9

Object Recognition

An essential part of the behaviour of animals and people is their ability to *recognise* objects, animals, and people that are important to their survival. People are able to recognise large numbers of other people, the letters of the alphabet, familiar buildings, and so on. Animals may need to recognise landmarks, suitable prey, potential mates or predators, and to behave in the appropriate way to each category.

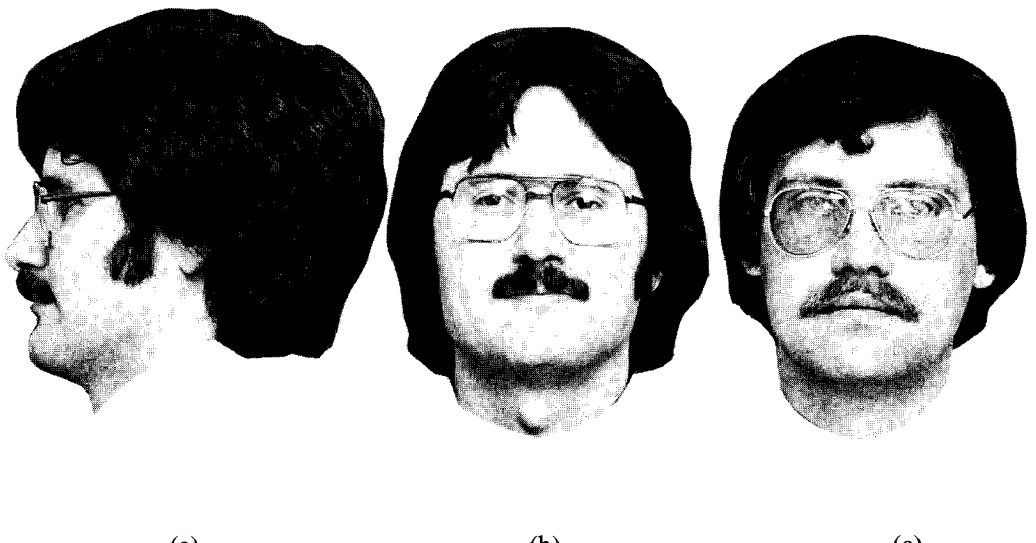
If we assume that the information available to a person or animal is a static two-dimensional image on the retina, a problem immediately arises in explaining visual recognition. Take the example of a person recognising letters of the alphabet: the problem is that an infinite number of possible retinal images can correspond to a particular letter, depending on how the letter is written, how large it is, the angle at which it is seen, and so on (Fig. 9.1). Yet somehow we recognise all these patterns of light as corresponding to the same letter. Or consider the problem of recognising a friend's face: the image of the face on the retina will depend on the lighting conditions and the distance, angle, and facial expression. Again, all these images are classified together, even though some (such as a full-face and a profile view) are quite dissimilar and more like the same views of different faces than they are like each other (Fig. 9.2).

These are both illustrations of the problem of *stimulus equivalence*: if the stimulus controlling behaviour is a pattern of light, or image, on the retina, then an infinite number of images are

FIGURE 9.1

All these different shapes are classified as the letter A.

FIGURE 9.2



(a) and (b) show two different views of the same person, Patrick Green. View (b) is in many ways more like picture (c), which is of a different person, than it is like view (a). Photographs by Sam Grainger.

equivalent in their effects, and different from other sets of images. Many influential treatments of object recognition assume that all the images corresponding to a particular thing, whether letter of the alphabet or face, have something in common. The problem is to find just what this is and how this thing in common is detected. It is this problem that we will be considering in this chapter.

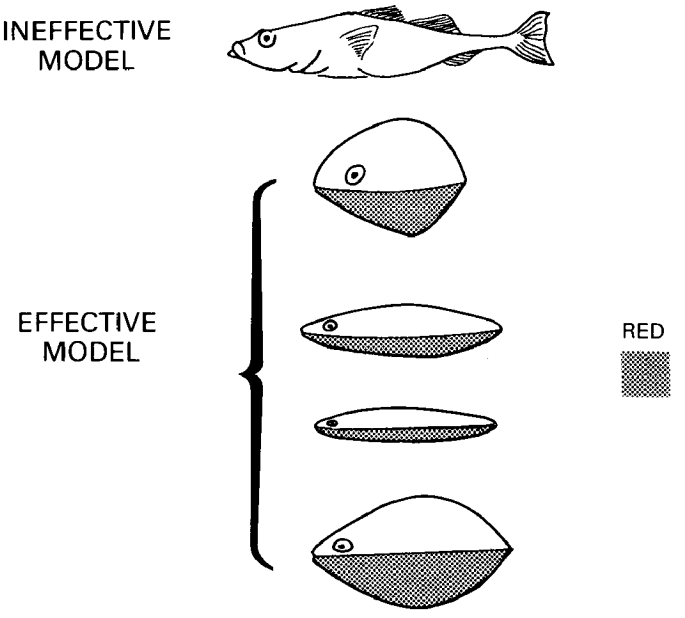
SIMPLE MECHANISMS OF RECOGNITION

Many animals, particularly simpler ones such as insects and fish, solve the stimulus equivalence problem by detecting something relatively simple that all images corresponding to a particular object have in common. A good example is the three-spined stickleback. Males of this species build nests and defend them against other males by performing threat displays. A stickleback must therefore be able to recognise rival males and discriminate them from other fish and from objects drifting by. The retinal images of rival males will obviously vary greatly, depending on the other fish's distance, angle, and posture, and it seems that classifying these images separately from those of other fish will need elaborate criteria.

In fact, as Tinbergen (1951) discovered, the stickleback manages successfully with quite simple mechanisms of recognition. Tinbergen observed the strength of sticklebacks' aggressive responses to a range of models and found that they would readily attack a crude model of another fish, *provided* it had the red belly colour characteristic of male sticklebacks. Indeed, a crude model with a red belly elicited more attack than an accurate one without (Fig. 9.3).

A feature of an object or animal—such as the red belly of a stickleback—that elicits a response from an animal is called a *key* or *sign* stimulus, and

FIGURE 9.3



An accurate model of a stickleback without a red belly (top) is less effective as a stimulus to elicit aggression from a male stickleback than any of the cruder models below. Adapted from Tinbergen (1951).

it greatly simplifies the problem of recognition. As long as red objects and fish with red markings are rare in the stickleback's environment, it can use the key stimulus to recognise rivals and does not need to use information about another fish's detailed structure and colouration.

The stickleback's recognition of a rival male does depend on more than just the presence of a patch of red of a certain size in the retinal image, as Tinbergen also found that a model with a red patch on its back was attacked less than one with an identical red patch on its belly, and that a model in the "head-down" posture of an aggressive fish was attacked more than one in a horizontal posture. Even so, the presence of this distinctive feature allows a much simpler means of recognition to be effective than would otherwise be the case.

Many other examples are known of key stimuli being important in the recognition by animals of other members of their species, and we will mention two other examples from Tinbergen's work. One is the recognition of female grayling butterflies by males. Tinbergen found that males would fly towards crude paper models moving overhead and that their response was not affected by the colour or shape of the model. The key stimulus turned out to be the pattern of movement of the model: males would fly towards it if it imitated the flickering and up-and-down movements of a butterfly, but not if it moved in a smooth way. Although butterflies do waste time chasing other males, or butterflies of the wrong species, this simple mechanism of recognition does prevent responses to other kinds of insect.

Another example is the recognition by nestling thrushes and blackbirds of their parents. When the parents bring food to the nest, the young birds turn towards them and gape, opening their mouths wide to be fed. Tinbergen found that gaping is elicited by a moving dark silhouette above the birds' eye level, of any shape and size. Presumably this simple mechanism of recognition is adequate because the chances of anything other than a parent resembling the key stimulus are low.

Key stimuli may also be important in the recognition of prey. Toads feed by snapping at

insects flying past them, capturing them with their long sticky tongues, and Ewert (1974) found that they recognise insects by fairly simple criteria, as they will snap at small cardboard squares. Although Ewert's experiments used moving targets, toads will also snap at stationary models (Roth & Wiggers, 1983). Although toads are selective for the size and speed of movement of model prey, these results show clearly that they are not able to recognise insects on the basis of finer details of their appearance.

Thus for some animals the problem of recognising significant objects may be reduced to the problem of detecting localised key stimuli or features that in the natural world are unambiguous cues to appropriate action. Such local features may be quite simple—it is easy to see how a "redness" detector might function in the stickleback, and not too difficult to conjecture how this might be coupled with a rather crude configurational analysis to explain observed preferences for the location of the red patch and the posture of the model. However, such mechanisms are also relatively inflexible, and depend for their success on the predictability of the natural environment. When a scientist introduces a red dummy fish, a paper butterfly, or pieces of cardboard into an animal's surroundings, the assumptions about the properties of mates or prey on which the perceptual mechanism relies are violated.

Other animals, especially primates, have more flexibility in their perception and action and are able to recognise and discriminate on the basis of more complex and subtle criteria. In these cases, as in human perception, the problem of how stimulus equivalence is achieved is a difficult one, as we will see in the remainder of this chapter.

MORE COMPLEX RECOGNITION PROCESSES

We may speculate that at least some behaviour in humans may be under the control of key stimuli. For example, it has been shown (e.g. Goren, Sarty, & Wu, 1975) that human neonates show innate following of face-like patterns, and we discuss this evidence in more detail in Chapter 16. On the whole, however, it is through a process of learning that we come to classify certain configurations as equivalent and distinct from others. The human infant learns to recognise the faces of its parents irrespective of angle, expression, or lighting. A mother will still be "mummy" to her child after she has curled her hair, and a father will still be "daddy" if he hasn't shaved for a few days. Later, the child will learn to distinguish teachers and friends from strangers, family pets from strays, and the long process of formal education enables most to decipher the intricacies of written language. What kinds of internal representations allow for the recognition of complex configurations, and what kinds of processes operate on the retinal image to allow access to these internal representations? These have been the questions posed in the study of human pattern and object recognition.

Much early work on pattern recognition focused on the problem of recognising alphanumeric patterns. There is good reason for such work, as researchers in computer science have had the applied aim of making computers able to recognise such patterns so that they might, for example, achieve automatic sorting of letters with hand-written postal codes. The emphasis on alphanumerics was unfortunate in other ways, because the problem of stimulus equivalence is rather different for alphanumerics than for objects. Letters must be recognised despite changes in their form, but they are only two-dimensional patterns, so that other problems in object recognition are minimised. Nevertheless, the area of alphanumeric recognition is worth discussing briefly because it serves to introduce certain theoretical approaches to the broader area of object recognition.

