
Visual Object Recognition:

Can A Single Mechanism Suffice?

Michael J. Tarr

Running Head: Can A Single Mechanisms Suffice?

Mailing address:
Department of Cognitive and Linguistic Sciences
Brown University
190 Thayer Street, Box 1978
Providence, RI 02912 USA

Tel: 401.863.1148
Fax: 401.863.2255
Email: Michael_Tarr@brown.edu
URL: http://www.tarrlab.org/
**VISUAL OBJECT RECOGNITION: CAN A SINGLE MECHANISM SUFFICE?**

“In actual, physical life I can turn as simply and swiftly as anyone. But mentally, with my eyes closed and my body immobile, I am unable to switch from one direction to the other. Some swivel cell in my brain does not work. I can cheat, of course, by setting aside the mental snapshot of one vista and leisurely selecting the opposite view for my walk back to my starting point. But if I do not cheat, some kind of atrocious obstacle, which would drive me mad if I persevered, prevents me from imagining the twist which transforms one direction into another, directly opposite. I am crushed, I am carrying the whole world on my back in the process of trying to visualize my turning around and making myself see in terms of “right” what I saw in terms of “left” and vice versa.”

—Vladimir Nabokov

How do humans recognize 3D objects? This simple question leads to surprisingly complex answers. Indeed, object recognition is sufficiently difficult that state-of-the-art computer vision systems can only perform the most rudimentary visual tasks and, even then, only under highly constrained conditions. At the heart of what makes visual recognition difficult are two factors. First, we live in a world made up of 3D objects, yet only receive 2D stimulation on our retinae as sense input. Second, we live in a highly variable world in which images of objects change constantly due to transformations in size, position, orientation, pose, color, lighting, and configuration. The challenge is to derive a consistent mapping from a potentially infinite set of images to a relatively small number of known objects and categories. It is a problem that the human visual system routinely and effortlessly solves.
How the mammalian brain solves the problem of visual recognition has been a topic of study since the early days of brain science. David Hubel and Torsten Wiesel (1959) received the Nobel Prize for their discovery that neurons in cat visual cortex that respond to boundaries between regions of light and dark are organized into columns according to their orientation preference. This critical result appeared to capture an important facet of visual processing – a visual system that is sensitive to edges positioned at different orientations in space. Once the particular orientations of edges are known, it seemed only a small step to “connect the dots” – joining edges into more complex descriptions of object shape. Edge-based representations appeared ideal for recognition: shape defining edges often capture the critical features of objects and remain relatively invariant over many image transformations. Thus, most vision scientists came to believe that the goal of vision was to derive or reconstruct an edge-based description of object shape.

It was this belief that drove David Marr to develop two ideas that have dramatically influenced the study of visual object recognition. The first was a computational theory in which Marr and Ellen Hildreth (1980) proposed what they saw as the processing constraints needed to build a successful edge detector. They observed that an implemented version of their detector behaved much like some subclasses of visual neurons (so-called “simple cells”). Once Marr had a plausible algorithm for finding local edges in an image, he was able to argue that early visual processing derived increasingly complex representations of shape built from such local edges – at the endpoint of this progression was a 2D description of object shape referred to as the full primal sketch. To Marr, however, the full primal sketch was inadequate as a representation for visual recognition. Although the full primal sketch might be invariant over changes in the appearance of an object’s surfaces due to variations in lighting or small shifts in orientation, as a 2D representation described in a frame of reference based on the observer (“viewer-centered”) the full primal sketch would still change dramatically with changes in object viewpoint. As such, recognition using only the full primal sketch seemed to require a potentially infinite number of 2D descriptions for each
3D object (one description for each unique view). Consequently, Marr believed that the representations underlying visual recognition should be insensitive to changes in viewpoint. This constraint led Marr to develop his second critical idea – 3D parts-based descriptions for object recognition.

Marr and Keith Nishihara (1978) proposed that the full primal sketch is used to derive a 3D representation by first adding information about the relative depths and orientations of surfaces and then by grouping such surfaces into 3D object parts. They argued that 3D volumes known as “generalized cylinders” formed an appropriate method for describing such parts. Critical to their theory was the fact that the visual frame of reference used in the generalized cylinder representation was based on the object itself (“object-centered”). As such, an object’s description would not vary with changes in the viewpoint of the observer relative to the object. To Marr and Nishihara the use of an object-centered reference frame seemed a computationally elegant solution to the problem of how to derive a consistent mapping from variable images to a single object or class. Indeed, to a great many others studying the problem of vision, Marr and Nishihara’s proposal seemed to offer the first tractable answer to the problem of visual recognition. This and related viewpoint-invariant parts-based models, most notably Biederman’s (1987) Recognition-By-Components theory (RBC), have had and continue to have enormous influence.

**Pandora’s Box**

Much like Pandora’s box, simple experiments can sometimes lead to unforeseen results. In 1984 most vision researchers took for granted some version of Marr and Nishihara’s model of object recognition and, in particular, the idea that mental representations of objects are encoded using an object-centered reference frame. Only a handful of experimental psychologists, however, had attempted to test aspects of this model. Steven Pinker and I observed that the central question of whether objects were represented in a viewer-centered or an object-centered reference frame was unanswered.
The few studies that had tried to address this issue had used familiar shapes (e.g., letters – see, Corballis, Zbrodoff, Shetzer, & Butler, 1978) as stimuli and therefore could not distinguish between a true object-centered representation and a viewer-centered representation composed of many specific views of each shape at each familiar vantage point. Either approach would predict equivalent performance in terms of response speed and accuracy across familiar orientations – where they differed was in what happened at unfamiliar viewpoints. Unfortunately, because the stimuli used in most experiments were highly familiar, subjects had, in all likelihood, already encountered them in many orientations. Furthermore, many of the stimuli used in such studies were highly distinctive from one another, for example, the “tail” on a ‘Q’ is not found in any other letter. Thus, apparent evidence for object-centered representations, e.g., equivalent recognition performance for familiar objects across different viewpoints, may be misleading.

Insert Figure 1 Here

To address these concerns Pinker and I designed a set of stimuli that fulfilled three criteria: (1) novelty; (2) no distinct local features; and (3) well-defined axes (to eliminate any need for viewpoint-dependent axis-finding procedures that might serve as a necessary precursor to recovering 3D parts). These simple 2D shapes (Figure 1a) were used in a straightforward test of object- vs. viewer-centered reference frames (Tarr & Pinker, 1989). Subjects were trained to associate nonsense names (e.g., “KIP”) with individual shapes shown at the upright and then practiced naming the same shapes at selected orientations in the picture-plane. During this practice we observed that when subjects first encountered the shapes in these new orientations, their naming performance varied as a function of the distance between the unfamiliar orientations and the trained upright orientation. Although this pattern of viewpoint dependence could be taken as evidence for viewer-centered representations, we reasoned that object-centered descriptions might take some time to develop. Indeed, as subjects became more and more practiced at naming the objects, their performance at all of the previously-unfamiliar orientations became equivalent. Much as with
familiar objects, this viewpoint invariance may be attributed to two quite different underlying causes – subjects learning a viewpoint-invariant object-centered representation for each shape or subjects learning multiple viewpoint-specific viewer-centered representations for each shape.

To distinguish between these two possibilities Pinker and I introduced a condition in which subjects were shown the now-familiar shapes in new, never-before-seen orientations. If subjects had learned the shapes in an object-centered format, then their performance at the new orientations should have been no different from their performance at the familiar orientations. Much to our surprise this is not what we found (we had been expecting to confirm Marr’s hypothesis!) – subjects’ performance at the new orientations varied systematically with the distance from the nearest familiar orientation. This result is exactly what is predicted if subjects were learning multiple orientation-specific shape representations at each familiar orientation – a “multiple-views” description (Figure 2). We also observed that the magnitude of this viewpoint dependence was almost identical at the beginning of the experiment when subjects were familiar with the shapes in only the upright orientation and at the end of the experiment when subjects were familiar with the shapes in multiple orientations. Thus, a single shape recognition mechanism, based on learning objects in specific viewpoints and then mentally transforming unfamiliar viewpoints to the familiar views (sometimes referred to as “normalization”) appeared to be at work.

