REPRESENTATION AND RECOGNITION IN VISION.
By Shimon Edelman.

As we move about the world and view objects from different distances and directions, the images projected onto the retina are correspondingly transformed—by translation, magnification, rotation and so on. How, despite rarely experiencing the same image twice, do we recognize an object as being the same? This problem has had a long and varied history involving disparate disciplines: philosophy and psychology; computer science, mathematics, and engineering; anatomy and physiology; and, to a limited extent, neurology. There is still no universally accepted solution. In broad terms, explanatory theories can be divided into two kinds: those that concentrate on the nature of the internal model or representation that the visual system forms of an object in the external world, and those that concentrate on how that representation alters with changes in the pose of the object or changes in observer viewpoint. The two approaches involve a trade-off. Thus, if the representation is assumed to alter with a change in viewpoint, then the visual system needs to apply an internal restoring, interpolating, or normalizing transformation for recognition to occur; and the more work that is done in forming the representation (the more invariant it is to viewpoint changes), the less work that needs to be done in normalizing it. In practice, factors such as the need to retain information about viewpoint changes and whether objects are to be compared simultaneously or sequentially influence the extent of the trade-off.

One of the first significant theoretical analyses of visual (and auditory) representation and recognition was undertaken in the 1940s by Walter Pitts and Warren McCulloch, who showed how a biological system could calculate constant properties—invariants—of entire images under changes in viewpoint. It was not at all clear, however, whether these global invariants could, or should, be computed; moreover, the required biological machinery seemed demanding, and the representations assumed to be produced took little account of the structure of the images presented to the organism.

In the decades that followed, a more object-centred approach to the problem of extracting invariants developed. The idea was to associate with each image a ‘structural description’, specifying the configuration of an object’s components or local features (e.g. lines and edges) in terms of discrete or qualitative relations between those features (e.g. ‘joined to’, ‘above’, ‘left of’). Such a description was naturally invariant to translations and magnifications of the image on the retina, and, depending on the choice of relations, to its rotations.

A particular structural-description theory of recognition, called ‘recognition-by-components’, has been advocated by Irving Biederman. Here, the object components are generalized cones, derived from contrasts of five edge cues in a two-dimensional image: curvature, collinearity, symmetry, parallelism and co-termination. Visual detection of these elementary properties is assumed to be invariant over viewing position. Experiments on picture priming, where speed and accuracy of recognition of a briefly presented picture is facilitated by prior presentation, suggested that priming depends on activation of these components. Changes in position, size, orientation-in-depth, or the particular lines and vertices present in the image did not affect the magnitude of the priming, as long as the same generalized-cone components were assumed to be activated.

Descriptions of the structure of images based on object parts are appealing because objects are often only part-seen, for example, from the front only, or occluded by other objects. Such descriptions naturally represent non-rigid or articulated objects, for example, a body or hand; and, phenomenologically, they accommodate our experience that objects, such as a cup or a face, are actually perceived as comprising parts. Nevertheless, there has been some debate
in the literature about how well object parts can be extracted from images and how well such descriptions account for data on the recognition of novel and familiar objects that have been rotated in depth.

An alternative to describing the structure of images in terms of combinations of generic primitive shapes is to describe them in terms of memory records of similar images. One starting point for this theory is that it is unreasonable for a visual system to treat all possible shapes within the same parametric framework: whereas similar shapes may need to be compared in detail, very different shapes need merely to be distinguished from each other. How, then, should similar images be described? For a given class of similar shapes, it is known from statistical pattern-recognition theory that a representation can be obtained in a common space spanned by a small set of reference shapes, generally referred to as basis functions. One way of generating basis functions is to apply the statistical technique of principal-components analysis, which has been used successfully to explain how the visual system approximates a solution to the theoretically simpler problem of surface colour constancy, that is, the constant colour appearance of an object’s surface independent of the illuminant on the scene. Another way of generating basis functions from a set of data is to expose an artificial neural network to prototypical examples drawn from that set.

In Representation and Recognition in Vision, Shimon Edelman sets out the case for such a network-based scheme called Chorus. In this scheme, vectors of proximities to a small number of prototypes are used to span the representation space. The prototypes are derived from recurring stable patterns of primitive features, which, by definition, correspond to frequently observed objects. Each prototype is represented by an interpolation mechanism similar to a basis-function neural network tuned to several of the object’s views, and is learned from examples.

The elementary building block of Chorus—a unit tuned to a specific view of a specific object—is inherently viewpoint-dependent, but, because the tuning of the individual units is wide (that is, they also respond, usually less well, to views that differ from the optimal), it is possible for a collection of such units to interpolate the entire space of views for a given object. Viewpoint invariance achieved in this way will be approximate and specific for the given object and the degree of viewpoint invariance will decrease when the objects to be discriminated are similar to each other. A system of several object-specific modules of this kind should exhibit some viewpoint invariance for novel objects, by interpolating between the view spaces of the familiar ones.

Support for Chorus has come from experiments with synthesized images of abstract, paperclip-like objects and more natural animal-like objects. In one experiment, human subjects were trained to discriminate between computer images selected from classes of monkey-like and dog-like objects. After subjects reached 90% correct on a fixed canonical view of each object, discrimination performance was tested for novel views that differed by up to 60° from the training view. Despite differing only parametrically, that is, in degree rather than in kind, these objects were recognized virtually independently of viewpoint, providing that the two classes were sufficiently dissimilar. These and the results of other experimental tests were taken as evidence for a theory of recognition by views and against a theory of recognition by components.

Which of these two theories is the more appropriate description of human object recognition is unlikely to be resolved by appeals to neurobiology. For example, recordings in macaque temporal cortex have shown specificity for face view and gaze direction, and the extent of any invariance to viewing conditions is limited to about ±30° rotation in depth. These results are certainly consistent with Chorus, but, as Edelman points out, it is possible that the function of the underlying units could be irrelevant to the process of object recognition.

Views-based theories suffer from certain foundational problems, as Edelman readily acknowledges. Thus, a principled way needs to be found for defining the set of objects constituting the training set and for dealing with occlusion and interference between the images of neighbouring objects in a scene. There are also some outstanding experimental issues; for example, how to account for psychophysical data showing that shape recognition and discrimination vary systematically with the position of the image on the retina. More critically, there is persistent evidence that small changes in object shape can influence visual perception in a way that seems more naturally explained by components-based theory than by views-based theory.

In Representation and Recognition in Vision, Edelman argues strongly for a particular approach to modelling human object recognition, focusing arguments and evidence to that end. It is a stimulating and informative book, but better suited to the sophisticated reader, and rather less so to the beginner hoping to achieve a broad understanding of a sometimes contentious field.

David Foster
Department of Optometry and Neuroscience, UMIST, Manchester, UK