TEMPLATE MATCHING

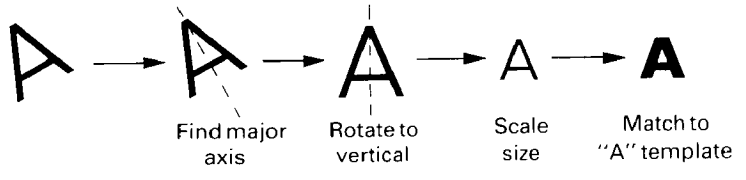
The simplest account that we could offer of how we recognise alphanumeric characters would be that of *template matching*. For each letter or numeral known by the perceiver there would be a template stored in long-term memory. Incoming

patterns would be matched against the set of templates, and if there were sufficient overlap between a novel pattern and a template then the pattern would be categorised as belonging to the class captured by that template. Within such a framework, slight changes in the size or angle of patterns could be taken care of by an initial process of standardisation and normalisation. For example, all patterns could be rotated so that their major axes (as discovered by other processing operations) were aligned vertically, with the height of the major axis scaled to unity (see Fig. 9.4). In addition, some pre-processing or "cleaning up" of the image would be necessary. Both humans and other animals (Sutherland, 1973) cope very well with broken or wobbly lines in the patterns they recognise.

Such a template-matching scheme could work provided that such normalising procedures were sufficient to render the resulting patterns unambiguous. Unfortunately this is almost impossible to achieve, even in the simple world of alphanumerics. An "R" could match an "A" template better than its own, and vice versa (see Fig. 9.5). The bar that distinguishes a "Q" from an "O" may be located in a variety of places (see Fig. 9.6). At the very least we would need more than one template for each letter and numeral, and it becomes difficult to see how children could learn letters and numbers in such a scheme.

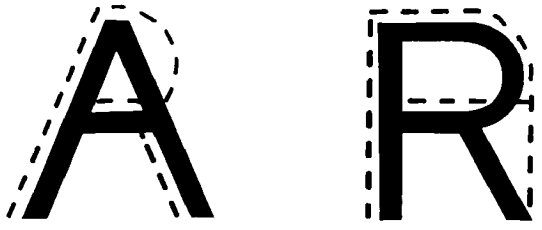
Template-matching schemes also fail to account readily for the facts of animal discrimination. Sutherland and Williams (1969) showed that rats trained to discriminate an irregular from a regular chequerboard pattern readily transferred this learning to new examples of random and regular patterns (see Fig. 9.7). As Sutherland (1973) points out, the configuration in Fig. 9.7d should match better with a "template" for pattern 9.7a than for b, but it is treated by the rats as though it were more like b than a. It is also difficult to see how a template-matching model could possibly be applied to the more general area of object recognition, where the problem of stimulus equivalence is magnified. However, a template-matching process can operate

FIGURE 9.4



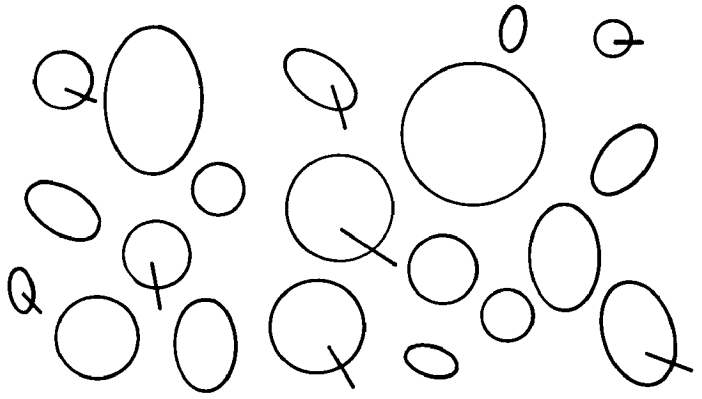
Before matching to a template, a pattern could be standardised in terms of its orientation and size. This could be done by finding the major axis of the figure, rotating this to vertical, and scaling its size to some standard.

FIGURE 9.5



The bold figures show possible templates for an A (left) and an R (right). The dashed figures show how an R (left) and an A (right) could match another letter's template better than their own.

FIGURE 9.6



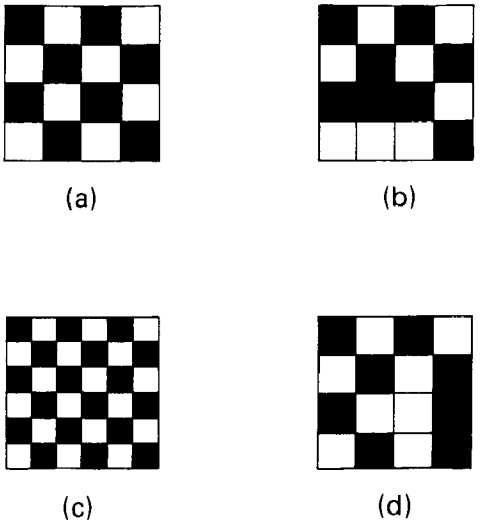
What distinguishes the Qs from the Os? Not the precise form of the circle, nor the precise location or orientation of the bar.

successfully if the form of the characters it must recognise can be constrained. Thus the computer that recognises account numbers on the bottom of cheques matches these to stored templates. The character set has been constrained, however, so that the numerals have constant form, and in addition are made as dissimilar to one another as possible to avoid any chance of confusion. The characters that humans recognise are not constrained in this way.

FEATURE ANALYSIS

When we consider how it is that we know the difference between an A and an R, or a Q and an O, it seems that there are certain critical features that distinguish one from another. The bar that cuts the circular body of a Q is essential to distinguish it from an O, whereas the precise form of the circle

FIGURE 9.7



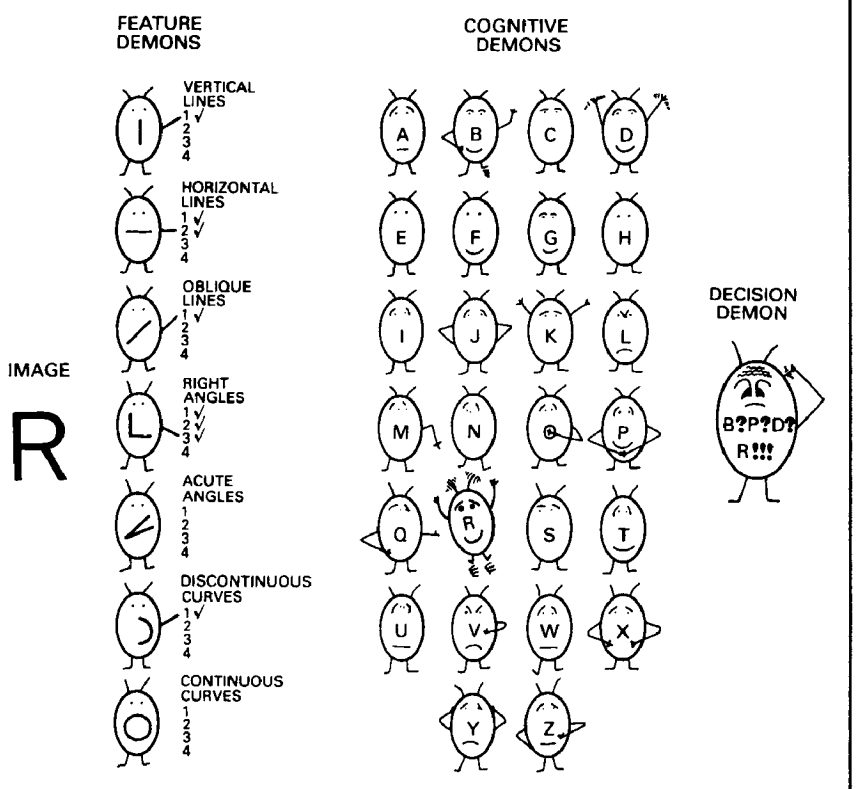
Rats trained to respond in one way to pattern (a), and another way to pattern (b), later treat pattern (c) in the same way as (a), and pattern (d) in the same way as (b). This is not consistent with a template-matching model (Sutherland & Williams, 1969). Reprinted with permission from the author and the Experimental Psychology Society.

is less crucial. Perhaps a model in which combinations of features were detected would be more successful than one based on templates.

Feature analysis models of recognition were popular with psychologists and computer scientists during the 1960s while physiologists such as Hubel and Wiesel were postulating "feature detectors" in the visual cortex of cats and monkeys (see Chapter 3). Perhaps the most influential model for psychology was Selfridge's (1959) Pandemonium system, originally devised as a computer program to recognise Morse Code signals, but popularised as a model of alphanumeric recognition by Neisser (1967), and Lindsay and Norman (1972). An illustration of a Pandemonium system is shown in Fig. 9.8.

The system consists of a number of different classes of "demon". The most important of these for our purposes are the *feature demons* and the *cognitive demons*. Feature demons respond selectively when particular local configurations (right angles, vertical lines, etc.) are presented. The cognitive demons, which represent particular letters, look for particular combinations of features from the feature demons. Thus the cognitive demon representing the letter H might look for two vertical and one horizontal lines, plus four right angles. The more of their features are present, the louder will the cognitive demons "shout" to the highest level, the decision demon, who selects the letter corresponding to that represented by the cognitive demon who is shouting the loudest. Thus in this system individual characters are represented as sets of critical features, and the processing of any image proceeds in a hierarchical fashion through levels of increasing abstraction. It is this kind of model that Barlow (1972) and others used to interpret the properties of simple cells in the visual cortex (see Ch.3, p.54). Simple cells were thought to be acting as the feature demons in the

FIGURE 9.8



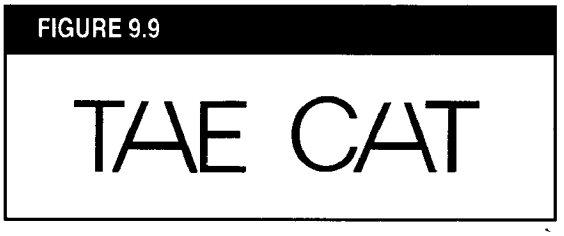
A Pandemonium system for classifying letters. Each of the feature demons responds selectively to a different feature in the image, and signals the number of features present to the cognitive demons. Each of the cognitive demons represents a different letter, and "shrieks" louder the more of its features are present. (Extra features inhibit the responses of cognitive demons.) The decision demon selects the letter that is being shouted the loudest. Liberally adapted from Selfridge (1959) and Lindsay and Norman (1972).

Pandemonium system, passing information on to cells that would supposedly respond to increasingly abstract properties. Such hypothetical cells were dubbed "Grandmother cells" or "Yellow Volkswagen detectors" to express the abstract nature of the stimuli exciting them.

A Pandemonium system can learn to give different weights to different features according to how well these features discriminate between different patterns, and in the next chapter we will consider a number of pattern recognition systems that learn in a similar way, by altering the weights between stimulus and response connections. A system of the Pandemonium type can in principle accommodate certain kinds of contextual effect. These are a ubiquitous feature of human pattern recognition, and Fig. 9.9 shows one example of how context affects the recognition of letters. The same shape can be seen as H or as A depending on the surrounding letters. Within a Pandemonium system we might allow higher-level demons to "arouse" those at lower levels that correspond to particularly likely patterns, so that they would need less sensory evidence to make them shout sufficiently loudly to win over the decision demon. Humphreys and Bruce (1989) give more details of a range of context effects in human pattern and object recognition.