Insert Figure 2 Here

This apparently straightforward finding of viewpoint dependence in object recognition has become a critical result in motivating the research I have pursued in the years since. At the core of this research is the question of how humans recognize objects. At a more specific level, answering this question will require understanding the contributions of three aspects of recognition:

• The image geometry for objects and object classes and how it changes with changes in 3D orientation, illumination, etc.;
• The level of categorization required for a given task, varying from coarse “basic-level” categorization to fine item-specific recognition;
• The differing degrees of perceptual expertise that observers have with specific object classes and how visual experience fine-tunes the recognition system to attain such expertise.

In the coming sections on three-dimensional object recognition, perceptual classification, and perceptual expertise I take up each of these issues in turn, focusing on aspects of visual recognition for which these issues are central. The actual research reviewed in these sections explores these different aspects of recognition through a variety of converging techniques, including computer-graphics psychophysics, brain imaging using functional Magnetic Resonance Imaging (fMRI), and neuropsychology with brain-injured subjects.

THREE-DIMENSIONAL OBJECT RECOGNITION

Given evidence for multiple-views across changes in picture-plane orientation, it was natural to wonder whether the result extended to the recognition of 3D objects rotated in depth. There are reasons to believe that rotations in depth might be treated differently from rotations in the picture-plane. The former produce geometric changes in the structure of the 2D image as 3D surfaces come in and out of view and change their orientation relative to the observer. In contrast, the latter have no impact on the 2D image structure, but do change the top-bottom and left-right relations between features relative to the observer. Thus, in the latter case viewpoint dependence may be a consequence of changes in the spatial relations within objects, rather than a fundamental organizing principle for visual representation.

Using what was (at the time!) state-of-the-art 3D software, I created a set of 3D stimuli that fulfilled the same criteria used in the original Tarr and Pinker study (Figure 1b; although similar to the objects used in Shepard and Metzler’s, 1971, classic study of mental rotation, my objects all contained a vertical axis clearly marked by a small “foot”). As in our earlier
study, I trained subjects to name each object from the near-upright orientation and then had subjects practice recognizing the same objects in a selected set of viewpoints. Here, however, the practice viewpoints were generated by rotations in depth around either the vertical or the horizontal axis, as well as by rotations in the picture-plane. Following practice, new, never-before-seen-viewpoints were generated by similar rotations interspersed among the now-familiar viewpoints. The results were astonishingly clean – for each axis of rotation there was a clear pattern of viewpoint dependence (Tarr, 1989, 1995). At the beginning of the experiment this pattern was systematically related to the distance from the single training view. Following practice and the introduction of new unfamiliar viewpoints, this pattern was systematically related to the distance from the nearest familiar view regardless of whether the view was generated by a rotation in depth or in the picture-plane. As before, the magnitude of this pattern was quite similar during both the initial recognition of new views and the recognition of additional new views following practice. Thus, together with my earlier findings, there seemed to be good evidence for generally applicable viewer-centered visual recognition mechanisms.

Beyond the basic finding of viewpoint dependence, there was one other piece of evidence that strongly implicated viewpoint-dependent recognition processes in both the 2D and 3D versions of my experiments. In the 2D case we had run a variant in which subjects learned a “standard” version of each shape, but then following practice, were asked to apply the same name to mirror-reflected versions of the familiar shapes. We were puzzled to find that subjects showed the same relatively slow performance regardless of the picture-plane orientation of the unfamiliar mirror-reflected shape. Pinker and I came to the realization that this is exactly what is predicted by a recognition process that is normalizing unfamiliar mirror-reflected shapes at any orientation to a familiar view of the familiar standard version – for any orientation of a mirror-reflected shape, the shortest path to its standard is a 180° flip-in-depth (Figure!3).

Insert Figure 3 Here
This result in itself does not provide definitive evidence for normalization between standard and mirror-reflected versions of shapes. One alternative is that when subjects encounter a mirror-reflected version of a familiar shape, they cannot deduce (consciously or unconsciously) a transformation that will align it with its standard and therefore resort to viewpoint-invariant object-centered features. In the 3D version of my experiment, however, I obtained results that ruled out this alternative – when subjects first encountered mirror-reflected versions of familiar 3D objects there was no rotation in 3D space that would bring the two into correspondence. Just as aligning mirror images of 2D shapes requires a 3rd dimension, aligning mirror images of 3D objects requires a 4th dimension. Thus, subjects either needed to employ a 4D mental transformation, turn the mirror-reflected object inside-out (in the same way that an inside-out left-handed glove matches a right-handed glove), or use some other strategy. In the 3D case subjects did not show equivalent performance across different unfamiliar viewpoints of mirror-reflected objects – rather they exhibited viewpoint dependence quite similar to that found for the recognition of standard versions of the objects in unfamiliar viewpoints. One possible alternative strategy in the face of a 4D transformation is to normalize the major vertical axis of the mirror-reflected objects into correspondence with the standard version and then observe whether the protruding parts were symmetrical between the pair. Regardless of the particular process used by subjects to compensate for changes in viewpoint (including “non-transformational” mechanisms such as those proposed by Perrett, Oram, and Ashbridge, 1998), the combination of findings from the 2D and 3D experiments provide compelling evidence for a multiple-views object representation that is matched to object images through viewpoint-dependent normalization procedures (Figure 2).

**Perceptual Classification Using Viewpoint-Dependent Mechanisms**

Object recognition, broadly defined, involves a wide range of tasks, including the recognition of specific individual objects and the classification of many different objects into a
single category. The results from my 2D and 3D recognition experiments suggest that some aspects of recognition rely on viewpoint-dependent mechanisms, but do not address whether this is the only process available for recognition. Indeed, given that I intentionally made the stimuli highly confusable, my results may speak more to object-specific recognition, so-called “subordinate-level” tasks, rather than object categorization, so-called “basic-level” tasks. This hypothesis became known as the “dual systems” approach (e.g., Jolicoeur, 1990) and posited that while subordinate-level recognition might involve viewpoint-dependent processes, basic-level recognition, often thought to be the “default” level of access, was based on viewpoint-invariant processes.

Evidence for this dichotomy was rather slim, but gained some support from a study by Irving Biederman and Peter Gerhardstein (1993) in which viewpoint-invariant recognition performance was obtained for several sets of 3D objects containing distinctive features or parts – similar to basic-level categorization. Based on their results, Biederman and Gerhardstein argued that human object recognition typically occurs at the basic-level where distinctive features are available to discriminate between object classes and therefore recognition is viewpoint invariant. In contrast, they argued that discrimination tasks on objects lacking distinctive features are the exception rather than the rule and therefore viewpoint-dependent recognition is rarely observed in the real world. My students, William Hayward, Isabel Gauthier, and Pepper Williams, and I wondered about this result. Although we had no strong belief that viewpoint-dependent processes extended to basic-level recognition, there were methodological issues in Biederman and Gerhardstein’s study that led us to question their conclusions. In particular, in some experiments they used familiar common objects that were likely to have already been learned at multiple viewpoints, thereby masking any viewpoint-dependent recognition processes. In other experiments they used line drawings of novel objects containing highly distinctive features, but a task in which subjects were required to remember only one object at a time over a series of trials – such
a task might have predisposed subjects to adopt a strategy in which they selectively relied on the local distinct features of the single object.

