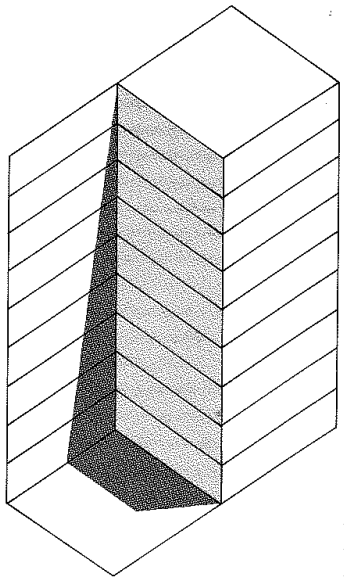

Figure 4–9. The strange reversal of this figure may, like the reversal of the Necker cube, be due to constraints embedded in the 2½-dimensional sketch.

is done in the 2½-D sketch proper and how much occurs as this immediate representation is computed into a three-dimensional representation of the kind that we remember (see the next chapter). Examples like the Penrose triangle, many of Escher's figures, and even Figure 4–9 probably depend on a mixture of effects, some local in the 2½-D sketch, and other effects due to a failure to construct an overall, consistent three-dimensional interpretation from a set of local views.

One final point that might be thought puzzling. Why should the Necker cube reversal occur when depicted in a random-dot stereogram? It might be argued that since stereopsis definitely assigns the edges all to a plane, the figure should be seen in two-dimensions and not in three. I think it is best to regard all contours in the 2½-D sketch as trying for a three-dimensional interpretation. The fact that the contours are put there by stereopsis rather than by, say, the primal sketch is unimportant.

CHAPTER 5

Representing Shapes for Recognition

5.1 INTRODUCTION

We come now to the final and perhaps most fascinating of the steps in our overall program, the transformation of shapes from a representation that is matched to the processes of perception into a representation that is suitable for recognition. There are many issues to be explored here, and this chapter, which rests heavily on Marr and Nishihara (1978), touches only the surface of some of them. Nevertheless, the main ideas are once more clear in outline, and I shall emphasize exactly what creating a shape representation that is suitable for recognition entails. This involves us in a discussion of what recognition is and how it comes about.

The single most important point is that we must now abandon the luxury of a viewer-centered coordinate frame on which all representations discussed hitherto have been based because of their intimate connection with the imaging process. Object recognition demands a stable shape description that depends little, if at all, on the viewpoint. This, in turn, means that the pieces and articulation of a shape need to be described not

relative to the viewer but relative to a frame of reference based on the shape itself. This has the fascinating implication that a canonical coordinate frame* must be set up within the object *before* its shape is described, and there seems to be no way of avoiding this. For some shapes, like a cigar, it will be easy to do this, and for others, like a crumpled newspaper, it will not.

Let us therefore look at these questions in detail. I shall reserve the term *shape* for the geometry of an object's physical surface. Thus, two statues of a horse cast from the same mold have the same shape. A *representation* for shape is a formal scheme for describing shape or some aspects of shape together with rules that specify how the scheme is applied to any particular shape. I shall call the result of using a representation to describe a given shape a *description* of the shape in that representation. A description may specify a shape only roughly or in fine detail.

5.2 ISSUES RAISED BY THE REPRESENTATION OF SHAPE

There are many kinds of visually derivable information that play important roles in recognition and discrimination tasks. Shape information has a special character, because unlike color or visual texture information, the representation of most kinds of shape information requires some sort of coordinate system for describing spatial relations. For example, the information that distinguishes the different animal shapes in Figure 5–1 is the spatial arrangement, orientation, and sizes of the sticks. Similarly, since left and right hands are reflections of each other in space, any description of the shape of a hand that is sufficient for determining whether it is left or right must in some manner specify the relative locations of the fingers and thumb.

Criteria for Judging the Effectiveness of a Shape Representation

There are many different aspects of an object's shape, some more useful for recognition than others, and any one aspect can be described in a number of ways. Although formulating a completely general classification

*A coordinate frame uniquely determined by the shape itself.

of shape representations is difficult, we can attempt to set out the main criteria by which they may be judged and the basic design choices that have to be made when formulating a representation.

