, 1995). As an alternative, one might abandon the notion of viewpoint-dependency as a guiding principle in favor of similarity metrics between to-be-recognized objects, rendering viewpoint-dependent effects a by-product of the way in which similarity is measured (Edelman, 1995; Perrett, Oram, and Ashbridge, 1998). The problem is that there is currently no reasonable notion of how to measure “similarity.” What is the correct feature set? How are features compared to one another? These questions are thus far unanswered, yet they are central to any theory of object recognition even if one sidesteps many other potentially difficult issues.

Given that there is no such thing as a definitive experiment and data exists that can in some sense invalidate every theory, what can be said about the current state of models of visual object recognition? To begin with, the debate between proponents of structural-description models and image-based models boils down to an argument about the molar features of object representations. On one side, researchers such as Biederman (1987) and Hummel (Hummel & Stankiewicz, 1996) have posited the use of 3D volumes that approximate the 3D appearance of individual object parts – an approach that has its origins in computer graphics (Foley and Van Dam, 1982) and long-standing popularity in the field of computer vision (Binford, 1971; Brooks, 1983; Marr & Nishihara, 1978). On the other side, theorists such as Tarr (Tarr & Pinker, 1989; Tarr, 1995), Poggio and Edelman (1990), and Bülthoff and Edelman (1992) have posited the use of local 2D and 2.5D image features – an approach that is rooted in what is known about the architecture of the primate visual system (Hubel & Wiesel, 1959).

At the same time, both camps effectively agree about many of the properties that are critical in any plausible model of recognition:

- The decomposition of an image into component features.
- The coding of the spatial relations between such features.
- Multiple views for single objects to encode “different” collections of features arising from different viewpoints.
- Generalization mechanisms to normalize over viewpoint and other changes in viewing conditions.
- Plasticity that can support recognition tasks ranging from the highly specific individual level to the categorical entry level.

In the end, things might not be so bleak after all. That there is any agreement at all about such a list suggests that vision scientists have made some progress.

References

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

Biederman, I. (1988). *Aspects and Extensions of a Theory of Human Image Understanding.* Paper presented at the Computational processes in human vision: An interdisciplinary perspective, Norwood, NJ.

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

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

Binford, T. O. (1971, December). *Visual perception by computer.* Paper presented at the IEEE Conference on Systems and Control, Miami, FL.

Booth, M. C. A., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex, 8*(6), 510-523.

Brooks, R. A. (1983). Model-based three-dimensional interpretations of two-dimensional images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5*(2), 140-149.

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

Bülthoff, H. H., & Edelman, S. (1993). Evaluating object recognition theories by computer graphics psychophysics. In T. A. Poggio & D. A. Glaser (Eds.), *Exploring Brain Functions: Models in Neuroscience* (pp. 139-164). New York, NY: John Wiley & Sons Ltd.

Edelman, S. (1993). Representing three-dimensional objects by sets of activities of receptive fields. *Biological Cybernetics, 70*, 37-45.

Edelman, S. (1995). Representation, similarity, and the chorus of prototypes. *Minds and Machines, 5*(1), 45-68.

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

Farah, M. J. (1990). *Visual Agnosia: Disorders of Object Recognition and What They Tell Us About Normal Vision*. Cambridge, MA: The MIT Press.

Farah, M. J. (1992). Is an object an object an object? Cognitive and neuropsychological investigations of domain-specificity in visual object recognition. *Current Directions in Psychological Science, 1*(5), 164-169.

Foley, J. D., & Van Dam, A. (1982). *Fundamentals of Interactive Computer Graphics*. Reading, MA: Addison-Wesley.

Gauthier, I., Anderson, A. W., Tarr, M. J., Skudlarski, P., & Gore, J. C. (1997). Levels of categorization in visual object studied with functional MRI. *Current Biology, 7*, 645-651.

Gauthier, I., Behrmann, M., & Tarr, M. J. (1999). Can face recognition really be dissociated from object recognition? *Journal of Cognitive Neuroscience, 11*(4), 349-370.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000a). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience, 3*(2), 191-197.

Gauthier, I., & Tarr, M. J. (1997a). Becoming a "Greeble" expert: Exploring the face recognition mechanism. *Vision Research, 37*(12), 1673-1682.

Gauthier, I., & Tarr, M. J. (1997b). Orientation priming of novel shapes in the context of viewpoint-dependent recognition. *Perception, 26*, 51-73.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nature Neuroscience, 2*(6), 568-573.

Gauthier, I., Tarr, M. J., Moylan, J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (2000b). Does visual subordinate-level categorisation engage the functionally defined Fusiform Face Area? *Cognitive Neuropsychology, 17*(1/2/3), 143-163.

Hayward, W. G., & Tarr, M. J. (1997). Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23*(5), 1511-1521.

Hayward, W. G., & Williams, P. (2000). Viewpoint dependence and object discriminability. *Psychological Science, 11*(1), 7-12.

Hayward, W. G., & Tarr, M. J. (2000). Differing views on views: Comments on Biederman & Bar (1999). *Vision Research, 40*, 3895-3899.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurons in the cat's striate cortex. *J. Physiol., 148*, 574-591.

Hummel, J. E. (1994). Reference frames and relations in computational models of object recognition. *Current Directions in Psychological Science, 3*(4), 111-116.

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

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

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

Humphreys, G. W., & Riddoch, M. J. (1984). Routes to object constancy: Implications from neurological impairments of object constancy. *Quarterly Journal of Experimental Psychology, 36A*, 385-415.