However, as a general model for human pattern and object recognition the Pandemonium system is unsatisfactory. Ultimately it rests on a description of patterns in terms of a set of features, which are themselves like mini-templates. One of the reasons that Pandemonium was so popular was that it seemed consistent with the neurophysiology of the visual cortex; but we have already seen that single cells cannot be thought of as "feature detectors" (see Ch.3, p.54). Although this may not matter for a purely psychological or computational theory of recognition, there are other problems. Feature-list descriptions fail to capture overall structural relations that are captured, but too rigidly, by more global templates. Thus the Pandemonium system depicted in Fig. 9.8 would confuse an F with ⊤ and a T with ⊥, confusions that humans typically do not make. In addition, the Pandemonium system, in classifying patterns, discards all information that distinguishes different instances of the same pattern. The output of the decision demon would be the same irrespective of the particular version of the letter A shown. We need a way of talking about recognition that allows us to describe the differences between patterns as well as being able to classify together those that are instances of the same type. We need to preserve such differences so that other kinds of classifications can be made. We recognise people's hand-writing, for example, by the particular shapes of the letters they produce. Thus we need a representational format that captures aspects of structure essential for the classification of an item but preserves at some other level structural differences between different instances of the same class.

FIGURE 9.9

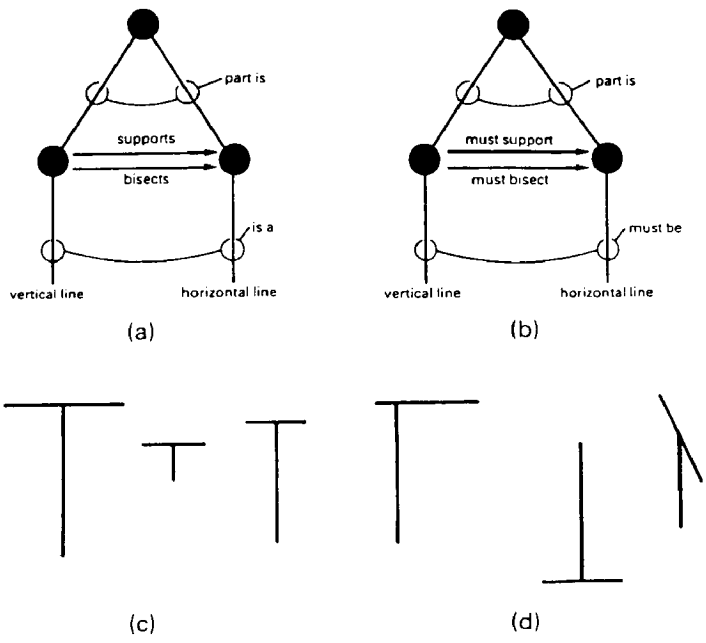


The same shape may be seen as an H in one context and an A in another (from a demonstration by Selfridge).

STRUCTURAL DESCRIPTIONS

A general and flexible representational format for human pattern and object recognition is provided by the language of structural descriptions. Structural descriptions do not constitute a theory of how recognition is achieved, they simply provide the right type of representation with which to construct such a theory. A structural description consists of a set of propositions (which are symbolic, but not linguistic, although we describe them in words) about a particular configuration. Such propositions describe the nature of the components of a configuration and make explicit the structural arrangements of these parts. Thus a structural description of a letter T might look like Fig. 9.10a.

FIGURE 9.10



(a) A structural description for a letter T. The description indicates that there are two parts to the letter. One part is a vertical line, the other a horizontal line. The vertical line supports and bisects the horizontal line. (b) A model for a letter T. This is like the description at (a), but the essential aspects of the description are specified. For something to be a T, a vertical line must support, and must bisect, a horizontal line, but the relative lengths are not important. (c) Shapes that would be classified as Ts by the model. (d) Shapes that would fail to be classified as Ts.

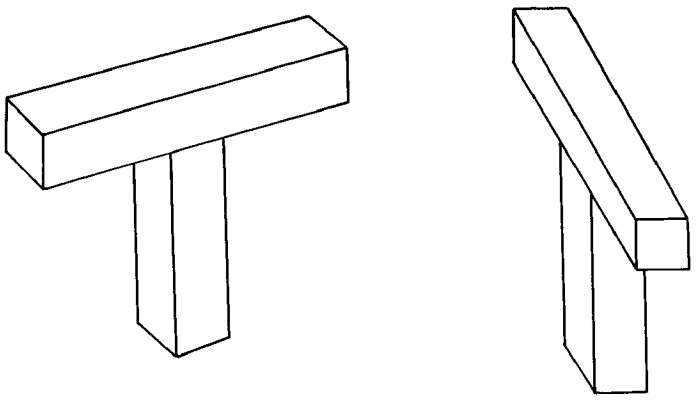
Using the language of structural descriptions it is possible to construct "models" for particular concepts and categories against which any incoming instance can be matched. Such models capture obligatory features of the structure but may be less particular about other details. Thus the "model" for a letter T might look like Fig. 9.10b. It is essential that a horizontal line is supported by a vertical line, and that this support occurs about half way along the horizontal line. But the lengths of the two lines are less important. Figure 9.10c shows examples that would be classified as letter Ts by this model, and 9.10d shows those that would fail.

Structural descriptions are also easier to apply to object recognition than templates or feature representations. A picture of an object can be described by a series of structural descriptions at increasing levels of abstraction from the original intensity distribution. There are thus a number of possible "domains" of description (Sutherland, 1973).

Take for example the two drawings shown in Fig. 9.11. These drawings can be described within a number of distinct domains, which can broadly be grouped together as being either "two-dimensional" or "three-dimensional". The 2-D descriptions describe the picture or image present, and this image can be described in increasingly abstract or global terms. It may be described as a collection of points of different brightnesses, as a collection of lines, or as a group of regions. These different levels of description are reminiscent of the different stages of elaboration of the primal sketch, through the aggregation of small edge segments up to larger contours or aggregated texture regions (see Chapters 5 and 6). Whatever the level of description in the 2-D domain, whether points, lines, or regions, the representations established for these two pictures would look very different. It is within the domain of 3-D description that the equivalence of these two pictures can be established. 3-D descriptions are couched in terms of surfaces, bodies, and objects. The two pictures shown in Fig. 9.11 are equivalent only at the level of an object description that is independent of the vantage point.

The description above again illustrates the thrust of Marr's term "2½-D" sketch for the representation of *surfaces*, from the point of view of the observer. Marr's 2½-D sketch falls

FIGURE 9.11



These two forms are quite different in terms of their two-dimensional description. They are equivalent only in the three-dimensional domain.

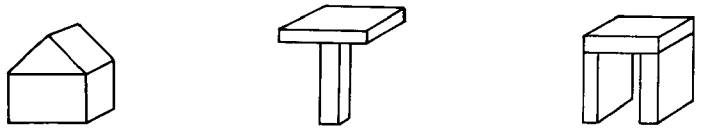
somewhere in between the 2-D and 3-D groups of descriptions in Sutherland's scheme.

Thus two different projections of the same object will have different structural descriptions in the picture domain, but will be equivalent in the object domain (see Fig. 9.11). Provided that structural descriptions are established at all levels simultaneously, we can capture both the equivalences between different views of the same object and their differences. Our problem now is to consider how structural descriptions at the 3-D level can be constructed, stored, and matched, and to examine the extent to which the construction of 3-D representations can proceed in a "bottom-up" fashion.

Winston (1975) provided an early illustration of the use of structural descriptions in object recognition to show how object concepts might be learned by giving examples. His program learns to recognise simple toy block structures such as those illustrated in Fig. 9.12, which contains examples of an "arch", a "pedestal", and a "house".

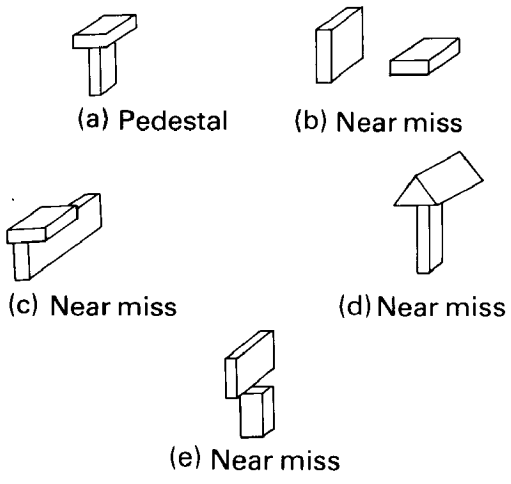
The computer program is presented with examples of each, as well as "near-misses", in order to build up models for each concept. The procedure for a pedestal might go as follows. First, the program would be presented with an example of a pedestal (Fig. 9.13a) to which it would assign the structural description shown in Fig. 9.14a. Thus a pedestal is described as having two parts, with one part being a "brick" and the other part being a "board", with the former supporting the latter. Then the program would be presented with the sequence of "near misses" shown in Fig. 9.13b–e. For Fig. 9.13b, the description would again show two parts, with one a brick and the other a board, but the

FIGURE 9.12



Three of the toy block structures learned by Winston's program. Adapted from Winston (1973) with his permission.

FIGURE 9.13

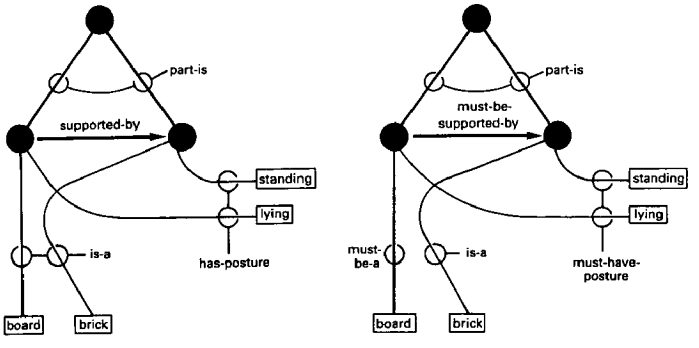


A pedestal training sequence. Adapted from Winston (1973) with his permission.

relationship between these is now different. The board is beside the brick, and the program was told that this is *not* a pedestal. By comparing this description of the near-miss with that of the structure labelled pedestal, the program can construct a model for a pedestal in which the support relation is made obligatory. For something to be a pedestal one part *must* be supported by the other. The other examples in the training sequence (Fig. 9.13c–e) further constrain the eventual model for a pedestal (Fig. 9.14b). The eventual model shows that for something to be a pedestal, an upright brick must support a lying board.