**Are Basic-Level Recognition Tasks Viewpoint Dependent?**

In order to assess the role of viewpoint-dependent recognition processes in discriminations that corresponded to the basic-level, we created several sets of novel 3D objects. Objects in each set were composed of combinations of parts based on the parts used by Biederman and Gerhardstein (1993). By using combinations of parts we were able to manipulate the similarity between objects within a given set. Specifically, in one set each object consisted of a single distinctive part and in several sets each object consisted of a central unique body-part with additional parts attached to it (examples of “multi-part” objects are shown in Figures 1c-d; the object in Figure 1d is based directly on an object used by Biederman & Gerhardstein). Although, these “additional parts” sometimes appeared in other objects in a given set, a particular part/position combination was never repeated. Thus, not only were objects in these latter sets unique in terms of their central parts, but the order of all of the parts in a given object was also diagnostic for the identity of that object. That is, although the objects were never explicitly divided into categories, each object within a set was characterized by a unique set of 3D parts corresponding to a shape-defined basic-level class (Rosch et al., 1976). Employing each set separately, we ran an extensive series of experiments using a variety of recognition tasks, including object naming, pairwise matching, and the single object memory task used by Biederman and Gerhardstein. To our astonishment the results were unequivocal in each and every case – viewpoint-dependent recognition performance was found regardless of the object set and the recognition task (Tarr, Williams, Hayward, & Gauthier, 1998; Hayward & Tarr, 1997). These effects, albeit smaller than those obtained for objects that did not contain distinctive features, were highly systematic across changes in 3D viewpoint and were unaffected by whether or not parts became visible or disappeared from view. We could reach only one
Visual Object Recognition

conclusion – visual object recognition, regardless of the level of categorization, is mediated by viewpoint-dependent mechanisms.

This finding of viewpoint dependence was potentially problematic for extant viewpoint-dependent models of recognition which assumed template-like object representations, for instance, describing objects in terms of the X-Y coordinates of features in the 2D image (Poggio & Edelman, 1990) or the output of simple image filters similar to those seen in early vision (Edelman, 1993). Such representations did not seem suitable for supporting the visual recognition of object classes in that this task requires generalizing from known instances of a class to novel instances of the same class. Put another way, class-level recognition of objects requires a many-to-one mapping in which objects of varying shape are treated as equivalent. Undifferentiated two-dimensional template-like representations seemed ill-suited for this task in that no mechanisms had been offered for relating the similarity of visually-different instances of each object class. Thus, two objects with only slightly different shapes might still be treated by the recognition system as completely unrelated. Within viewpoint-invariant part-based models, however, the class-level recognition is solved by using qualitative descriptions of object parts (Biederman, 1987). Consequently, visually-similar instances of an object class are treated as related (or even identical), because they give rise to the same part-based representation, regardless of small variations in object shape. The challenge before us was to understand how viewpoint-dependent multiple-views object representations could support basic-level, as well as subordinate-level, recognition (see Tarr & Bülthoff, 1998).

Isabel Gauthier and I were intrigued by several computational models that used “multi-dimensional feature interpolation” to generalize from one instance of a class to new, never-before-seen instances of the same class (Beymer & Poggio, 1996; Moses, Ullman, & Edelman, 1996). This ability to generalize would allow a single model to account for the recognition of individual objects, that is, subordinate-level tasks, and for the recognition of object classes, that is, basic-level tasks. Such models assumed that viewpoint-dependent
object representations were composed of large numbers of viewpoint-dependent local features and that recognition of any object whether identical to a known object or unfamiliar involved measuring the local similarity between such features. Thus, as an alternative to globally attempting to compare input images to views of an objects, these models proposed local comparisons in which features of each object representation could “vote” for their presence in the image (Perrett et al., 1998, provide neurophysiological evidence for similar “accumulation of evidence” mechanisms playing a role in viewpoint-dependent object recognition). The visual system could then simply tally up the votes for each object or class and decide on a winner (a particularly promising instantiation of this approach is presented in Riesenhuber & Poggio, 1999; the importance of this model for understanding visual recognition is discussed in Tarr, 1999).

Do Viewpoint-Dependent Mechanisms Support Within-Class Generalization?

Gauthier and I were interested in testing aspects of these models behaviorally. In particular, we asked whether viewpoint-dependent recognition processes could generalize between instances of a class. To this end we created a set of 2D shapes similar to the shape shown in Figure 1a. We taught subjects individual names for the shapes and then had them repeatedly name each shape in several different orientations (a subordinate-level discrimination task among members of a basic-level class). Gauthier and I observed that when subjects named a shape in a given orientation and this trial was preceded by a trial in which the same shape had appeared in the same orientation, there was no effect of orientation – performance was equivalent regardless of the actual orientation of the shape (Gauthier & Tarr, 1997a). This result suggested that the visual recognition system had residual activation about the viewpoint-specific appearance of object shape from the previous trial that facilitated recognition on the subsequent trial. That is, there was visual generalization from a 2D shape to the identical 2D shape.

This result might not seem surprising. What is more important is that, as a set, the 2D shapes constituted a basic-level object class. Therefore, it was also possible to ask
whether there was visual generalization from one instance of the basic-level shape class to a different instance of that same class. We designed a second experiment to test whether orientation information about one member of the class (a named shape) would transfer to another member of this class (a differently named shape). That is, although the task was subordinate-level naming, predictions were made about how the recognition of one member of a class would be affected by the prior recognition of another member of that same class. Thus, the experiment investigated the object representation common to two or more individuals – the basic-level for those particular exemplars. In this experiment we manipulated the degree of similarity between the preceding shape and the subsequent shape. We again found that repeating the same shape in the same orientation facilitated viewpoint-invariant recognition performance. Crucially, we also found that repeating a visually-similar shape (same basic-level category), but not a visually-dissimilar shape (different basic-level category), in the same orientation also facilitated performance that was independent of orientation (a similar orientation-dependent priming effect has been reported for the assignment of figure and ground; Gibson & Peterson, 1994). Thus, we obtained evidence for viewpoint-specific transfer from one instance of a class to a new member of that class. We concluded that recognition processes generalize across visually-similar objects appearing in the same viewpoint even when the recognition task is to discriminate between these objects. Thus, based on evidence that viewpoint-dependent information about an individual object generalizes to other, visually-similar objects, that is, other instances of the same basic-level object class, viewpoint-dependent mechanisms can mediate basic-level as well as subordinate-level recognition.

Gauthier and I also created a set of 3D objects (Figure 1c) to investigate aspects of this question. The 3D set consisted of 6 pairs of differently-named objects (“cohorts”) that within a pair shared the same distinct central part and the same attached parts, but with the attached parts in different arrangements so that the two members of each pair could be distinguished. Thus, each pair formed a distinct object class. We were interested in two
specific predictions of a viewpoint-dependent class-general model of recognition. First, the processes we had observed for generalization across rotations in the picture-plane should also apply to rotations in depth. Second, in our earlier study subjects were always shown all of the objects in all of the test views – here we wondered whether a known view of one member of a class would facilitate the recognition of the other instance of the class that had never been seen in that view. Gauthier and I ran an experiment in which one member of each pair was shown to subjects in several viewpoints (different for each pair) and its cohort was shown to subjects only in one viewpoint. Following training, subjects named both members of each pair in a wide range of viewpoints, both familiar and unfamiliar. For the object actually seen at multiple viewpoints we observed a pattern of viewpoint dependence similar to that seen in my original 3D experiments – recognition performance was systematically related to the distance from the nearest familiar view. At the same time we found that the cohort objects that had only been trained in one viewpoint showed a similar pattern of viewpoint dependence (Tarr & Gauthier, 1998). It was as if subjects had actually seen these objects from the viewpoints in which their cohorts had been seen and, therefore, were able to use such information to recognize these objects in what were completely unfamiliar viewpoints.

