Encyclopedia of Psychology
In press. A. E. Kazdin (Ed.), Washington, DC: American Psychological Association

Visual Pattern Recognition

Michael J. Tarr
Brown University

Perceiving the visual world around us is one of the most basic acts of everyday experience. Although pattern recognition may seem effortless, it is actually a complex problem--so much so that the visual areas responsible for this process occupy up to one-half of our cortex. Fundamental to our perception is the transformation of the light array that falls on our retinae into coherent surfaces and objects. How this is done is still a matter of some debate, but results from psychophysical, neuropsychological, and physiological studies point towards a remarkably adaptive system that supports a wide range of recognition tasks.

THE FLEXIBILITY OF HUMAN RECOGNITION

One of the hallmarks of the human pattern recognition system is its extreme flexibility. As observers we are able to identify objects under a wide array of conditions that confound even the most powerful computer vision systems. For example, we can recognize objects at many different categorical levels. Often, however, it is assumed that objects are first identified at the *entry level* (Jolicoeur, Gluck, & Kosslyn, 1984)--defined as the name which is generated or matched most rapidly to a given object, e.g., "apple" or "bird" (although less typical instances may be identified at a more specific level, e.g., "penguins").

While entry-level recognition is certainly an important element of everyday recognition, it is not the only level at which objects are recognized. We often identify objects at a more specific level, sometimes referred to as the *subordinate level*, e.g., a "McIntosh apple" or a "white-breasted nuthatch." Such finer discriminations require additional perceptual analysis and thus typically take longer than entry-level judgments. Beyond subordinate-level recognition, we also recognize objects at the *individual*- or *exemplar*- level, e.g., "the McIntosh apple I brought for lunch." Making such judgments requires specific information about a given individual object and consequently even greater perceptual analysis.

The fact that we are capable of performing any of these tasks with remarkable precision indicates that visual recognition is best thought of as a continuum ranging from rather coarse to incredibly fine perceptual discriminations (Tarr & Bülthoff, 1995). However, as pointed out by Marr and Nishihara (1978), there is a tradeoff between the information that captures the more general and less variable properties of objects and the information that captures the finer and more variable details of objects. Thus it is not clear that a single recognition system can support categorization at many different levels.

PATTERN PERCEPTION

Before objects can be identified at any level, the contours and surfaces defining them must be grouped into coherent wholes or patterns. Remember that what arrives at our eyes is an undifferentiated array of light intensities, but what we perceive are surfaces and objects. At the most elementary level it is clear that our visual system performs a *feature analysis* at every location in the scene. What this means is that the optic array is transformed into a description of local visual properties, for example, the orientation of edges or the color over restricted regions. This recoding provides important information about what is happening in the scene in terms of visual properties that we "care" about, but does not tell us how to put such features together to form more complex percepts. Thus, it is typically assumed that, following feature detection, principles of *perceptual organization* are used to combine local features into more global structures, for example, extended contours or textured 3D surfaces. Principles such as *similarity* (similar features are grouped together) or *good continuation* (features that form a straight line or a smooth curve are grouped together) were originally elucidated by the *Gestalt school of psychology* during the early 1900's and, remarkably, they are still considered fundamental today.

Beyond such simple principles, there are many other processes that appear to contribute to the perception of complex patterns (for a good overview of many aspects of pattern perception, see the readings in Rock, 1990). Two of the best known are *structure-from-motion* and *shape-from-shading*. In the former case, 3D surfaces can be perceived because local 2D motions can be integrated according to the principle that they must all arise from a single moving rigid object whose surface orientations only change gradually. Likewise, by attending to the change in shading (or texture for that matter) over a surface and again assuming only gradually shifts in orientation, we can perceive the 3D shape of an object. Even these principles, however, are insufficient to specify completely the complex nature of a typical scene. What is ultimately necessary is that we separate each object from both other objects and the

background--a process known as *figure-ground* segregation. Presumably we rely on cues such as discontinuities in color, texture, or shape, but the precise mechanisms for accomplishing figure-ground are still poorly understood--indeed, this is one of the reasons why computer vision systems are so bad at object recognition.

RECOGNIZING OBJECTS IN A CHANGING WORLD

Compounding the already difficult task of segmenting individual objects out of a complex scene is the variability we encounter in viewing conditions at different moments in time. The recognition system must contend with images of objects that vary with changes in almost any viewing parameter, including occlusion, illumination, orientation, 3D viewpoint, position, size, or configuration (Figure 1). Almost any source of variability may affect recognition performance with recent evidence suggesting that the degree to which a change impairs recognition depends on the categorical level of the recognition judgment. For example, increasing the similarity between the actual target object and other potential target objects (as in increasingly subordinate-level tasks) typically increases recognition costs across changes in viewpoint (Tarr, 1995). As discussed in the following section, the bulk of object recognition research has focused on variability due to changes in viewpoint--presumably because rotating an object in depth produces such dramatic changes in the image.