Our choice of a pedestal to illustrate this process of learning a structural model from examples was deliberate. The pedestal is like a three-dimensional letter T (see Fig. 9.10), and the structural description for a pedestal is very similar to that described for a T, except that the parts of the pedestal are themselves three-dimensional objects like a brick and a board, instead of the horizontal and vertical lines in the letter T. Thus, this kind of

FIGURE 9.14



(a) A description of the pedestal in Fig. 9.13a. (b) A model for a pedestal built up after training on a sequence of pedestals and near-misses. Adapted from Winston (1973) with his permission.

representation can be used for two-dimensional written characters, or three-dimensional objects.

To return to Winston's program, a process similar to that used for the pedestal can be used to derive a model for a house (Fig. 9.12). Here the eventual model would specify that a brick must support a wedge (the roof). As for the pedestal, both the support relations and the nature of the objects are quite tightly specified. However, in the case of an arch (Fig. 9.12) there is more flexibility. Although the upright structures in the arch model *must* be bricks and *must not* touch each other, the structure they support can be a brick, or a wedge, or maybe even any object at all. An arch is still an arch whatever the shape at the top.

Winston's program is here operating in the object domain. It can accept any projection of a brick or wedge and label these accordingly. However, the structural descriptions for brick and wedge must themselves be specified at a different level of the program. At an even lower level, the line drawing that serves as input must be parsed into separate objects using the procedures described in Chapter 6. The initial stages of the program make use of programs like Guzman's (see Ch.6, p.120) to group regions of the picture together.

The problems with Winston's system are buried within these low-level programs that furnish the descriptions on which the learning program operates. As we noted in Chapter 6, scene analysis programs of the kind developed by Guzman, Clowes, and Waltz work by making use of the constraints inherent in the kinds of scene they describe. But the constraints of the mini-world of matt prismatic solids are not the constraints of the natural world. Although something similar to Winston's learning program might provide a theory of visual object classification, we need a better way of furnishing structural descriptions for such procedures to operate on—one that is not restricted to an artificial world.

To do this, we must return to consider the fundamental problem of object recognition. To recap, the projection of an object's shape on the retina depends on the vantage point of the viewer. Thus, if we relied on a *viewer-centred* coordinate system for describing the object (one in the picture domain, to use Sutherland's terminology), descriptions would have to be stored for a number of different vantage points. Later in this chapter and in the next one we will consider some recent theories of recognition that do involve the storage of discrete viewpoints, an approach that is now gaining considerable empirical and computational support, at least for certain kinds of recognition task.

However, if we can describe the object with reference to an *object-centred* coordinate system, (i.e. build a structural description in the "object" domain) then it would be possible to reduce the number of object models stored, ideally to only a single one per distinguishable object. This was what Winston attempted to do with an artificial world.

The problem is then to find a way of describing the object within its own coordinate system without confining the discussion to an artificial world, and/or using knowledge of an object-specific kind. If one has to rely at the outset on object-specific knowledge then we would have to know what an object was before we could recognise it—an obvious paradox. However, it seems likely that knowledge of some constraints is essential to parse objects—the question is, how specific are these?

MARR AND NISHIHARA'S THEORY OF OBJECT RECOGNITION

Marr and Nishihara (1978) outlined the foundations for one possible solution to this problem. An object must be described within a frame of reference that is based on the shape itself. To do this, we must be able to set up a canonical coordinate frame (a coordinate frame that is determined by the shape itself) for the shape before the shape has been described.

The appropriate set of descriptive elements (primitives) for describing a shape will depend in part on the level of detail that the shape description is to capture. The fingers of a human hand are not expressed in a system that uses primitives the size of arms and legs. To get around this problem, Marr

and Nishihara suggest that we need a modular organisation of shape descriptions with different-sized primitives used at different levels. This allows a description at a "high" level to be stable over changes in fine detail, but sensitivity to these changes to be available at other levels.

First we need to define an *axis* for the representation of a shape. Shapes that are elongated or have a natural axis of symmetry are easier to describe, and Marr and Nishihara restrict their discussion to the class of such objects that can be described as a set of one or more *generalised cones*. A generalised cone is the surface created by moving a cross-section of constant shape but variable size along an axis (see Fig. 9.15). The cross-section can get fatter or thinner provided that its shape is preserved. The class of generalised cones includes "geometric" forms like a pyramid or sphere, as well as natural forms like arms and legs (roughly). Objects whose shape is achieved by growth are often describable by one or more generalised cones, and so we can talk about object recognition in the natural world, rather than an artificial one. In the discussion that follows we will generally be talking about the recognition of shapes composed of more than one generalised cone, so that there will be more than one axis in the representation. For example, a human figure can be described as a set of generalised cones corresponding to the trunk, head, arms, and legs. Each of these component generalised cones has its own axis, and together these form the component axes for a representation of a human.

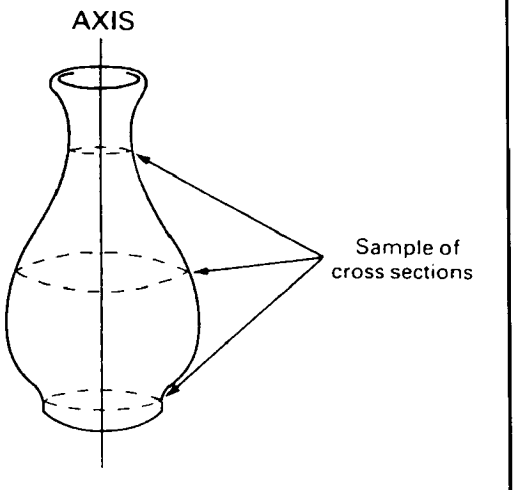
A description that uses axis-based primitives is like a stick figure. Stick figures capture the relative lengths and dispositions of the axes that form the components of the entire structure. The relative thicknesses of these components (e.g. the human trunk is thicker than a leg) could also be included in the representation, although for simplicity we will omit this detail here. Information captured by such a description might be very useful for recognition as stick figures are inherently modular. We can use a single stick to represent a whole leg, or three smaller sticks to represent the upper and lower limb segments and the foot. At a still finer level, we can capture the details of toes with a set of much smaller sticks. At each level of description we can construct a 3-D model where each 3-D model specifies:

1. A single-model axis. This provides coarse information about the size and orientation of the overall shape described.
2. The arrangements and lengths of the major component axes.
3. Pointers to the 3-D models for the shape components associated with these component axes.

This leads to a hierarchy of 3-D models (illustrated in Fig. 9.16), each with its own coordinate system.

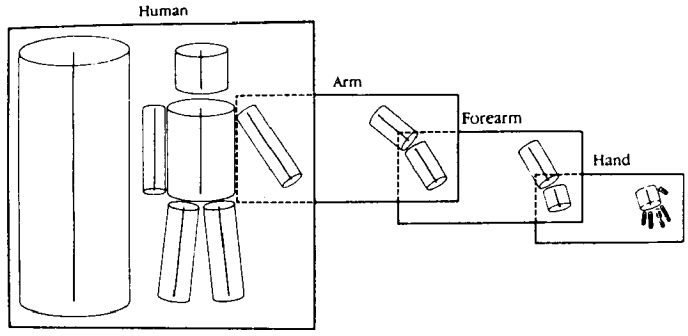
The first "box" in Fig. 9.16 shows the single-model axis for a human body with the relative dispositions of the component axes (corresponding to head, body, legs, and arms). The axis that corresponds to the arm forms the major axis for the "arm model" (next box in the figure), in which the component axes of upper arm and forearm are shown, and so on through to the details of the fingers of a human hand. Such a hierarchy of 3-D

FIGURE 9.15



One example of a generalised cone. The shape is created by moving a cross-section of constant shape but variable size along an axis.

FIGURE 9.16



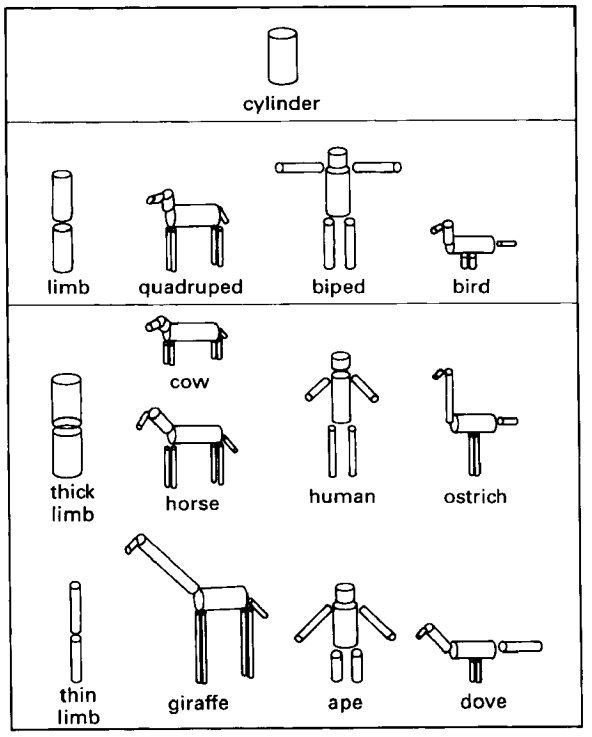
A hierarchy of 3-D models. Each box shows the major axis for the figure of interest on the left, and its component axes to the right. From Marr and Nishihara (1978). Reprinted with permission of The Royal Society.

models is called a 3-D model description. Recognition is thought to be achieved when a match is established between a 3-D model description derived from an image, and one of the stored catalogue of 3-D model descriptions corresponding to known objects. These may in turn be organised hierarchically, in terms of the specificity of their descriptions (see Fig. 9.17). Thus a human figure can be matched to the general model for a biped, or the more specific model for a human. Ape and human are distinguished by the relative lengths of the component axes in the model description for a biped.

At this point we should note that there is some limited evidence for the psychological validity of axis-based representations. For example, Humphreys (1984) asked subjects to decide whether or not two presented objects were the same shape (both elongated triangles or both parallelograms). Humphreys found that when subjects did not know exactly where the second shape would appear relative to the first, judgements were faster if the orientations of the major axis of the shape was preserved, suggesting that this aspect of the shape played a role in the comparison process. Although such results lend some support to Marr and Nishihara's theory, axis-based descriptions do not seem to be constructed when the position of the second shape is known in advance (Humphreys, 1984), nor is there evidence that axis-based descriptions are used for all elongated shapes (e.g. Quinlan & Humphreys, 1993). However, although the evidence for the primary role of axis-based representations is limited, it is also the case that these studies have explored the perception of 2-D shapes rather than the 3-D objects addressed in Marr and Nishihara's theory. Humphrey and Jolicoeur (1993) reported that the identification of line drawings was markedly disrupted when the objects were depicted with their main axis oriented directly towards the viewer so that the main axis appeared foreshortened. This disruptive effect of foreshortening occurred even though the main components of the objects were salient at all viewing angles. Lawson and Humphreys (1996) used a matching task with line drawings of objects rotated in depth. With relatively long intervals between the stimuli there was little effect of the angle between consecutive objects until the to-be-matched stimulus had its main axis foreshortened. These studies lend some support to Marr's theory that object recognition would be disrupted if the major axes of elongation of the object is not visible.