Before we could conclude that there was indeed viewpoint-specific transfer between visually-similar cohorts we needed to run a critical control. An alternative account was that similar object geometry for both members of a cohort, in particular for new views generated by rotations in depth, produced the common performance patterns found for both members of each pair. We tested this possibility by running an experiment in which subjects were again trained to name both members of each pair, but where both members of a cohort pair were shown only in a single, common viewpoint during training (in the previous experiment one member of each pair was seen in several viewpoints during training). As before, we then tested their recognition performance across a wide range of viewpoints. The elegance of this control is that the cohort objects shown in only one
viewpoint in the first experiment were shown in exactly the same viewpoints during training and test in this experiment – the only difference between experiments was the viewpoints in which the other member of each pair was shown. Compared to our first experiment we found different results – subjects’ recognition performance for the cohort objects was highly viewpoint dependent, but now related to the distance from the single trained viewpoint. Most importantly, there was a dramatic difference between this experiment and the first experiment in terms of naming performance around the cohort views. In the first experiment the cohort objects showed facilitation as if they had actually been seen in these views, here there was no such facilitation. Taken together, our 2D and 3D studies indicate that viewpoint-dependent mechanisms are capable of class generalization and can support a range of recognition tasks spanning the continuum from basic-level categorization to item-specific recognition.

What Object Properties Mediate Viewpoint Dependence?

Although the results reviewed to this point provide evidence for the wide applicability of viewpoint-dependent mechanisms, they do not address the specifics of what comprises a viewpoint-dependent representation. What is necessary is a format that is sufficiently flexible to adapt to tasks at many different categorical levels. My colleague Heinrich Bülthoff and I had observed that although the modal finding across many studies was viewpoint-dependent performance, the magnitude of this effect appeared to vary. We hypothesized that the mediating variable was the discriminability of the stimuli that were to be distinguished. We were interested in examining what kinds of representational features played a role in modulating these viewpoint-dependent processes.

Insert Figure 4 Here

To that end we (Tarr, Bülthoff, Zabinski, & Blanz, 1997) created four sets of novel 3D objects based on “paperclip” objects used in studies by Bülthoff and Shimon Edelman (1992): objects comprised of five “tubes” connected end to end (Figure 4a); objects
comprised of a distinctively-shaped part in the central position and two tubes attached to each end of this part (Figure 4b); objects comprised of a distinct sequence of three different parts with one tube attached to each end (Figure 4c); and objects comprised of a distinct sequence of five different parts (Figure 4d). For the objects with 3 and 5 distinctive parts some parts appeared in more than one object, thereby making the local features, but not the part configurations, of a given object confusable with other objects in the set. The angles between tubes or parts were also varied for each object in each set so that along with differences in part shape, each object could be distinguished on the basis of the configuration of parts.

Using both a pairwise matching task and a naming task we trained subjects to recognize objects in one viewpoint and then examined recognition performance across rotations in depth. Results for the five-tube and single distinctive part sets were consistent with our previous work – the magnitude of viewpoint dependence was quite large in the five-tube condition where no distinctive shape features were available, while the magnitude of viewpoint dependence was much smaller in the single distinctive part condition (Tarr, Bülthoff, Zabinski, & Blanz, 1997). What is critical is the magnitude of this effect in the three distinctive part and five distinctive part conditions. Viewpoint-invariant part-based models (e.g., Biederman, 1987; Biederman & Gerhardstein, 1993) hypothesize that the building blocks of object representations are qualitatively-defined parts, that is, 3D volumes that are recovered from the 2D image on the basis of viewpoint-invariant contrasts such as parallel/non-parallel and straight/curved. Critically, such parts are both individually and as a group recognized in a viewpoint-invariant manner (because individual parts are so recognized)\(^1\). Therefore, these models appear to predict that adding distinct parts as in the three and five distinctive part conditions should, at a minimum, be no different than the single distinctive part condition and, at best, even less sensitive to changes in viewpoint given the additional distinguishing representational elements (for example, single distinctive part objects necessarily differ in only one part, but the three distinctive part objects always
differed from one another by at least two parts). In contrast, Bülthoff and I (Tarr & Bülthoff, 1995) hypothesized that the building blocks of object representations are viewpoint-dependent features, for example, local surface patches that include texture and color. Critically, these features are both individually and as a group recognized in a viewpoint-dependent manner. At the same time, we pointed out that highly-distinctive local features (although not configurations of such features) can circumvent this process by uniquely specifying the identity of an object, for instance, if each object in a set was painted with a differently colored spot (Tarr & Bülthoff, 1995). Therefore, we predicted that the most distinctive condition would be the single distinctive part case where the local features within each central part were distinctive from those in the other objects in the set. In contrast, in the three and five distinctive part conditions, the additional parts might be distinctive as elements of a configuration, but as local features they would add confusion across objects in that at least one other object in the set would contain the same features. Consequently we predicted that viewpoint dependence would increase in magnitude in the three distinctive part condition and would be even larger in the five distinctive part condition – possibly reaching the same level observed in the five-tube condition. This is precisely what we found in both the matching and naming tasks, indicating that object recognition involves viewpoint-dependent processes that are sensitive to the distinctiveness of local image features.

**Insert Figure 5 Here**

A similar conclusion can be reached from an experiment I ran with William Hayward (Hayward & Tarr, 1997). Here we had subjects simply recognize isolated parts across rotations in depth. For each part we generated two rotations of equal magnitude: one in which all of the local image features were distorted and one in which the local image features changed qualitatively (Figure 5). Subjects learned a target view and then recognized one of the two rotated views in a sequential-matching task. Part-based models unquestionably predict that recognition should be equivalent across these two conditions in that the same
part description should be derived in each instance. In contrast, the image feature model predicts that changes in the local features should affect performance. In the experiment performance was consistent with a local feature account – subjects were faster to recognize the objects when the rotation produced only a distortion of the features present in the original image as compared to when the particular configuration of features present in the original image changed into a new configuration of features with the rotation. Although there is clearly a great deal of work to be done regarding the types of features that form viewpoint-dependent object representations, our results suggest that the geometry of objects is described in terms of configurations of local features – depending upon the distinctiveness of the features necessary to perform a given discrimination recognition performance may appear more or less viewpoint dependent. Thus, we can hypothesize that a single visual recognition mechanism is sufficient to mediate the continuum of categorization tasks, from basic-level classification to item-specific recognition, with which we are presented.

**Perceptual Expertise and the Fine-Tuning of Recognition Mechanisms**

Up to this point I have focused on how object geometry and the specificity of the level of categorization can influence recognition performance. It is clear, however, that these two factors alone are insufficient to account for human recognition competence. A third factor, the degree of experience or expertise an observer has with a given object class is equally important for understanding recognition behavior. Perhaps the most salient example of this factor of recognition is face recognition – humans are unquestionably more expert at recognizing human faces at the individual level than they are at recognizing members of any other stimulus class. There are many reasons for this high level of expertise. Two of the most critical are the social significance of faces and the fact that we are apparently predisposed to be interested in individual faces from the first moments of birth onwards (Johnson & Morton, 1991). The fundamental question is whether humans are simply
biologically predisposed to detect faces and then learn them using more generic visual recognition processes or whether humans have face-specific recognition mechanisms distinct from other recognition processes. Given the picture I have painted of human visual recognition as a flexible system capable of supporting a wide range of tasks, it has been my working hypothesis (in particular in collaboration with Isabel Gauthier) that while face recognition is certainly the most complex discrimination task most of us ever learn to perform, it is still part and parcel of general recognition mechanisms, albeit mechanisms that have been tuned to recognize specific faces through many years of experience with objects.

There have been several recent claims to the contrary. Researchers from both the behavioral and cognitive neuroscientific domains have garnered evidence that seems to suggest that faces are “special” and are processed by a recognition system distinct from that used for non-faces (Kanwisher, 2000). Gauthier and I, however, observed that many of these studies, regardless of domain, had a common flaw – they tended to confound stimulus class, faces vs. objects, with the level of categorization, item-specific level vs. basic-level, and the level of expertise, expert vs. novice (Tarr & Gauthier, 2000). To better understand the relationship of face recognition to normal object recognition we embarked on a series of experiments using behavioral psychophysics, fMRI, and brain-injured subjects. The underlying theme to all of these studies is careful control of the level of categorization and the level of expertise in addition to manipulation of the stimulus class.