Accessibility

Can the desired description be computed from an image, and can it be done reasonably inexpensively? There are fundamental limitations to the information available in an image—for example, regarding its resolution—and the requirements of a representation have to fall within the limits of what is possible. Moreover, a description that is in principle derivable from an image may still be undesirable if its derivation involves unacceptably large amounts of memory or computation time.

Scope and uniqueness

What class of shapes is the representation designed for, and do the shapes in that class have canonical descriptions in the representation? For example, a shape representation designed to describe planar surfaces and junctions between perpendicular planes would have cubical solids within its scope, but would be inappropriate for describing a billiard ball or a comb. If the representation is to be used for recognition, the shape description must also be unique; otherwise, at some point in the recognition process, the difficult problem would arise of deciding whether two descriptions specify the same shape. If, for example, we chose to represent shape using polynomials of degree n, the formal description of a given surface would depend on the particular coordinate system chosen. Since we would be unlikely to use the same coordinate system on two different occasions without observing some additional conventions, even the same image of a surface could give rise to very different descriptions.

Another example would be to represent a shape by a large collection of small cubes, packed together so as to approximate the shape as closely as possible. If the cubes were sufficiently small, the shape could be approximated quite accurately so that the scope of such a representation would be quite broad. On the other hand, a small shift of, say, half the side of a ⅛-in "minicube" could significantly change the representation of a shape, thus violating the uniqueness condition. If we used 1-ft cubes instead, the uniqueness problem would be greatly alleviated (a human might be represented by just six of them stacked up), but at considerable cost to other aspects of the representation.

Stability and sensitivity

Beyond the above scope and uniqueness conditions lie questions about the continuity and resolution of a representation. To be useful for recognition, the similarity between two shapes must be reflected in their descriptions, but at the same time even subtle differences must be expressible. These opposing conditions can be satisfied only if it is possible to decouple stable information that captures the more general and less varying properties of a shape from information that is sensitive to the finer distinctions between shapes.

For example, consider a stick figure representation that uses the three-dimensional arrangement and the relative size of sticks as primitive elements to describe animal shapes, as in Figure 5–1. The size of the sticks used gives one control over the stability and sensitivity of the resulting stick figure description. Stability is increased by using larger sticks; a single stick provides the most stable description of the whole shape, describing only its size and orientation. A description built of smaller sticks, on the other hand, would be sensitive to smaller, more local details, such as the extremities of an animal's limbs. Although such details tend to be less stable, they can nevertheless be important for making fine distinctions between similar shapes.

Choices in the Design of a Shape Representation

We can now relate the effects of different designs of shape representation to our three performance criteria. It is worth repeating once more that the most fundamental property of a representation is that it can make some types of information explicit, and this property can be used to bring the essential information to the foreground allowing smaller and more easily manipulated descriptions to suffice. We shall consider three aspects of a representation's design here: (1) the representation's coordinate system; (2) its primitives, which are the primary units of shape information used in the representation; and (3) the organization that the representation imposes on the information in its descriptions.

Coordinate systems

The most important aspect of the coordinate system used by a representation is the way it is defined. If locations are specified relative to the viewer, we say the representation uses a viewer-centered coordinate system. If locations are specified in a coordinate system defined by the viewed object,

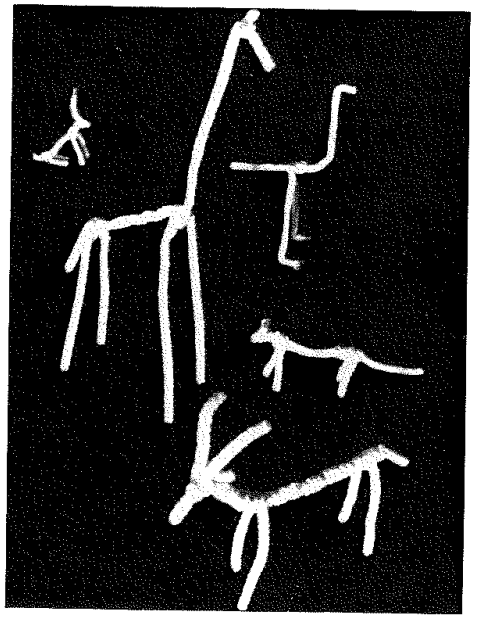

Figure 5–1. These pipe cleaner figures illustrate several of the points developed in this chapter. A shape representation does not have to reproduce a shape's surface in order to describe it adequately for recognition; as we see here, animal shapes can be portrayed quite effectively by the arrangement and relative sizes of a small number of sticks. The simplicity of these descriptions is due to the correspondence between the sticks shown here and natural or canonical axes of the shapes described. To be useful for recognition, a shape representation must be based on characteristics that are uniquely defined by the shape and that can be derived reliably from images of it. (Reprinted by permission from D. Marr and H. K. Nishihara, "Representation and recognition of the spatial organization of three-dimensional shapes," *Proc. R. Soc. Lond. B 200,* 269–294.)

the representation uses an object-centered coordinate system. There are, of course, several versions of each type.