How could such 3-D model descriptions be derived *prior* to accessing the catalogue? The problem is to derive the axes from an image *without* knowing what object it is that the image represents. A possible solution is provided by Marr's (1977) demonstration that we can make use of the occluding contours of an image to find the axis of a generalised cone, provided the axis is not too foreshortened. The only assumption needed is that

FIGURE 9.17



A catalogue of 3-D model descriptions at different levels of specificity. Redrawn from Marr and Nishihara (1978) with permission of The Royal Society.

these contours come from a shape that is comprised of generalised cones.

We have already seen, in Chapter 6, how Marr's early visual processing program derived contour information from an image without knowing what shape it is looking for. Occluding contours in an image are those that show the silhouette of the object (see the outline of the head of the bear in Fig. 6.28, or the donkey in Fig. 9.20). As Marr points out, silhouettes are infinitely ambiguous, and yet we interpret them in a particular way (1982, p.219):

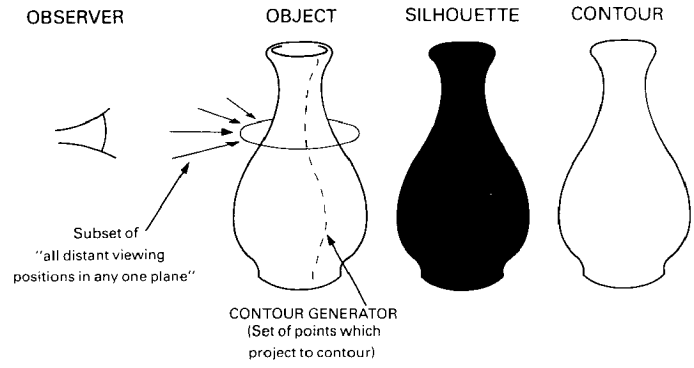
> Somewhere, buried in the perceptual machinery that can interpret silhouettes as three-dimensional shape, there must lie some source of additional information that constrains us to see silhouettes as we do. Probably ... these constraints are general rather than particular and do not require a priori knowledge of the viewed shapes.

Let us examine the assumptions that Marr suggests allow us to interpret silhouettes so consistently:

1. Each line of sight from the viewer to the object should graze the object's surface at exactly one point. Thus each point on a silhouette arises from one point on the surface being viewed. We can define the *contour generator* as the set of points on a surface that projects to the boundary of a silhouette (see Fig. 9.18).
2. Nearby points on the contour in an image arise from nearby points on the contour generator on the viewed object.
3. All the points on the contour generator lie in a single plane (see Fig. 9.19).

This third is the strongest assumption, but is necessary in order to distinguish convex and concave segments in the interpetation process. If

FIGURE 9.18

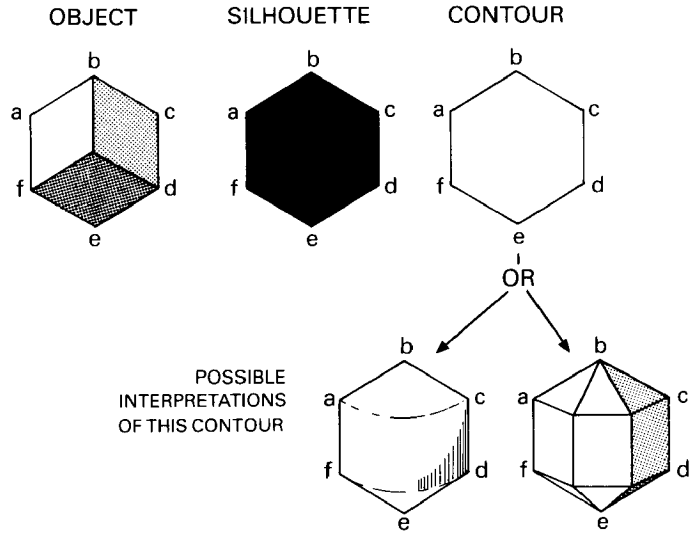


An object, its silhouette and its contour. The set of points that projects to the contour (the contour generator) is shown. For this figure, all three assumptions (see text) hold for all distant viewing positions in any one plane. Adapted from Marr (1977) and Marr (1982) with permission of The Royal Society.

this assumption is violated, then the wrong conclusion might be reached. For example, the occluding contour in the image of a cube, viewed corner on, is hexagonal (see Fig. 9.19). Because we assume the contour generator is planar, we could interpret such a silhouette wrongly. In the absence of any other information from internal lines or motion, we might interpret the contour as belonging to a spindle shape like one of those drawn, or simply as a flat hexagon. In fact, the points on the cube that gave rise to this contour do not lie in a single plane. It is this assumption of a planar contour generator that may lead us (wrongly!) to interpret the moving silhouette of someone's hands as the head of a duck, or an alligator, while playing shadow games.

Marr has shown that if a surface is smooth, and all the above assumptions hold for all distant

FIGURE 9.19



A cube viewed corner on gives rise to the silhouette and contour shown. The contour generator (a-b-c-d-e-f) is not planar. This silhouette might be seen simply as a hexagon, or interpreted as one of the spindle shapes shown.

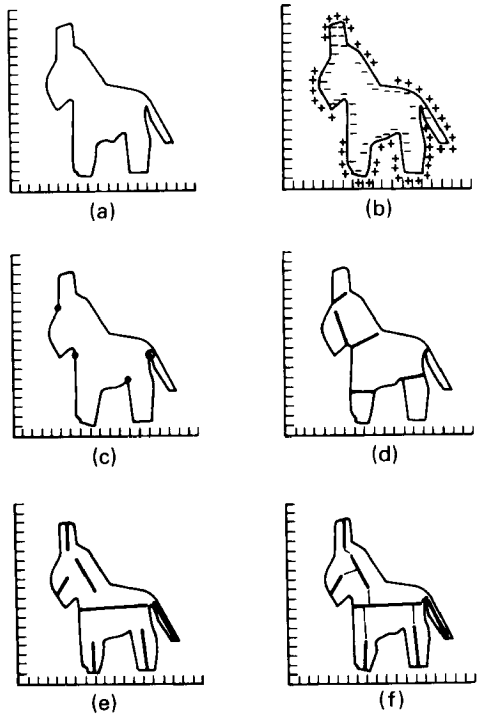
viewing positions in any one plane (see Fig. 9.18), then the viewed surface is a generalised cone. Thus shape can be derived from occluding contours *provided* the shape is a generalised cone, or a set of such cones.

Vatan (cited by Marr, 1982) has written a program to derive the axes from such a contour. Figure 9.20 shows how his program derives the component axes from an image of a toy donkey. The initial outline was formed by aggregating descriptions from the raw primal sketch, in the same way as for the teddy bear's head (Ch.6, p.125). From this initial outline, convex and concave segments are labelled and used to divide the "donkey" into smaller sections. The axis is derived for each of these sections separately, and then these component axes are related together to form a "stick" representation for the entire figure.

Now these axes derived from occluding contours are viewer-centred. They depend on the image, which in turn depends on the vantage point. We must transform them to object-centred axes, and Marr and Nishihara (1978) suggested an additional stage to achieve this by making use of the "*image-space processor*". The image-space processor operates on the viewer-centred axes and translates them to object-centred coordinates, so that the relationships between the different axes in the figure are specified in three, instead of two dimensions. Use may be made of information from stereopsis, texture, and shading to achieve this, but it may also be necessary to use preliminary matches with stored 3-D-model description to improve the analysis of the image. Thus, for recognition, Marr does envisage that there is a continuous interplay between the derivation of an object's description and the process of recognition itself (1982, p.321):

> We view recognition as a gradual process that proceeds from the general to the specific

FIGURE 9.20



(a) An outline of a toy donkey. (b) Convex (+) and concave (–) sections are labelled. (c) Strong segmentation points are found. (d) The outline is divided into a set of smaller segments making use of the points found at (c) and rules for connecting these to other points on the contour. (e) The component axis is found for each segment. (f) The axes are related to one another (thin lines). Redrawn from Marr and Nishihara (1978) with permission of The Royal Society.

and that overlaps with, guides, and constrains the derivation of a description from the image.

In summary then, Marr and Nishihara outlined a scheme in which an *object-centred* representation, consisting of an axis-based structural description, could be established from an image and used to access a stored catalogue of 3-D-model descriptions in order for recognition to be achieved. Once an initial match has been established, use may then be made of downward-flowing information to refine the analysis of the image. These ideas of Marr's were speculative; only a few isolated details of these derivation and recognition processes have been specified sufficiently clearly to implement them; and the system itself rests on a number of assumptions and observations about the perception of stick figures and silhouettes that have a rather ad hoc flavour. Nevertheless, in the years since Marr and Nishihara's (1978) theory, there have been a number of developments of these basic ideas.

BEYOND GENERALISED CONES

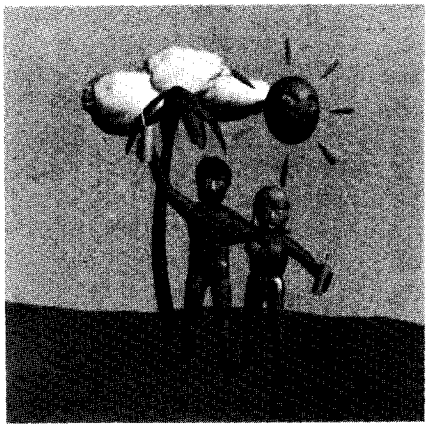
An important step in the development of Marr and Nishihara's theory was the suggestion that complex occluding contours formed from objects comprising several generalised cones are segmented at regions of sharp concavity. In Chapter 6, we described the work of Hoffman and Richards (1984) who have illustrated the importance of such concavities in segmenting contours to reveal parts for recognition, and thereby have supported one aspect of Marr and Nishihara's theory. However, Hoffman and Richards' scheme is independent of the nature of the "parts" within the image. It will work if these are generalised cones, but it will work too if they are quite different kinds of shapes. Since Marr and Nishihara's theory of recognition was formulated, a number of authors have suggested extensions to their basic approach, to encompass a wider range of shapes among the component parts.

For example, Pentland (1986a) proposed a more flexible system of volumetric representation than can be achieved with generalised cones. Pentland suggests that most complex natural shapes are comprised of *superquadric* components and that these might be the basic components that we recover when analysing images of natural objects. Superquadrics include simple shapes such as spheres and wedges, and all kinds of deformations on these shapes that preserve their smoothly varying form and that do not introduce concavities. Figure 9.21 shows a scene constructed with superquadric components. Pentland's theory is an interesting development for computer vision and graphics, but no evidence has been offered for its *psychological* plausibility.

In contrast, Biederman (1987a) has offered a theory of human object recognition that is clearly related to early ideas of Marr and others, although with some key differences, and which he supports with evidence from a variety of psychological experiments.