**Behavior.**

One of the most important pieces of evidence cited in favor of face-specific mechanisms is what is referred to as “configural sensitivity.” Although this sometimes confusing term has been defined in many ways, the essential idea is that features within the representation are spatially located relative to one another with a great deal of specificity. For example, James Tanaka and Martha Farah (1993) reported that subjects were faster to recognize a part of an individual’s face, e.g., Bob’s nose, when it was shown in the context of the face as originally
learned, e.g., Bob’s nose in Bob’s face, as compared to recognizing the same part when it was shown in the face with other parts in altered spatial positions, e.g., Bob’s nose in Bob’s face with the eyes moved further apart. Crucially, they observed that the same effect could not be obtained for either inverted faces or for non-face objects, for instance, recognizing doors of houses with the windows either in the same position or relocated. Although this result might seem to provide some support for face-specific effects in part recognition, Gauthier and I felt that houses were an inadequate control for faces. In particular, few if any of Tanaka and Farah’s subjects were likely to be house experts, and even if they were, houses form a far more heterogeneous class as compared to faces.

**Insert Figure 6 Here**

To provide a better control set for faces, Gauthier and I supervised the creation of what became known as the “Greebles” (created by Scott Yu). Greebles were designed as a set of 60 objects that shared similar parts in similar spatial configurations (Figure 1e). What made Greebles unique was that, as with faces, they were also organized hierarchically so that an individual Greeble could be recognized at multiple levels of categorical difficulty – a coarse level referred to a gender, an intermediate level referred to as family, and at the individual level (Figure 6). Gauthier and I designed a test of configural sensitivity similar to that employed by Tanaka and Farah. Subjects learned to name individual parts of each Greeble and were then tested on the recognition of such parts in either intact or altered configurations of the other parts. The critical manipulation in our study was not the use of Greebles *per se*, but how much experience subjects had at recognizing them. One group of subjects were tested as novices, that is, they had almost no practice at recognizing Greebles. A second group of subjects were tested as experts, that is, they had extensive practice recognizing Greebles. To ensure that subjects were sufficiently expert, we used the rule that subjects had to practice recognizing Greebles until they were just as fast to name individual Greebles as they were to identify the Greeble gender or family – a fact of recognition behavior that is true for identifying faces (Tanaka, in press; Tanaka & Gauthier,
1997), as well as other domains of expertise (Tanaka & Taylor, 1991). Our results supported the hypothesis that expertise, not faces per se, was responsible for the configural sensitivity observed in face recognition. Experts, but not novices, showed configural sensitivity in the recognition of Greeble parts, but only for Greebles in the trained upright orientation (Gauthier & Tarr, 1997b). That is, as with faces, we found that reported that subjects were faster to recognize a part of an individual Greeble, for instance the bottom protruding part of the top left Greeble shown in Figure 6, when it was shown in the context of the Greeble in which it was originally learned, for instance, the configuration of the top left Greeble as shown in Figure 6, as compared to recognizing the same part when it was shown in the same Greeble with some of its parts in altered spatial positions, for instance, the top left Greeble in Figure 6 with the left and right parts moved forward. In a subsequent study we have compared a wide range of putatively face-specific behavioral effects across Greeble novices and Greeble experts and have consistently obtained face-like patterns of performance with experts, but not novices (Gauthier, Williams, Tarr, & Tanaka, 1998). For example, we found that practice recognizing picture-plane inverted Greebles had no impact on expert performance, but affected novice Greeble recognition. Although experts were faster overall at recognizing Greebles, practice with inverted Greebles did not diminish the advantage in recognition times for upright as compared to inverted Greebles. In contrast, the same amount of practice completely eradicated the difference in response times between upright and inverted Greebles for Greeble novices. Thus, we observe the same difficulty in encoding inverted Greebles for Greeble experts that is seen for inverted faces and for experts with other homogeneous object classes for their domain of expertise (Diamond & Carey, 1986).

Neuroimaging.

A second source of evidence cited in favor of face-specific mechanisms comes from recent work in brain imaging, and, specifically fMRI. Several groups have been interested in the question of whether there is an identifiable neural module for face processing in the
human brain (Kanwisher, McDermott, & Chun, 1997; Puce, Allison, Gore, & McCarthy, 1995; Sergent, Ohta, MacDonald, 1992). A typical imaging study has involved having subjects either recognize or passively view familiar faces in one condition and common objects in the baseline condition. By subtracting the activation (as indicated by the degree of blood oxygenation in different areas of the brain) of the object condition from the face condition, one can localize the brain regions where additional processing occurs for faces relative to objects. Across multiple studies, researchers consistently found that one region of visual cortex, the fusiform gyrus, is especially active in face recognition.

Again Gauthier and I wondered whether the best controls had been applied in these studies. Along with issues raised regarding the level of expertise, an obvious confound with stimulus class was the level of categorical access. Subjects in these experiments had to identify faces at the individual level (“Bob”), but objects only at the basic level (“bird”). In the studies using passive viewing a similar problem existed – the default level of access for faces is almost certainly the individual level, while for common objects it is typically the basic level. What we proposed was a face recognition study without faces. We took a large collection of pictures of familiar common objects that could be readily named at both the basic and subordinate levels, verified that the default level of access for each was the basic level, and used the objects in an fMRI study. We compared brain activity in two conditions: one in which subjects matched a basic-level label to each picture (e.g., “bird” followed by a picture of a pelican) and one in which subjects matched a subordinate-level label to each picture (e.g., “pelican” followed by the same picture of a pelican). To control for possible differences in the semantic processing of basic- and subordinate-level labels we also ran a purely semantic task at both levels and subtracted the semantic activation from the picture-matching activation for both conditions. What we found bore out our hunch regarding earlier studies – as shown in Figure 7, across 8 subjects we obtained a bilateral pattern of activation for the subordinate-level recognition of objects above and beyond the basic-
level recognition of objects that was remarkably similar to that previously found for faces (Gauthier, Anderson, Tarr, Skudlarski, & Gore, 1997c).

**Insert Figure 7 Here**

Although our imaging results point towards a brain region that mediates subordinate-level recognition regardless of stimulus class, it is possible that by averaging across subjects we masked differences in the area mediating face processing and the area mediating object processing *within individual subjects*. Thus, while the face and object areas may appear similar on average, individuals may have separable regions. Gauthier and I have recently addressed this concern by locating individual subjects’ “face areas” using passive viewing of faces vs. objects. We then ran the same individuals in a basic-level vs. subordinate-level recognition task with common objects. The individual results replicated our earlier group-averaged results even when the “face area” was defined precisely for each subject. Thus, we were able to observe the same pattern of brain activation for subordinate minus basic-level visual recognition in individual subjects (Gauthier et al., 2000b). This replication also included somewhat stronger methods (e.g., auditory presentation of object labels) and more stringent data analyses that established that the same voxels active in face recognition were active in subordinate-level recognition of non-face objects. Thus, within the resolution limits of fMRI, there appears to be no reason to posit brain regions specialized for the recognition of faces.

Gauthier and I were also interested in whether the specificity of the categorization judgment was the only factor driving activation in the fusiform gyrus. From our Greeble studies we knew that manipulating the level of expertise could produce a variety of different “face-specific” behavioral effects (including both configural sensitivity and inversion effects) with non-face objects (both novices and experts performed subordinate-level discriminations, so the level of access could not account for our earlier results). Thus, expertise level was a second factor confounded with the stimulus class in some imaging studies. The logical, but risky, study to run was one in which we created Greeble experts
and used fMRI to monitor the reorganization of visual cortex over the acquisition of expertise. Such a design would allow us to study the role of expertise separately from any effects arising from the level of categorization.