For recognition tasks, viewer-centered descriptions are easier to produce but harder to use than object-centered ones, because viewer-centered descriptions depend upon the vantage point from which they are built. As a result, any theory of recognition that is based on a viewer-centered rep-

resentation must treat distinct views of an object essentially as distinct objects. Thus this approach requires a potentially large store of descriptions in memory in exchange for a reduction in the magnitude and complexity of the computations required to compensate for the effects of perspective.

Minsky (1975) has suggested that this number of descriptions might be minimized by choosing appropriate shape primitives and views to be stored in memory. Clearly much can be accomplished by this approach in some circumstances. For example, suppose squirrels need to distinguish trees from other objects but do not need to identify particular trees by their shape. They may be able to note some general characteristics of the appearance of a vertical tree trunk on the ground nearby that do not depend on the vantage point. In a representation based on these characteristics, all trees in the squirrel's environment would produce essentially the same description.

For more complex recognition tasks involving the arrangement of an object's components, however, any viewer-centered representation is likely to be sensitive to the object's orientation. For example, consider the many orientation-dependent appearances of a human hand, even if the fingers and thumb remain fixed with respect to each other. In order to distinguish a left hand from a right by using a viewer-centered representation, this problem would have to be treated as many separate cases, one for each possible appearance of a hand.

The alternative to relying on an exhaustive enumeration of all possible appearances is to use an object-centered coordinate system and thus to emphasize the computation of a canonical description that is independent of the vantage point. Ideally, only a single description of each object's spatial structure would have to be stored in memory in order for that object to be recognizable from even unfamiliar vantage points. However, an object-centered description is more difficult to derive, since a unique coordinate system has to be defined for each object, and, as I mentioned earlier, that coordinate system has to be identified from the image before the description is constructed.

Primitives

The primitives of a representation are the most elementary units of shape information available in the representation, which is the type of information that the representation receives from earlier visual processes. For instance, the 2½-D sketch is an example of a representation whose primitives carry information about local surface orientation and distance (relative to the viewer) at thousands of locations in the visual field. We can separate two aspects of a representation's primitives; the type of shape information they carry, which is important for questions of accessibility, and their size, which is important for questions of stability and sensitivity.

There are two principal classes of shape primitives, surface-based (two-dimensional) and volumetric (three-dimensional). As we have already seen, surface information is more immediately derivable from images. The simplest primitives useful for surface descriptions would specify just the location and size of small pieces of surface. More elaborate surface primitives like those used in the 2½-D sketch could include orientation and depth information as well.

On the other hand, volumetric primitives carry information about the spatial distribution of a shape. This type of information is more directly related to the requirements of shape recognition than information about a shape's surface structure, and this often means that much shorter and therefore more stable descriptions can still satisfy the sensitivity criterion. The simplest volumetric primitive specifies just a location and a spatial extent, and corresponds to a roughly spherical region in space. By adding a vector to this information, a roughly cylindrical region can be specified, whose length is indicated by the length of the vector and whose diameter is indicated by the spatial extent parameter of the primitive. A second vector could indicate a rotational orientation about the first vector, making it possible to specify a pillow-shaped region whose cross section along the first vector is thicker in the direction of the second vector. The additional vector could alternatively be used to specify the direction and magnitude of a curvature in the axis of the cylindrical region.

The complexity of the primitives used by a representation is limited largely by the type of information that can be reliably derived by processes prior to the representation. While the information-carrying capacity of primitives can be increased arbitrarily, there is a limit to the amount that is useful, since very detailed primitives will be derived less consistently by those earlier processes. In the extreme case, descriptions in a shape representation would consist of a single primitive. Such a representation would satisfy the uniqueness and stability conditions only if the information carried by the primitive was derived consistently by the processes supplying it. If this were so, however, those processes would already have accomplished shape recognition in specifying the primitive, and there would be no need for the representation.