In Biederman's theory, complex objects are described as spatial arrangements of basic component parts. These parts come from a restricted set of basic shapes such as wedges and cylinders. Biederman calls these shape primitives

FIGURE 9.21



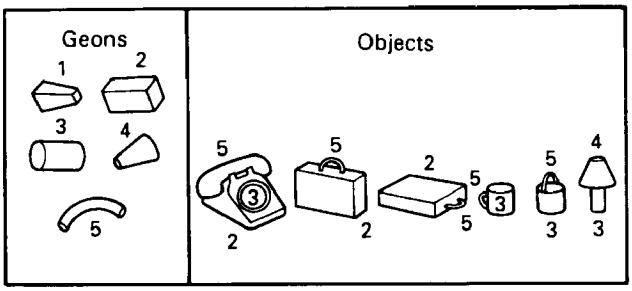
A scene comprised of superquadrics. Reproduced from Pentland (1987) with permission.

"geons" (a short-hand for the phrase geometric ions), suggesting an analogy with words, which are also constructed from combinations of primitives—phonemes. Like Marr and Nishihara, Biederman suggests that the first stage of object description involves the segmentation of the occluding contour at regions of sharp concavity. This divides the contour into a number of parts, which can then be matched against representations of the primitive object shapes (geons). The nature and arrangements of the geons found can then be matched with structural models of objects. The representation of each known object is a structural model of the components from which it is constructed, their relative sizes, orientations, place of attachment, and so forth (see Fig. 9.22). Where members of the same basic object category (e.g. piano) may have quite different shapes (e.g. grand piano vs upright piano) then more than one structural model would be stored for the object.

The main point of departure of Biederman's theory from Marr and Nishihara's is the suggestion that geons are defined by properties that are *invariant* over different views. According to this theory, it is not necessary to make use of occluding contours to recover an axis-based *three-dimensional shape* description. Instead, each different kind of geon has its own "key" features in the 2-D primal sketch level representation. Thus in Biederman's theory, unlike Marr's, object recognition can be achieved directly from the 2-D (primal sketch) level representation with no need to construct an explicit representation of 3-D shape. Biederman argues that there a number of "nonaccidental" properties of edges in images that can be used as reliable cues to related properties of edges in the world (cf. Kanade, 1981; Lowe, 1987). The "nonaccidental" principle is an assumption that when a certain regularity is present in an image, this is assumed to reflect a true regularity in the world, rather than an "accidental" consequence of a particular viewpoint. We can illustrate this with the example of a straight line in an image. This will usually result from an extended straight edge in the world, but it *could* result from other "accidental" consequences of viewpoint; for example, a bicycle wheel viewed end-on will give rise to a straight line image, even though it is actually curved. The nonaccidental assumption would lead to the wrong answer in this case, but will usually be correct, and the general assumption is required in order to constrain the interpretation of essentially ambiguous image data. The nonaccidentalness assumption leads to assertions such as that curved lines in images result from curved edges in the world, parallel edges in an image derive from parallel edges in the world, symmetry in the image signals symmetry in the world, and so forth. Nonaccidental properties include collinearity, curvilinearity, symmetry, parallelism, and cotermination (see Fig 9.23).

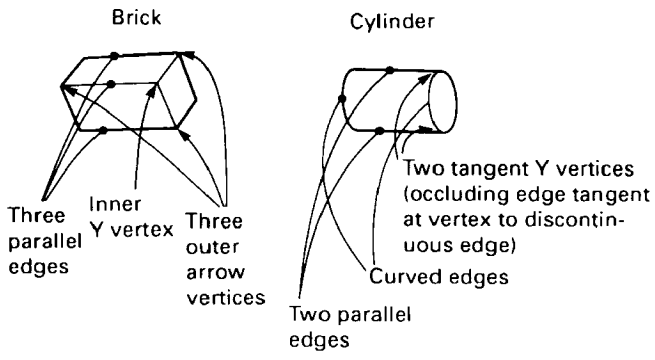
A geon is identified by a particular set of defining features (such as parallel edges) that can be accessed via these nonaccidental properties. Biederman suggests that the assumption of nonaccidental properties could explain a number of illusions such as the Ames chair (see Ch.4, p.71), and "impossible" objects, where, for example, the

FIGURE 9.22



A selection of the volumetric primitives called "geons" (left-hand panel) are used to specify objects in the right-hand panel. The relations between the geons are important, as shown by the difference between a pail and a cup. Reproduced from Biederman (1987b) with permission © 1987 IEEE.

FIGURE 9.23



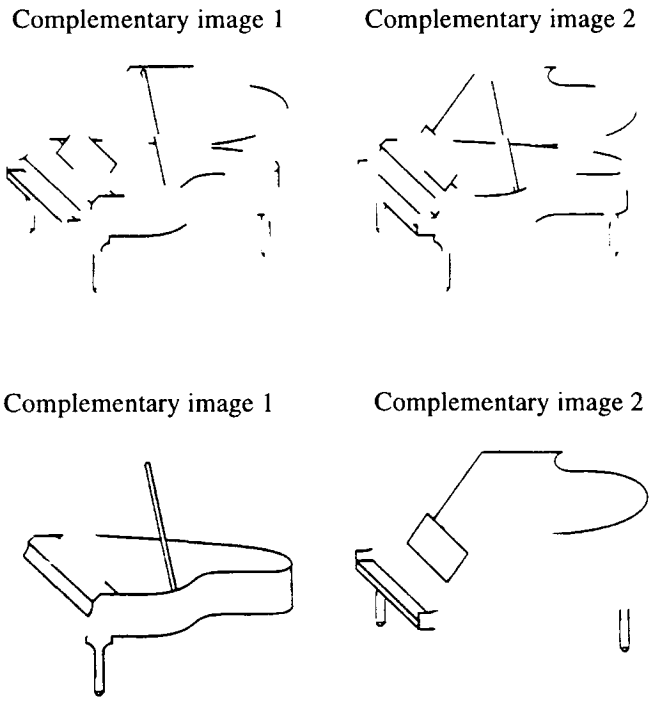
Nonaccidental differences between a brick and a cylinder. From Biederman (1987a). Copyright © 1987 by the American Psychological Association. Reprinted with permission.

cotermination assumption is violated. Biederman also provides new evidence for the importance of concavities in defining the part structures of objects (cf. Hoffman & Richards, 1984). Biederman (1987a) describes experiments in which objects were presented with regions of contour deleted—either at places where there were concavities in the occluding contour that should help define the parts structure, or from segments between these concavities. Contour deletion had a far greater detrimental effect on recognition when information about concavities was removed than when this was preserved. Biederman and Ju (1988) also produced evidence that supported the proposal that it is *edge* properties, rather than surface or texture properties, that are used to classify objects into basic categories. In Biederman and Ju's experiments object recognition was affected rather little by whether or not appropriate or inappropriate colour was added to line-drawn objects in a recognition test, suggesting that the recognition processes ignore such surface properties.

Biederman has also performed a series of empirical studies that appear to support the geon theory of object recognition. When a picture of an object is presented twice for naming, the naming latency on the second occurrence is much faster than on the first. This speeding up of responses from one presentation to the next is known as *repetition priming*. Biederman and Cooper (1991) investigated how repetition priming is affected by a change in the way a line drawing of an object is depicted. The amount of priming obtained is reduced when the repeated presentation shows a different exemplar of the category that would access a distinct structural model (e.g. an upright piano followed by a grand piano), compared with the amount of priming shown when the same exemplar is repeated (e.g. another picture of an upright piano). This difference gives a measure of "visual" priming at the level of the structural model itself over and above additional "conceptual" priming that might occur as a result of re-accessing the same object meaning or category label. Biederman and Cooper (1991) showed that the magnitude of this visual priming of object identification was unaffected if the second view of the same object exemplar showed the same object components, represented by complementary but nonoverlapping image edge features. However, visual priming was reduced if the depicted components (geons) themselves were changed from first to second presentation (when different volumetric parts of the same object exemplar are shown on the two occasions). (See Fig. 9.24.)

Further experiments have shown that priming is invariant over other changes that alter the image but preserve its components, such as size, location, and moderate changes in viewpoint. In contrast, these same manipulations do affect memory for line-drawn pictures (Biederman & Cooper, 1992; Cooper, Schacter, Ballesteros, & Moore, 1992; Humphrey & Khan, 1992), suggesting that variations in object components that are irrelevant for

FIGURE 9.24



Examples of materials used in Biederman and Cooper's (1991) experiment. The top panel shows two complementary images of a piano, created by deleting alternate segments of contour in each image. The amount of visual priming obtained when one member of this pair was followed by the other was as great as when identical images were repeated. The bottom panel shows two complementary images produced by deleting alternate geon components. The amount of priming obtained when one member of such a pair was followed by the other was much reduced and attributed to conceptual rather than visual processes. Adapted from Biederman and Cooper (1991) with permission of the author.

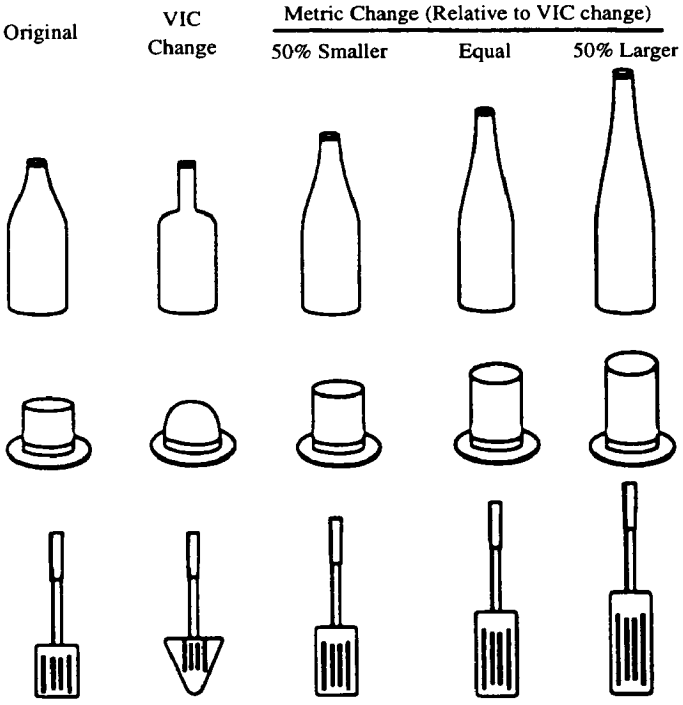
identity may be processed and maintained by other parts of the visual system, perhaps those to do with spatial layout and action. The location of an object in the visual field does not affect its identity, but will affect how an observer reacts to it (e.g. if reaching out to grasp it, or ducking to avoid being hit by it).

Cooper and Biederman (1993; see also Biederman, 1995) furnished other evidence supporting the geon theory. In one study, people were asked to decide whether two objects shown successively were the same or different in name. When objects shared the same name (e.g. both were wine goblets), the two exemplars could differ in terms of the geon shown (e.g. the bowl of the goblet could have rounded or straight sides) or they could differ in a way that did not involve any change in non-accidental properties and hence geons (e.g. the bowl of the goblet could be stretched in the second view compared with the first) (see Fig. 9.25). They found that matching was slowed more (and became more error-prone) by a change in geon than by other metric changes that left the geons unchanged, suggesting that it is the *categorisation* of the shape parts, rather than *holistic or metric* properties of shape, that determines ease of matching.