**Insert Figure 8 Here**

As part of her Ph.D. dissertation, Gauthier (Gauthier et al., 1999b) compared the brain regions active in the processing of faces and Greebles when subjects were novices and when subjects were experts. For both types of stimuli we used two methods to localize the areas of extrastriate cortex specifically associated with processing a given class of objects: passive viewing of faces or Greebles as compared to passive viewing of common objects; pairwise matching of upright faces or Greebles as compared to pairwise matching of inverted faces or Greebles. The activation patterns for novices were as predicted based on earlier imaging and behavioral studies of face recognition – a “face-specific” area in fusiform gyrus that was present for faces vs. objects and for upright vs. inverted faces, but not for Greebles vs. objects or upright vs. inverted Greebles. For Greeble experts a quite different and somewhat remarkable pattern was obtained. As shown in Figure 8, experts exhibited much more focal activation in visual cortex and, in particular, activation in right fusiform gyrus in the precise location associated with face processing as measured in our two face recognition tasks (Greebles minus Objects and Upright Greebles minus Inverted Greebles). We also monitored changes in subjects’ behavioral processing of Greebles during the acquisition of expertise and replicated our earlier results demonstrating that Greeble experts show a range of “face-specific” effects when recognizing Greebles.

Gauthier followed up on our Greeble expertise study with an elegantly designed experiment in which she examined the brain mechanisms recruited by *extant* experts. She reasoned that if putatively “face-specific” extrastriate areas are recruited through experience, then perceptual expertise with almost any homogeneous category should produce a neural activation pattern similar to that obtained for faces and Greebles (in experts). Indeed, this is
exactly what Gauthier, Skudlarski, Gore, and Anderson (2000a) found for both bird and car experts. Gauthier et al.’s results are particularly compelling in that they report activation in face-selective brain areas for bird experts when recognizing birds, but not cars, and for car experts when recognizing cars, but not birds – an interaction revealing that the preferred category depends directly on expertise. Reinforcing the connection between expertise in activation in face-selective brain regions, is a correlational analysis between the bird and car experts’ behaviorally assessed levels of expertise and their relative level of neural activation for birds and cars as measured in the right fusiform gyrus. Gauthier et al. (2000a) found r’s of 0.82 and 0.75 respectively for bird and car experts – astoundingly high correlations between a standard psychophysical test and the specific activation pattern obtained in a neuroimaging study.

Thus, behavioral and concurrent neural evidence is helping us to understand the mechanisms that lead to functional specialization in visual cortex. Rather than being organized along conceptual object categories, specific cortical areas are part of a highly plastic visual recognition system that can be tuned to perform efficient fine-level visual discriminations. Our manipulation of factors such as the level of categorization and the degree of expertise have revealed some of the elements of this complex system. Further understanding has been obtained by analyzing the responses of category-selective cortical regions other than fusiform gyrus. In particular, we have identified a region (bilaterally) in occipital lobe that is also selective for faces and other domains of expertise (Gauthier et al., 2000c). This functionally-defined region is located relatively early in the visual pathway, indicating that the mechanisms responsible for expert recognition of objects from homogeneous classes are widely distributed.

Neuropsychology.

A third and highly compelling source of evidence cited in favor of face-specific mechanisms comes from neuropsychology. Following brain injury due to stroke, head impact, or other insult some individuals appear to be dramatically impaired at visually
recognizing objects even though their early perceptual mechanisms are intact. A particular version of this syndrome provides evidence for a specialized face processing system – these brain-injured subjects, known as “prosopagnosics,” show impaired face recognition, but are relatively good at non-face object recognition. Although researchers have pointed out that this deficit may apply more generally to the discrimination of visually-similar members of any homogeneous category (Damasio, Damasio, & Van Hoesen, 1982), there has been a general consensus that prosopagnosic subjects provide one of the strongest pieces of evidence for face-specific recognition mechanisms. In a test of whether prosopagnosia is a face-specific impairment, rather than a consequence of the item-specific nature of face recognition, Martha Farah and her colleagues compared the recognition of faces and homogeneous classes of common objects, chairs or eyeglasses, for normal control subjects and one prosopagnosic subject. They found that in comparison to the controls the prosopagnosic subject’s recognition performance was disproportionately worse for faces relative to objects. Their conclusion was that there exist face-specific neural mechanisms and that prosopagnosic impairment “cannot be accounted for as an impairment of within-category discrimination” (Farah, Levinson, & Klein, 1995, p.673).

Farah et al.’s conclusion seemed at odds with the conclusions that Gauthier and I had reached on the basis of our behavioral and neuroimaging studies. Two open issues led us to question whether there might be more to prosopagnosic subjects’ deficits than face-specific impairment. First, although discriminating between chairs or eyeglasses implicated subordinate-level recognition, there may have been local distinct features within both classes that allowed the prosopagnosic subject to perform better with common objects relative to faces. In contrast, control subjects with normal recognition processes would not need to rely on such local feature-based strategies and would show good recognition regardless of stimulus class. Second, we felt that it was important to consider prosopagnosic subjects’ performance using measures other than percent correct, for example response time or a bias-free measure such as sensitivity (which takes into account
the ratio between the number of times that the subject responds correctly to a “yes” trial, called a “hit,” and the number of times that the subject responds incorrectly to a “no” trial, called a “false alarm”). This latter point is critical in that prosopagnosic subjects may expend more effort attempting to identify objects as compared to faces. They also may believe that they are poorer at face recognition relative to common object recognition (e.g., a response bias that would change their accuracy for faces as opposed to other object categories, but would not affect their discrimination sensitivity for either faces or objects) – in our interactions with specific prosopagnosic subjects we had in fact noted that they do show stronger response biases than uninjured subjects, as well as speed-accuracy trade-offs.

In collaboration with Marlene Behrmann, we embarked on a series of experiments that systematically varied the level of categorization for common objects, Greebles, snowflakes (a highly homogeneous category), and faces (Gauthier, Behrmann, & Tarr, 1999a). In each condition we ran identical experiments with uninjured controls and at least two prosopagnosic subjects and recorded both response times and sensitivity. Two results stand out. First, we found that the apparent disproportionate impairment for faces as compared to non-face objects could be replicated if we looked only at percent correct, but that response times showed a trade-off in that the prosopagnosic subjects took much longer to recognize the non-face objects relative to faces. Second, we observed that if we controlled the amount of time subjects could view stimuli from each class, the same prosopagnosic subjects revealed similar impairments, as measured by sensitivity, for recognizing both faces and non-face objects. Important to our hypothesis, when sensitivity for faces and non-face objects was equated, it became obvious that the prosopagnosic subjects’ deficit became progressively more pronounced at the more specific levels of recognition regardless of the object category. To summarize our results, we have obtained evidence that independent of object category, our two prosopagnosic subjects are far more sensitive to the manipulation of the level of categorization as compared to our control
subjects. Thus, apparent face recognition deficits may be better explained as deficits in recognizing objects at more specific levels of discrimination.

Taken together, our behavioral, imaging, and neuropsychological work serves to implicate both the level of categorization and the level of perceptual expertise as important factors in visual recognition tasks. Indeed, it is our hypothesis that the interaction of these two factors is sufficient to explain the impressive specialization of face recognition mechanisms in visual cortex. Future studies will continue to investigate this issue, for example using far more sophisticated classes of novel objects (Figures 1f, 1g, and 1h are examples of our newest creations: “Fribbles” “YUFOs” and “Pumpkins” – these categories have properties that make them less “face-like” – for instance, YUFOs do not have the 2-1-1 part structure of faces and Pumpkins are asymmetric across all possible axes). If we also consider object geometry we have the foundations for forming a complete picture of recognition competence. Each of these aspects of recognition varies along a continuum that cannot be explained by simple dissociations between cognitive or neural systems. More likely is that we need to consider the interaction of all three factors and how a single recognition system can be flexible enough to adapt to the wide range of recognition contexts that we encounter everyday. It is this problem that we turn to next.