Figure 1. Examples of some of the variability the visual recognition system must overcome. From left to right we can still recognize the fan despite: partial occlusion, a change in illumination, a change in viewpoint, and a change in configuration.

CURRENT THEORIES OF OBJECT RECOGNITION

How the visual system compensates for variation in the image forms the core of almost all current theories of object recognition. As a starting point most theories assume either relatively *image-invariant* or relatively *image-dependent* representations. Viewpoint is often taken to be the diagnostic case and theories typically predict either small and discrete performance costs across changes in viewpoint (Biederman, 1987) or large and continuous performance costs across changes in viewpoint (Tarr, 1995). One might think that because we are able to recognize known objects quite well from almost any viewing position that object representations must be viewpoint invariant. In fact, since we are already familiar with real-world objects from many different viewpoints, it is just as likely that we have learned multiple viewpoint-specific representations for each known object or class.

To investigate these alternatives, Tarr and Pinker (1989) taught observers to name novel objects from a single orientation. When they tested generalization to new picture-plane orientations, Tarr and Pinker found that observers were fastest at the familiar orientation and progressively slower at unfamiliar orientations further and further from the trained orientation. With practice, however, observers became equally fast at all known orientations. Tarr and Pinker hypothesized that this learning was analogous to the apparent viewpoint independence exhibited for known objects--invariance obtained by virtue of multiple views. This hypothesis was tested by introducing additional new orientations for the now-familiar objects. While observers continued to show equivalent performance for all familiar orientations, they again took longer to recognize the objects in unfamiliar orientations--now, however, with performance dependent on the *nearest familiar* orientation. Similar results have also been obtained for novel 3D objects rotated in depth (Tarr, 1995).

Converging evidence for multiple image-dependent representations--often referred to as *view-based* models--comes from physiological research. For example, Logothetis, Pauls, & Poggio (1995) trained monkeys to recognize novel objects from several different 3D viewpoints. With practice the monkeys, like humans, became equally good at recognizing the objects from any of the training viewpoints. When Logothetis et al. recorded the responses of cells in the inferior temporal cortex (IT) of the monkeys they found that many cells responded selectively to a previously novel object and, crucially, maximally to a single viewpoint that had been shown during training. As with recognition performance in humans there was a gradual decrease in a given cell's response as the preferred object was rotated away from the familiar viewpoint. Thus, an ensemble of neurons, each tuned to a different viewpoint, may represent a 3D object.

While such results are intriguing, many aspects of view-based models are underspecified. For instance, there is as yet no clear definition of what features are used to represent each view of an object. Although many theorists have used simplified features (such as vertices specified in linear image coordinates), they are quick to point out that view-based models are unlikely to rely on such features. In particular, features based on

spatial coordinates such as pixels are highly unstable; rather it is presumed that higher-order features such as surface patches, edge features, or bounding contours are used, albeit in a viewpoint-specific manner.

The best known alternative to view-based models are *structural-description* models that typically assume image-invariant representations (Marr & Nishihara, 1978; Biederman, 1987). The fundamental assumption of these models is that objects are represented in terms of features that are stable over changes in the image, for example, parts described as 3D volumes. A second assumption is that configurations of parts are described relative to one another rather than relative to the observer or the world. Marr and Nishihara (1978) assumed that observers could use the major axes of an object to recover the shape of almost any part--because the parts were 3D volumes described in an viewpoint-independent manner, once abstracted away from the image, a description of a given object was identical regardless of the viewing position. More recently, Biederman (1987) has proposed a related scheme in which there is only a limited set (~30) of qualitatively-defined 3D volumes, e.g., "brick" or "cone." While a restricted set of parts may make object representation more tractable in a computational sense, it limits the theory to entry-level recognition (since many variations in fine structure are mapped into a single volume). More fundamentally, both of these schemes, as well as other models based on 3D volumes, have been dogged by the question of how to recover descriptions of parts from images--at present there is no workable solution to this problem.

One element that appears to be missing from both view-based and structural-description models is how to account for the flexibility of recognition across different categorical levels. It has often been claimed that structural descriptions are best suited to entry-level categorization, while view-based models are best suited to subordinate-level or individual recognition. Such a hypothesis is not entirely satisfactory if recognition is to be thought of as continuum of different levels of access. Where does one draw the boundary between one process and the other? Moreover, while it is known that damage to parts of the visual system can result in *object agnosia*--an inability to visually identify certain types of objects--there is little evidence to suggest that agnosic subjects simply loose the ability to recognize objects at either the entry-level or the subordinate-level (the pattern expected if one of the two systems was completely removed). Rather agnosics seem to show a more complex pattern of sparing and loss as if selective deficits occur according to the types of processing subsystems that are impaired in the individual (although, as reviewed below, there are a number of cases in which subjects apparently lose the ability to recognize human faces--raising the possibility of class-specific recognition systems). Thus, there is not much support for separable recognition systems for different levels of object identification. More likely is that a single system can be fined-tuned in response to perceptual experience, thereby mediating multiple categorical levels. Indeed, theorists have begun to consider the possibility of a single recognition system that can support both coarse categorical judgments and finer discriminations, adaptively selecting the most appropriate features according to the task at hand.