VIEWPOINT-DEPENDENT RECOGNITION

The theories of object recognition we have discussed here emphasise the recognition of objects irrespective of viewpoint. In fact, there is evidence that not all views of objects are equally easy to recognise. Palmer, Rosch, and Chase (1981) described how each of the different objects they examined appears to have a "canonical" viewpoint, which is often, though not always,

FIGURE 9.25



Examples of the geon-changed (VIC change) and metric-changed shapes used by Cooper and Biederman (1993). Object–name matches were disrupted more by a change of geon than by a metric change even when the metric change was 50% greater (see far-right column) than the amount of metric change that was rated as subjectively equal to the geon change. Reprinted by permission of the author.

something like a three-quarters view. People asked to imagine such objects report imaging them in their canonical views, and people asked which view they would choose to photograph, or which view of an object is "best", select canonical views. Importantly, Palmer, Rosch, and Chase also found that these canonical views could be named more quickly than other views, suggesting that such views play a privileged role in object recognition.

The advantage of canonical viewpoint could quite easily be accommodated by the theories discussed earlier, even though these stress the recognition of objects independent of viewpoint. Marr and Nishihara (1978) emphasise that certain viewpoints will conceal important major axes that are needed to derive a shape description. For example, a top view of a bucket conceals the axis of elongation, which is probably crucial to its description. For Biederman, certain views may conceal the nonaccidental properties that define the "geons", and other views may reveal them better.

Biederman and Gerhardstein (1993) conducted a series of experiments using the repetition-priming method to investigate whether object recognition was invariant across viewpoint. In their experiments they examined how priming was affected by a change in orientation of the object from the view experienced in the first phase. They found that, provided different viewpoints revealed the same geon components, the amount of repetition priming was affected very little by an angular change of up to 135° between its first and second presentation. If successive viewpoints revealed different geons then the amount of priming was affected more greatly.

Other experiments, however, seem to reveal a much greater dependence on viewpoint than do Biederman's. For example, Bülthoff and Edelman (1992; also Edelman and Bülthoff, 1992) showed that when people were asked to try to recognise rather complex unfamiliar shapes they showed very poor abilities to recognise them in novel

viewpoints, even when they had been studied under conditions that ought to have promoted the formation of a 3-D viewpoint-invariant description. There is some dispute about whether such effects of viewpoint-dependence arise only when objects are drawn from a very restricted set within which there are no distinguishing "geons" (see Biederman & Gerhardstein, 1993; and Tarr and Bülthoff, 1995, for discussion). However, effects of viewpoint-dependence have also been found with the kinds of familiar object categories studied by Biederman. For example, Lawson, Humphreys, and Watson (1994) reported experiments in which subjects were required to identify an object from a series of briefly presented pictures. Priming effects were strongly influenced by the visual similarity of successive views, a result that should not be expected if each recognisable view contacts a viewpoint-independent description (see also Lawson & Humphreys, 1996).

A number of authors therefore suggest that our usual ability to recognise objects across a range of viewpoints arises as a result of our experiencing and storing different viewpoints separately, rather than through the recognition of viewpoint-invariant features (Biederman) or the storage of a viewpoint-invariant model (Biederman, Marr). If discrete viewpoints are stored, recognition of novel views may be achieved by alignment of a novel image with one of those stored (e.g. see Bülthoff & Edelman, 1992; Tarr, 1995; Tarr & Pinker, 1989; Ullman, 1989). Theories of viewpoint-dependent recognition of objects are developing rapidly (e.g. see Edelman, 1995). However, it is important to note that object recognition is but one of the tasks accomplished by vision. If recognition can be achieved directly from 2-D features, as Biederman suggests, or through storing a number of viewpoint-specific exemplars (Tarr & Bülthoff, 1995) or "prototypes" (Edelman, 1995), this does not imply that 3-D descriptions of objects are not constructed to guide other actions, such as picking up the object. Different kinds of representation are needed for different kinds of visual task, and even within the task of object recognition it is possible that flexible representational systems are used depending on task demands (see Tarr, 1995, for a discussion).

Whatever the resolution of the rather intense debate about the mechanism by which viewpoint-invariance is achieved, the theories of Marr and Nishihara, and Biederman are all rather limited in scope, because they can only account for the recognition of *basic* categories of object from different configurations of parts. Humans can recognise much more subtle distinctions within classes of objects that share a similar configuration. We can recognise our individual dogs and houses, not just tell a dog from a horse or a house from a church. This ability to recognise objects from within a basic object category is at its most developed when we come to consider recognition of the human face.

DISCRIMINATING WITHIN CATEGORIES OF OBJECTS: THE CASE OF FACE RECOGNITION

Human faces must all be similar in overall configuration because of the other functions of signalling (e.g. expressions) and sensing (e.g. seeing) that they subserve (see Chapter 16). Individual identity must be stamped on this basic configuration. What do we know about the basic form of the representations used to tell one individual face from another?

In contrast to basic-level object recognition, face recognition is not very successful if based on simple "edge" features alone, and seems to require information about surface characteristics such as the pigmentation and/or the texture of skin and hair. One example arises from an experiment by Davies, Ellis, and Shepherd (1978), who showed that famous faces were very poorly recognised from outline drawings that traced the features of faces.

Bruce et al. (1992a) replicated this observation in an evaluation of Pearson and Robinson's (1985) algorithm for sketching images of faces (see Ch.5, p.88). They found that famous faces were quite difficult to recognise when presented as sketches made using Pearson and Robinson's (1985) "valley detecting" algorithm alone, but that the addition of the component that blacks in areas that were dark in the original photograph (see Fig 5.11) restored

recognition of these computer-generated sketches to a level comparable to that obtained with the original photographs.

Moreover, Bruce et al. (1994) showed that repetition priming of faces was considerably reduced if there was a change in the image characteristics between the first and second presentation of faces. Priming was reduced if faces were initially seen as photographic images, and then tested as sketches produced by the Pearson and Robinson (1985) algorithm, or vice versa, compared with the amount of priming produced when the format of the images remained constant between the prime and test phases of the experiment. The viewpoint, expression, and face features remain the same between the photographic and sketch versions—what varies is the details of the grey levels across the image. This sensitivity to image format in face priming is in apparent contrast to basic-level object recognition, where Biederman and Cooper (1991) found that priming was insensitive to changes in the image features.

Another observation is difficult to explain if edge features form the basis of the representational primitives used for face recognition. Faces are extremely difficult to recognise from photographic negatives (e.g. Phillips, 1972), although a negative of a face preserves the spatial layout of edges from the original image. Bruce and Langton (1994) were able to show that this impairment of face recognition did not occur when three-dimensional surface shapes of faces were negated (see Fig. 9.26). This finding suggests that the critical factor in the negation effect is the reversal of the relative brightness of pigmented areas such as hair and skin, which are absent from such surface shapes.

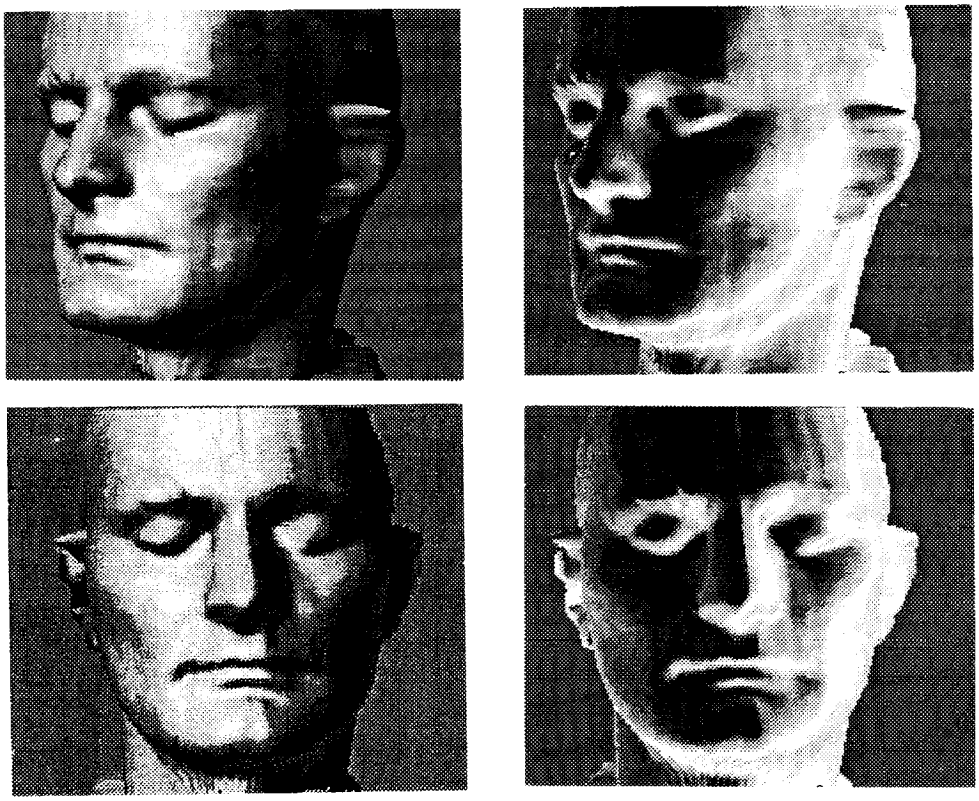
These studies of the identification of line-drawn and negated faces suggest that the surface properties of skin and textured areas such as hair—in particular their relative lightness and darkness—play an important role in face recognition. This need not imply that faces are represented in a radically different way from other objects, as object recognition also seems more dependent on surface properties when the task of discriminating within categories becomes more difficult. For example, Price and Humphreys (1989) showed that when objects to be recognised were drawn from structurally similar categories (such as animals or vegetables), there was a greater advantage in recognising them if they were coloured appropriately rather than inappropriately.

So, one contrast between face recognition (and, perhaps, within-category recognition more generally) and the recognition of basic object types is the extent to which representations preserve information about surface properties. Moreover, a further difference seems to arise in the extent to which different kinds of object discrimination involve decomposing the object shapes into parts, or analysing them more holistically.

The theories of Marr and Nishihara, and Biederman, which we have discussed at length in this chapter, emphasise the decomposition of object shapes into discrete parts, followed by the identification of these parts and their spatial relationships. In contrast to the evidence for a part-based representational scheme for objects, face representation seems to be more "holistic", or at least the relationships between parts (their configuration) seems to be more important in the coding of faces than in that of most objects.