IMPlications FOR MODELS OF RECOGNITION

My research to date has focused on elucidating the factors that are critical to understanding the cognitive and neurological bases of human object recognition. It is my contention that the results we have obtained over the past several years implicate a single highly plastic visual recognition system. The challenge over the coming years is to develop computational models that can account for these remarkable abilities. In a first attempt to simulate some of the complexity of human recognition competence, as part of his Ph.D. dissertation, Pepper Williams (1997) developed a neural-network model for shape recognition. His goal was to develop a model that could recognize a set of novel, complex
multi-part objects. Williams found that his relatively simple network, dubbed “WHOA,” (now called “TEA”; see http://www.tarrlab.org/tea) was able to replicate a wide range of behavioral effects, including the generalization from known instances of an object category to new instances. Importantly, in addition to categorical knowledge, WHOA still showed item-specific sensitivity, thereby providing some evidence that a single system may be sufficient for seemingly disparate recognition tasks. Williams was also able to apply the WHOA, with no modifications, to the problem of learning Greebles. We found that the model did surprisingly well at simulating the onset of perceptual expertise and was again able to account for both generalization across category members and individuation within categories (Gauthier et al., 1998).

**Insert Figure 9 Here**

Although a single architecture may be sufficient for spanning a range of categorical levels, it is still likely that different elements of object representations may play different roles in recognition. A working hypothesis is that different spatial scales (analogous to blurry vs. sharp images – Figure 9) are differentially weighted at different levels of categorization. The essential idea is that complete images may provide too much information given that the problem of basic-level classification is to map many instances of a class onto a single category. There are reasons to believe that blurry, high-contrast images very similar to object silhouettes provide a description that is relatively stable over members of a perceptual class. Reasons include the fact that the WHOA model is one of several computational models (going back to Blum’s classic 1967 paper) that have successfully used silhouettes to perform object classification and the fact that there are known to be separable neural pathways for transmitting blurred high-contrast visual information and detailed lower-contrast information.

To investigate the role of silhouettes in recognition, we developed two lines of research: one in which we asked whether silhouettes contain enough information about objects to support visual recognition and one in which we examined whether there is sufficient
information in object silhouettes to separate object categories. As part of his Ph.D. dissertation, William Hayward (1998) compared the recognition of common objects across rotations in depth in a sequential-matching tasks. After viewing an intact image of an object, the same object was presented again at either a 60° rotation that showed similar surfaces and parts to the original, but a quite different silhouette, or a 180° rotation that showed quite different surfaces and parts, but a silhouette that was a mirror-reflection of the original's silhouette. We found that subjects were actually faster and more accurate to recognize objects as being the same given the similar silhouettes as compared to similar surfaces or parts. This rather surprising result, as well as other data obtained using somewhat different recognition tasks (e.g., see Peterson, 1994), indicates that silhouettes appear able to mediate some aspects of visual recognition (although we have ample evidence that silhouettes are not the only type of information represented).

To get more directly at the question of what role silhouettes might play in recognition, Florin Cutzu and I developed several simple methods for measuring silhouette similarity – these ranged from computing point-by-point correspondences along the boundary of the silhouette to measuring the area of overlap between silhouettes. We conjectured that simply by clustering object instances based on silhouette similarity we would be able to separate most instances of one object class from instances of a second object class. To provide a strong test of our model we used silhouettes of cats and dogs – two highly similar categories. Importantly, exactly these stimuli had been used in a study in which it was demonstrated that human infants were capable of perceptually differentiating between these basic-level classes (Quinn & Eimas, 1994). Thus, we knew that babies were able to able to tell the cats from the dogs based purely on visual information – could our model do the same? Our results were quite clear – despite the high degree of similarity between cats and dogs (telling cars from chairs would not have been much of a challenge) we found that our relatively simple measure of silhouette similarity provided good separation between
object classes (Cutzu & Tarr, 1997). Thus, we have some evidence that silhouette-like information is sufficient for basic-level categorization.

The possibility that silhouette-like information forms an important level of visual representation leads to an intriguing conjecture – limiting early visual input to silhouette-like information may be essential for forming stable perceptual categories. It is known that human infants are born somewhat myopic and prefer high-contrast images. Other than being an accident of development, is there any potential benefit to this state? If a developing visual system were to receive complete, fully-detailed images, each object would appear quite distinct from all previously-seen objects and a many-to-one mapping might not arise. In contrast, if the visual system initially receives only coarse information then perceptual “bins” corresponding to categories may emerge (Figure 9). As the visual system develops, increasingly finer information will become available, thereby allowing in-place coarse categories to be refined into subclasses and specific instances for within-category recognition tasks. Thus, while silhouette-like information may mediate the recognition of perceptual categories in adults, it may be even more important for acquiring such categories in the first place. In the study of language it has been suggested that “starting small” (Elman, 1993) in terms of memory span is essential for learning syntactic categories. Similarly, “starting blurry” may be essential for learning visual categories.

Some evidence for “starting blurry” has been gathered by Quinn, Eimas, and Tarr (2001). We found that 3- to 4-month-old infants could perceptually categorize cats and dogs given only their silhouettes. Interestingly, our experiments also revealed that infants were better at categorically separating the two species when using silhouette information from the heads as compared to the bodies of the animals. What is not yet known is whether this advantage is a consequence of intrinsically more information in the heads or a pre-wired preference to attend to heads and faces (see Johnson & Morton, 1991). Regardless, our results indicate that general shape or external contour information that is centered about the
head is sufficient for young infants to form individuated perceptual categories of cats and dogs.

As already pointed out, silhouette-like information comprises only one component of object representations. To support more specific levels of recognition, finer surface details and variations in shape must be considered (Tarr & Bülthoff, 1998). There are a number of recent models of recognition that have proposed object representations based on large collections of viewpoint-dependent local features. Indeed, there has been a remarkable trend towards the local view-dependent feature approach by several independent research groups (e.g., Edelman, 1995; Fukushima, 2000; Lowe, 2000; Riesenhuber & Poggio, 1999; Ullman & Sali, 2000). By allowing the repertoire of features to be quite broad, including local surface patches, local measures of color and brightness, oriented edges, and contour configurations, many of the details that are necessary for within-category recognition may be captured. At the same time the feature set may include more global regions that are sensitive only to high-contrast boundaries, thereby capturing the silhouette-like level of information. Thus, a single view of an object might include thousands of features at multiple spatial scales. Categorical recognition could be mediated by measuring similarity across the coarse levels of information – essentially adjusting the threshold for what counts as a match. For example, votes might be tallied across views of all known objects with similar silhouettes. More specific levels of categorization could be mediated by measuring similarity across finer and finer levels of information – increasing the threshold. As the information required for a given discrimination becomes more and more specific to a particular object, the number of features that will vote will become progressively more narrow, thereby implicating fewer and fewer known objects.

Representing objects as collections of viewpoint-dependent features leads to several fundamental questions. First, individual features are rarely distinctive enough to uniquely specify a single object or class – it is only the configuration of features that allows effective recognition. How then are local features within the representation related to one another?
One straightforward answer is the object representation system is sensitive to the spatial co-occurrence of individual features. The more frequently any set of features are seen together at specific spatial positions, the more tightly they will be linked to one another (a form of what is known as “Hebbian learning”). Object representations will be comprised of features whose spatial positions are more or less strongly related to one another – features that co-occur quite often in a given configuration will become strongly interdependent with the presence of a subset of the features activating the entire ensemble. For example, the surfaces found on a single part of an object will appear together quite frequently and will become strongly associated. In contrast, features that do not co-occur very often will be connected only weakly or not at all (Wallis & Rolls, 1997). For example, the surfaces found on different parts of an articulated object, will not appear together nearly as often as surfaces on the same part, and thus, will be only weakly associated. This simple statistical learning mechanism may provide an explanation for the configural sensitivity found in cases of perceptual expertise, including face recognition (Gauthier & Tarr, 1997b; Rhodes, Brake, & Taylor, 1989). The acquisition of expertise is marked by extensive practice differentiating similar instances from a single class – consequently many class-level features will co-occur in the same configuration with great frequency, for example, the eyes, nose, and mouth of human faces. Such oft-seen features will become tightly interdependent as the system is fine-tuned by experience. Thus, relocating the position of one such feature will impact the recognition of the other features – much as has been found for parts of human faces and for parts of Greebles when recognized by experts.