FACE RECOGNITION

Although the notion of a unified recognition system is appealing, there are certain phenomena that point towards specialized recognition mechanisms. One of the most notable examples is the case of face recognition, where brain-injured patients, brain imaging studies, and behavioral results all appear to indicate a face-specific recognition system. Perhaps the most compelling piece of evidence is the phenomenon of *prosopagnosia*--a syndrome in which brain injury to visual cortex results in a profound inability to recognize individual faces (see Farah, 1992). While it has been argued that prosopagnosic subjects are more impaired at recognizing faces relative to other objects, it is possible that face recognition occurs at a more subordinate level as compared to common object recognition. Indeed, almost all prosopagnosics have some difficulties recognizing non-face objects.

A second piece of evidence for face-specific processing is the finding using functional magnetic resonance imaging (fMRI) that certain areas of IT are more active for face recognition as compared to common object recognition (Sergent, Ohta, & MacDonald, 1992). As mentioned, faces and common objects, however, are generally recognized at different categorical levels--faces at the individual level and common objects at the class level. What happens when common objects are also recognized at a more specific level? Using fMRI Gauthier et al. (1997) observed that the same areas of IT found to be more active for face recognition are also more active for subordinate-level recognition of common objects as compared to entry-level recognition of the same objects. Therefore, this area of IT is more plausibly involved in finer levels of recognition, regardless of stimulus class, rather than in face recognition *per se*.

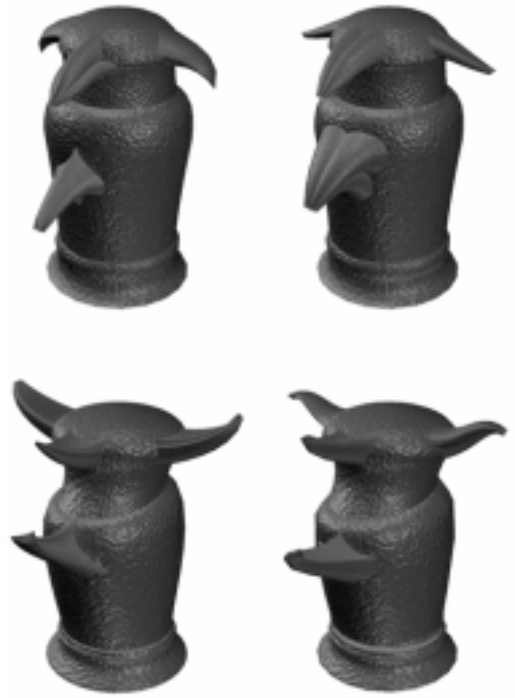

Figure 2. Examples of Greeble objects used to study perceptual expertise. Four Greebles from one family are shown, the top two being of one gender and the bottom two being of the other gender.

Finally, there are many studies that have reported "face-specific" behavioral effects. One of the most interesting is that of "holistic" processing for faces. Tanaka and Farah (see summary in Farah, 1992) found that observers were poorer at recognizing part of a trained face, e.g., "Bob's nose," if other parts of the face, e.g., the eyes, were transformed from the original configuration. This configural sensitivity is surprising in that the recognition of an individual part would seem to be independent of other features. In contrast, identifying individual parts of trained houses did not produce this effect, suggesting face specificity. On the other hand, observers are almost always perceptual experts at individual face recognition, but rarely so for other classes of objects (exceptions being birdwatchers and the like). To test whether perceptual expertise rather than the stimulus class produces configural sensitivity, Gauthier and Tarr (1997) created a novel class of objects--"Greebles" (Figure 2). Observers unfamiliar with Greebles did not show configural sensitivity. Greeble experts (created through 10 hours of training), however, showed configural sensitivity in identifying individual Greeble parts. Thus, factors such as the degree of perceptual expertise, rather than the stimulus class, are apparently responsible for what was previously thought to be face-specific processing.

BIBLIOGRAPHY

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

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*, 164-169.

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

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

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

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.

Rock, I. (Ed.). (1990). *The Perceptual World*. New York, NY: W. H. Freeman and Company.

Sergent, J., Ohta, S., & MacDonald, B. (1992). Functional neuroanatomy of face and object processing: A positron emission tomography study. *Brain*, *115*, 15-36.

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*, 55-82.

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*, 1494-1505.

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