Jolicoeur, P. (1985). The time to name disoriented natural objects. *Memory & Cognition, 13*, 289-303.

Jolicoeur, P. (1990). Identification of disoriented objects: A dual-systems theory. *Mind & Language, 5*(4), 387-410.

Jolicoeur, P., Gluck, M., & Kosslyn, S. M. (1984). Pictures and names: Making the connection. *Cognitive Psychology, 16*, 243-275.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience, 3*(8), 759-763.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosc., 17*, 4302-4311.

Lawson, R., & Humphreys, G. W. (1996). View specificity in object processing: Evidence from picture matching. *Journal of Experimental Psychology: Human Perception and Performance, 22*(2), 395-416.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology, 5*(5), 552-563.

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

Lowe, D. G. (2000). Towards a computational model for object recognition in IT Cortex. In S.-W. Lee & H. H. Bülthoff & T. Poggio (Eds.), *Biologically Motivated Computer Vision* (Vol. 1811, pp. 20-31). Berlin: Springer-Verlag.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. of Lond. B, 200*, 269-294.

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

McMullen, P. A., & Farah, M. J. (1991). Viewer-centered and object-centered representations in the recognition of naturalistic line drawings. *Psychological Science, 2*, 275-277.

Mel, B. (1997). SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation, 9*, 977-804.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature, 335*, 817-820.

Murphy, G. L., & Brownell, H. H. (1985). Category differentiation in object recognition: Typicality constraints on the basic level advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 11*, 70-84.

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

Perrett, D. I., Mistlin, A. J., & Chitty, A. J. (1987). Visual neurones responsive to faces. *Trends in Neuroscience, 10*(96), 358-364.

Perrett, D. I., Oram, M. W., & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: An account of generalisation of recognition without mental transformations. *Cognition, 67*(1,2), 111-145.

Perrett, D. I., Rolls, E. T., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research, 47*, 329-342.

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

Price, C. J., & Humphreys, G. W. The effects of surface detail on object categorization and naming. *The Quarterly Journal of Experimental Psychology, 41A*, 797-828.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019-1025.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382-439.

Sanocki, T., Bowyer, K. W., Heath, M. D., & Sarkar, S. (1998). Are edges sufficient for object recognition? *Journal of Experimental Psychology: Human Perception and Performance, 24*(1), 340-349.

Selfridge, O. G. (1959). *Pandemonium: A paradigm for learning.* Paper presented at the Symposium on the Mechanisation of Thought Processes, London.

Shepard, R. N., & Metzler, J. (1971). Mental Rotation of three-dimensional objects. *Science, 171*, 701-703.

Shepard, R. N., & Cooper, L. A. (1982). *Mental Images and Their Transformations*. Cambridge, MA: The MIT Press.

Srinivas, K. (1995). Representation of rotated objects in explicit and implicit memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 1019-1036.

Tanaka, K. (1996). Inferotemporal cortex and object vision, *Ann Rev Neuroscience* (Vol. 19, pp. 109-139). Palo Alto, CA: Annual Reviews.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology, 46A*, 225-245.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology, 23*, 457-482.

Tanaka, J., Weiskopf, D., & Williams, P. (In press). The role of color in object recognition. *Trends in Cognitive Science*.

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

Tarr, M. J. (1999). News on views: Pandemonium revisited. *Nature Neuroscience, 2*(11), 932-935.

Tarr, M. J. (in press). Visual Object Recognition: Can a Single Mechanism Suffice? In M. A. Peterson & G. Rhodes (Eds.), *Analytic and Holistic Processes in the Perception of Faces, Objects, and Scenes*. New York: JAI/Ablex.

Tarr, M. J., & Black, M. J. (1994). A computational and evolutionary perspective on the role of representation in vision. *Computer Vision, Graphics, and Image Processing: Image Understanding, 60*(1), 65-73.

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

Tarr, M. J., & Bülthoff, H. H. (1998). *Object Recognition in Man, Monkey, and Machine*. Cambridge, MA: The MIT Press.

Tarr, M. J., Bülthoff, H. H., Zabinski, M., & Blanz, V. (1997). To what extent do unique parts influence recognition across changes in viewpoint? *Psychological Science, 8*(4), 282-289.

Tarr, M. J., & Gauthier, I. (1998). Do viewpoint-dependent mechanisms generalize across members of a class? *Cognition, 67*(1-2), 71-108.

Tarr, M. J., & Gauthier, I. (2000). FFA: A Flexible Fusiform Area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience, 3*(8), 764-769.

Tarr, M. J., Kersten, D., & Bülthoff, H. H. (1998). Why the visual system might encode the effects of illumination. *Vision Research, 38*(15/16), 2259-2275.

Tarr, M. J., & Kriegman, D. J. (2001). What defines a view? *Vision Research, in press*.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21*(28), 233-282.

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint-dependent. *Nature Neuroscience, 1*(4), 275-277.

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

Ullman, S. (1996). *High-Level Vision*. Cambridge, MA: The MIT Press.

Ullman, S., & Basri, R. (1991). Recognition by linear combinations of models. *IEEE PAMI, 13*(10), 992-1006.

Ullman, S., & Sali, E. (2000). Object classification using a fragment-based representation. In S.-W. Lee & H. H. Bülthoff & T. Poggio (Eds.), *Biologically Motivated Computer Vision* (Vol. 1811, pp. 73-87). Berlin: Springer-Verlag.

Wallis, G. (1996). Using spatio-temporal correlations to learn invariant object recognition. *Neural Networks, 9*(9), 1513-1519.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*(1), 141-145.