Second, if different views of objects are composed of different sets of features, how are different views within the representation related to one another? Interestingly a principle similar to that used to link features may be used to link different views of an object. The single most likely image to occur following a view of an object is another view of that same object. Therefore, a recognition system that is sensitive to the temporal co-occurrence of sets of features would learn, over experience, to link those views that arise from single
objects – thereby forming a coherent, organized multiple-views representation (Figure 2). Neurophysiological evidence suggests that neurons in visual cortex are indeed sensitive to co-occurrence over time – remarkably even in those cases where there is no geometric similarity between the views that become associated (Miyashita, 1988; Wallis & Baddeley, 1997; see the Bülthoff & Bülthoff chapter in this volume for additional discussion of this issue).

CONCLUSIONS

A relatively simple experiment exploring whether visual recognition is based on viewpoint-dependent or viewpoint-independent information has led to an extensive research program employing psychophysical and neuropsychological methods. At the core of this program has been the idea that there is a complex interaction between three aspects of recognition: the appearance of images of objects as they vary in the environment; the level of categorical specificity required for a given task; and the degree of experience the perceiver has with specific object classes. In my laboratory we have investigated each of these issues, asking whether a single visual recognition system is sufficient to account for the continuum of behaviors seen in each case. Converging evidence provides a preliminary “yes” to this question. View-based, local-feature representations can account for the recognition performance observed across changes in viewpoint in subordinate-level recognition tasks. Similar viewpoint-dependent mechanisms appear capable of supporting basic-level recognition tasks and the recognition of new instances of familiar categories. Finally, the complete range of recognition tasks, including face recognition, can be accounted for by considering the degree of perceptual expertise. Thus, humans appear to have a single highly adaptable visual recognition system that can be fine-tuned by experience to support a spectrum of recognition behaviors (for a similar view see Schyns, 1998). Although there is, as always, much work to be done, we have begun to illuminate some of the properties of this remarkable system.
ACKNOWLEDGEMENTS

Many more details about this work can be obtained at http://www.tarlab.org Much of the research presented in this chapter was supported by NSF Award SBR-9615819. Thanks to Isabel Gauthier, William Hayward, and Pepper Williams for not only collaborating on much of the research, but providing valuable feedback on several drafts. Thanks also to my other collaborators over the years: Marlene Behrmann, Heinrich Bülthoff, and Steven Pinker.
REFERENCES


Peterson, M. A. (1994). Shape recognition can and does occur before figure-ground organization. *Current Directions in Psychological Science, 3*, 105-111.


FOOTNOTES

1 Biederman’s (1987) part-based model does not predict complete viewpoint invariance. Rather, he hypothesizes that recognition is viewpoint-invariant only so long as the same part-based description may be recovered. Thus, self-occluding objects for which rotations in depth alter which parts are visible and occluded will require separate part descriptions for each unique configuration of parts and will be recognized in a viewpoint-invariant manner only when the same parts remain visible. The stimulus images used in both Tarr et al. (1997) and Hayward and Tarr (1997) always showed the same configurations of parts and therefore Biederman’s theory would predict complete viewpoint invariance for the conditions tested.

2 Note that parts-based models only make this prediction under viewing conditions in which the part description is equally recoverable from both views. The same is true for predictions of viewpoint invariance across views of more complex objects. In all of the studies reported here attempts were made to ensure that every view clearly showed the viewpoint-invariant features that are putatively used to recover 3D parts in Biederman’s (1987) model. Thus, if this model were correct, all views should have been recognized equally well, which was not the case. That being said, it is difficult to definitively state that two views are equally “good” in terms of part recovery. However, if small gradations in pose across non-accidental views lead to differences in performance, then the recognition process is viewpoint dependent and inconsistent with extant viewpoint-invariant part-based models.
There is one published case in which brain injury produced the opposite deficit: intact face recognition with impaired object recognition (Moscovitch, Winocur, & Behrmann, 1997). Although such “double dissociations” are often held up as the strongest evidence in favor of separable systems, it is possible to obtain a double dissociation within a single system (Plaut, 1995). Moreover, our claims regarding the lack of face-specific processing should not be taken as an argument that brain areas preferential for faces do not exist. The reasons why such areas exist are simply explainable by factors other than faces as a “special” object class. However, given the existence of such areas, we should be able to render them preferential for other domains of expertise – a result we have found in several studies (Gauthier, 1999b, 2000a). Finally, it is also worth noting that the patient studied in Moscovitch et al. (1997) was unable to recognize Greebles – something he should be able to do if separable mechanisms for face (and other expert) recognition exist and are intact in his case (Behrmann, Gauthier, & Tarr, unpublished data).
**Figure Captions**

Figure 1. The evolution of the novel stimulus in our lab.

Figure 2. A possible multiple-views representation of a motorcycle – 3D objects are represented as a set of viewpoint-specific models.

Figure 3. A 180° “flip-in-depth” is always the shortest path to align a mirror-reflected version of a 2D shape with its standard in a recognition task. Apparently the human visual system (unconsciously at least) knows this too. Adapted from Tarr and Pinker (1989).

Figure 4. The novel 3D objects we created to assess the kinds of features used in viewpoint-dependent recognition. Objects are examples from the: a) five-tube set; b) single distinctive part set; c) three distinctive part set; and d) five distinctive part set. Adapted from Tarr, Bülthoff, Zabinski, and Blanz (1997).

Figure 5. Equal magnitude rotations for a single part that produce different local image features. The 30° rotation from center to left yields a qualitative shift from a cusp with a tangent line into an arrow vertex; the 30° rotation from center to right yields only a distortion of the cusp with a tangent line. Human observers are more sensitive to qualitative shifts in local features as compared to simple distortions in local features. Adapted from Hayward and Tarr (1997).

Figure 6. Examples of Greebles. Four Greebles from one family, two of each gender, are shown in the left-hand box. To the right a Greeble from a different family is shown for each gender.

Figure 7. The regions traced by the ovoids approximate those brain regions typically associated with face recognition based on earlier neuroimaging studies (Kanwisher et al., 1997; Puce et al., 1995). The fMRI maps show the activation we obtained for the subordinate-level recognition of common objects over and above basic-level recognition. The darker “splotches” represent positive activation for this comparison overlaid on structural maps of the human brain. Progressively greater positive activation...
is depicted by progressively brighter areas within the dark regions. See Gauthier et al. (2000b) for a discussion of how these activation maps correspond to the putatively face-selective regions of visual cortex and an interpretation of the additional regions of activation seen in the right panel.

Figure 8. Brain activation in response to faces and Greebles for three Greeble novices and three Greeble experts. The left panels show the brain activation we obtained using fMRI for three novices when the passive viewing of common objects was subtracted from the passive viewing of faces or Greebles. The right panels show the activation for three experts in the same tasks. Only the voxels showing more activation for faces or Greebles than objects are shown (darker regions with higher activation depicted as brighter areas within these regions). The dashed-line squares denote the middle fusiform gyrus bilaterally (functionally defined) and the lateral occipital gyrus foci for one expert (bottom right). Adapted from Gauthier et al. (1999b).

Figure 9. Descriptions including surface detail, subtle changes in brightness, and fine shape differences may be necessary for making within-category discriminations. In contrast, descriptions including only high-contrast blurred information – similar to silhouettes – may support basic-level categorization. This figure illustrates the point that given complete images, instances of a given category are still quite different from one another, but given silhouette-like images, such instances become more similar. Interestingly, human infants receive visual input that is closer to the bottom row, suggesting that they may begin life by learning coarser visual categories and then move towards differentiating instances within these categories only during the later stages of development.
Figure 1
Figure 2
Visual Object Recognition

Figure 3
Figure 4
Figure 9