Size is the other aspect that influences the information that the representation's primitives make explicit. In particular, information about features much larger than the primitives used is difficult to access, since it is represented only implicitly in the configuration of a larger number of smaller items. For example, consider how the arm of the human shape would be described in a surface representation like the 2½-D sketch. The

representation here is essentially what one would get by covering the surface with fish scales, each specifying a local surface orientation. Only information about small patches of surface is present, so a rather sophisticated analysis of a large assembly of these patches is required to make explicit the presence of the arm shape itself. A stick figure representation, on the other hand, can specify an arm explicitly with a single stick primitive of the appropriate size. Similar arguments can be applied to the representation scheme based on small cubes, discussed earlier; larger-scale shape information is not immediately available from such a representation.

At the other end of the scale, features of a shape that are much smaller than the primitives used to describe it are not just inaccessible, they are completely omitted from the description. For example, the fingers of a human shape are not expressible in a stick figure description that uses only primitives the size of the arms and legs. And even the arms and legs would be inexpressible in terms of 1-ft cubes. Similarly, surface details much smaller than the basic surface primitives used in the 2½-D sketch would be inexpressible in that representation. Thus the size of the primitives used in a description determines to a large degree the kind of information made explicit by a representation, the information made available but not directly obtainable, and the information that is discarded.

Organization

The third design dimension is the way shape information is organized by a representation. In the simplest case, no organization is imposed by the representation and all elements in a description have the same status. The local surface representation provided by the 2½-D sketch is one such example, and another would be our pile of minicubes that approximates a three-dimensional shape.

Alternatively, the primitive elements of a description can be organized into modules consisting, for example, of adjacent elements of roughly the same size, in order to distinguish certain groupings of the primitives from others. A modular organization is especially useful for recognition because it can make sensitivity and stability distinctions explicit if all constituents of a given module lie at roughly the same level of stability and sensitivity.

5.3 THE 3-D MODEL REPRESENTATION

We have formulated the requirements for a representation for shape recognition in terms of the criteria of accessibility, scope and uniqueness, and stability and sensitivity. We concluded that the design of a suitable representation should involve an object-centered coordinate system, include but

perhaps not be limited exclusively to volumetric shape primitives, and impose some kind of modular organization on the primitives involved in a description. These choices have strong implications, and a limited representation, called the *3-D* (three-dimensional) *model representation,* can be defined quite directly from them.

Natural Coordinate Systems

Our first objective is to define a shape's object-centered coordinate system. If it is to be canonical, it must be based on axes determined by salient geometrical characteristics of the shape, and conversely, the scope of the representation must be limited to those shapes for which this can be done. A shape's natural axes may be defined by elongation, symmetry, or even motion (for example, the axis of rotation); thus, the coordinate system for a sausage should be defined by its major axis and the direction of its curvature, and that of a face by its axis of symmetry. Objects with many or poorly defined axes, like a sphere, a door, or a crumpled newspaper, will inevitably lead to ambiguities. For a shape as regular as a sphere, this poses no great problem, because its description in all reasonable systems is the same. A door has four distinguished axes, defined by the directions of its length, its width, and its thickness and also by the axis on which it is hinged. Since the number of descriptions is small and doors are important, we could deal with each of the four possible descriptions of a door as a separate case. This would not be true of a crumpled newspaper, however, which is likely to have a large number of poorly defined axes.

At present, the problems we understand best are those involving the determination of axes based on a shape's elongation or symmetry (Marr, 1977a), and for the sake of simplicity we shall restrict the scope of the 3-D model representation to shapes that have natural axes of this type. One large class of shapes that satisfy this condition is the generalized cones, which we have already met and studied in Section 3.6 and illustrated in Figure 3–59. This class of shapes is important to us not because the surfaces are conveniently described—they may actually not be at all simple (Hollerbach, 1975)—but because such shapes have well-defined axes. This critical feature helps to define a canonical object-centered coordinate system, which is of course the central and most difficult task we face here.

In real life, a wide variety of common shapes is included in the scope of such a representation, because objects whose shape is achieved by growth are often described quite naturally in terms of one or more generalized cones. The animal shapes depicted in Figure 5–1 provide some examples—the individual sticks are simply axes of generalized cones that approximate the shapes of parts of these animals.