The main observation favouring the holistic processing of faces is that it seems to be difficult or impossible to encode a particular part, or "feature", of an upright face without some influence from other, more distant features. It is not just that the spatial arrangement of face features is important—after all, we have seen that the spatial arrangement of geons is crucial for the definition of an object. For faces, it seems either that the internal description of the parts themselves is influenced by that of other parts, or that parts are not made explicit in the description that mediates face identification. For example, Young, Hellawell, and Hay (1987) took pictures of famous faces and divided them horizontally across the centre. They showed that subjects were able to identify these halves in isolation. When halves of different faces were recombined, however, it became extremely difficult for subjects to name the people who contributed to the composites if these were aligned—new (and unfamiliar) faces seemed to emerge from the combination of the top half of, say, Margaret Thatcher's face and the bottom half of, say, Princess Diana's. However, when the

FIGURE 9.26

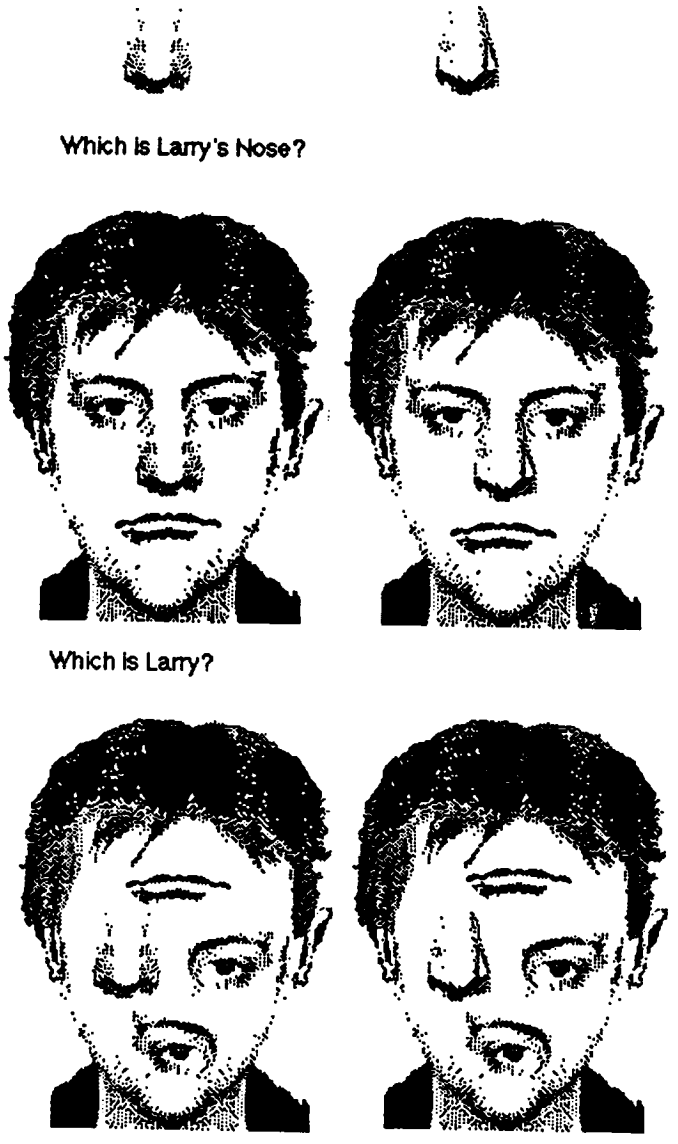


Examples of the surface images used by Bruce and Langton (1994) to explore effects of negation in the absence of surface pigmentation. Positive (left) and negative (right) versions are shown of two of the different viewpoints used in the experiments. Reprinted from Bruce and Langton (1994). © 1994 Pion Ltd. Used by permission.

composite faces were presented upside down, subjects' abilities to identify the halves improved.

Further evidence for the specific use of nondecomposed facial properties in face identification has been obtained by Tanaka and Farah (1993). They asked subjects to learn the identities of individuals constructed from Mac-a-Mug, an electronic "kit" of face features, available for the Macintosh computer. After learning the faces, subjects were asked questions such as "Which is Larry's nose?", where they had to choose the nose that went with the face they had learned to identify as Larry (see Fig. 9.27). Subjects were much better at making this judgement when the noses were shown in the context of the whole face, then when presented in isolation. However, this advantage for presentations of the whole face was not shown when identities had initially been learned for scrambled faces, upside-down faces, or houses (in the latter case, questions about windows, doors, etc. replaced those about face features such as the nose). These results suggest that memory for intact, upright faces is not based on a representation in which parts are made explicit, in contrast to memory for jumbled or inverted faces. Note, though, that the results do not necessitate the view that facial representations are nondecomposed; the results are also consistent with the idea that memory representations for faces are based on emergent, configural descriptions in

FIGURE 9.27



Examples of isolated part, intact face, and scrambled face test items used by Tanaka and Farah (1993). Subjects in these experiments were better able to distinguish the correct version of a feature (such as Larry's nose) when it appeared in the context of the whole face (centre) than on its own (top row) or in a scrambled face (bottom row). Reproduced from Tanaka and Farah (1993) with permission of the authors and the Experimental Psychology Society.

which parsed features are *no longer* represented independently (Bruce & Humphreys, 1994).

This evidence suggests that, even if face identification does involve part-decomposition, there may be a difference in the relative importance of parts versus their configuration, in the representation of basic kinds of objects versus faces. One theory suggests that the relative emphasis on configural and/or holistic processing of faces emerges as a function of expertise with this

object class (e.g. see Carey, 1992), and is orientation-specific. Upside-down faces, which are very difficult to recognise, seem to induce a more parts-based analysis compared with upright faces (see also Bartlett & Searcy, 1993; Rhodes, Brake, & Atkinson, 1993; Young et al., 1987). Diamond and Carey (1986) showed that people who were dog experts also showed dramatic effects of inversion of dog pictures, comparable to the effects of inverting faces, and suggested that the special "configural" mode of processing faces was something that might emerge with expertise within any class of objects sharing the same basic-level configuration. On this argument, face recognition is "special" only in so far as it is a task of within-category recognition at which we are all highly expert, and face recognition can be used to exemplify the more general process of within-category object recognition. (For further discussion and evidence about whether or not face recognition involves specific mechanisms or neural networks not shared with other objects, see Bruce & Humphreys, 1994).

Of course, in Biederman's terms, objects sharing the same overall configuration must share the same geon structural description, and thus some other way of discriminating that which is based on holistic and/or surface properties must be invoked.

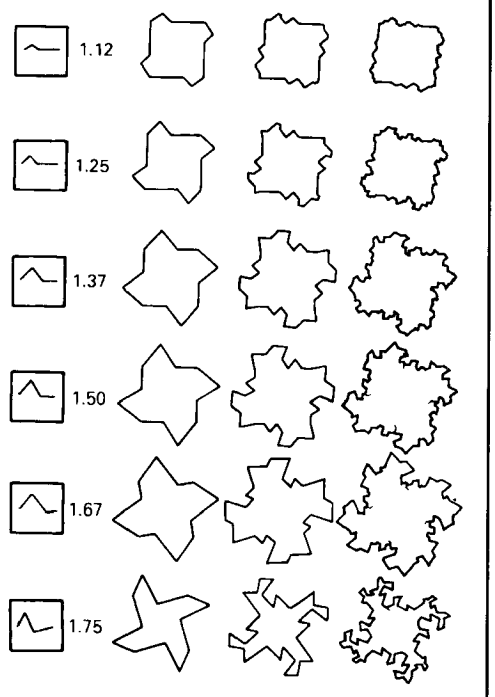
However, even within the domain of basic level object recognition, some workers have produced evidence seeming to favour more holistic over part-based object description schemes. Cave and Kosslyn (1993) showed that the identification of objects was severely disrupted by the scrambling of the spatial arrangement of the overall shape, a result that would be expected on a part-based as well as a holistic coding scheme. However, they also found that it mattered rather little how the objects were divided into parts. Dividing objects in ways that coincided with natural part boundaries (i.e. ways that kept geons intact) produced little advantage over dividing them in ways that did not maintain natural part boundaries. It was only when exposure durations were extremely short that there was an advantage for the natural over the unnatural part divisions. Cave and Kosslyn suggest that people can use parts such as geons as the building blocks for recognition but that they do not need to do so. One problem with Cave and Kosslyn's study, however, is that naturally parsed geons may serve as "objects" for perceptual identification in their own right, which may then compete for identification, thereby disadvantaging the identification of the compound objects. Objects divided in other ways (not into natural parts) would not suffer such competition.

FRACTALS

A further class of objects that are not readily characterised in the simple, part-based way envisaged by Marr and Nishihara, Biederman, and others include naturally rough, crumpled, or branching objects such as trees and clouds, and many textures such as rocky or sandy terrain. Some such "rough" patterns can be described as having a fractal structure (Mandelbrot, 1982). Fractals are patterns that have a fractional dimensionality. For example, a plane is two-dimensional whereas a cube is three-dimensional. A fractal pattern of dimension 2.1 would be almost smooth, like a plane, but with a slightly bumpy surface. As the fractal dimension increased towards 3, the surface would become increasingly craggy. Fractal patterns also have a recursive structure—they look the same at different scales. Figure 9.28 shows some examples of fractal patterns. Pentland (1986b) showed that human perception of the "roughness" of a surface was highly correlated with its fractal dimension as this ranged between 2 and 3 in the way described above, but did not compare the predictive power of fractal statistics with that of any other variable. Cutting and Garvin (1987) showed that ratings of the complexity of patterns like those shown in Fig. 9.28 are well predicted by their fractal pattern statistics, but also found that other variables, such as the number of sides, were equally good predictors of perceived complexity.

Pentland (1986b) describes how fractal-based methods can be used to segment natural images into different regions and objects (cf. Chapter 6), and describes how objects more natural-looking than those shown in Fig. 9.21 can be built by adding

FIGURE 9.28



Examples of fractal patterns derived from different "generators" (left column) whose fractal dimension varies from 1.12 to 1.75. The patterns generated along the rows vary in terms of their depth of recursion—the extent to which the generation process is repeated at different scales. Reproduced from Cutting and Garvin (1987) with permission of the Psychonomic Society, Inc.

together superquadric components using a fractal generation process to roughen the surface. Although this is of considerable interest as a computer graphics application, and shows how a basic "part-based" shape description could be extended so that it could apply to more natural objects, it remains to be seen whether the human visual system makes use of any such system when recognising objects.

CONCLUSIONS

In this chapter we have outlined some of the problems posed by the recognition of objects from retinal images, and have seen how contemporary work in cognitive science has attempted to overcome these problems. We are still a long way from developing a computer program that can recognise everyday objects with the ease that we do, and some way off understanding how we ourselves perform everyday tasks of natural object recognition. The theories of recognition we have discussed in this chapter differ in the extent to which objects are thought to be recognised via abstract models that are viewpoint independent, or by the storage of particular instances or viewpoints seen on distinct occasions. In the next chapter, we will consider how recent connectionist models of object recognition can give a feel for how apparently "abstract" representations might be built up from discrete encounters with objects in the world.