Axis-Based Descriptions

To be useful for recognition, a representation's primitives must also be associated with stable geometrical characteristics. The natural axes of a shape satisfy this requirement, and we shall therefore base the 3-D model representation's primitives on them. A description that uses axis-based primitives can be thought of as a stick figure, like those depicted in Figure 5–1, but one must be careful to think of the stick as a local coordinate axis. While only a limited amount of information about a shape is captured by such a description, that information is especially useful for recognition. We shall further limit the information carried by these primitives to pertain just to size and orientation. This will enable us to develop the 3-D model representation with a minimal commitment to inessential details. More elaborate details, such as curved axes or the tapering of a shape along the length of its axis, will not be included here.

The concept of a stick figure representation for shape is not new. Blum

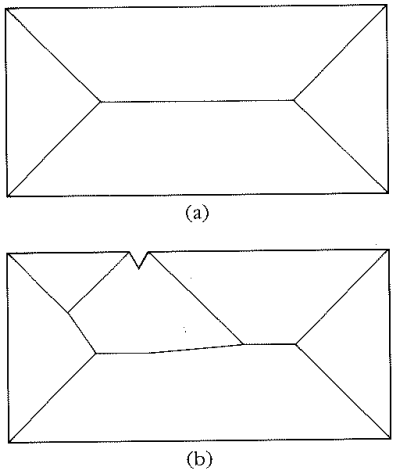

Figure 5–2. Blum's (1973) grassfire technique for recovering an axis from a silhouette. It can be thought of as lighting a fire at the boundary, the axis being defined as where two configurations meet. However, the technique is undesirably sensitive to small perturbations in the contour. (a) Shows the Blum transform of a rectangle, and (b) of a rectangle with a notch. (Reprinted by permission from G. Agin, "Representation and description of curved objects," Stanford Artificial Intelligence Project, memo AIM-173, Stanford University, Stanford, California.)

(1973), for example, has studied a classification scheme for two-dimensional silhouettes based on a "grassfire" technique for deriving a kind of stick figure from those shapes (see Figure 5–2), and Binford (1971) introduced the generalized cone for three-dimensional shapes. These representations have an important limitation, however; they do not impose a modular organization on the information they carry. For example, each part of the arm of a human shape can correspond to at most one stick in these representations; it would not be possible to have both a single stick corresponding to the whole arm and three smaller sticks corresponding to the major segments of the arm in the same description.

Modular Organization of the 3-D Model Representation

The modular decomposition of a description used for recognition must be well defined—such a decomposition must exist and it should be uniquely determined. In the 3-D model representation as specified so far, this is best achieved by basing the decomposition on the canonical axes of a shape. Each of these axes can be associated with a coarse spatial context that provides a natural grouping of the axes of the major shape components contained within that scope. We shall refer to a module defined this way as a *3-D model.* Thus, each 3-D model specifies the following:

1. A model axis, which is the single axis defining the extent of the shape context of the model. This is a primitive of the representation, and it provides coarse information about characteristics such as size and orientation about the overall shape described.

2. Optionally, the relative spatial arrangement and sizes of the major component axes contained within the spatial context specified by the model axis. The number of component axes should be small and they should be roughly the same size.

3. The names (internal references) of 3-D models for the shape components associated with the component axes, whenever such models have been constructed. Their model axes correspond to the component axes of this 3-D model.

Each of the boxes in Figure 5–3 depicts a 3-D model with the model axis on the left and an arrangement of the component axes on the right. The model axis of the human 3-D model makes explicit the gross properties (size and orientation) of the whole shape with a single primitive. The six component axes corresponding to the torso, head, and limbs can

each be associated with a 3-D model containing additional information about the decomposition of that component into an arrangement of smaller components. Although a single 3-D model is a simple structure, the combination of several in this kind of organizational hierarchy allows one to build up a description that captures the geometry of a shape to an arbitrary level of detail. We shall call such a hierarchy of 3-D models a *3-D model description* of a shape.

The example in Figure 5–3 illustrates the important advantages of a modular organization for a shape description. The stability of the representation is greatly enhanced by including both large and small primitive descriptions of the shape and by decoupling local spatial relations from

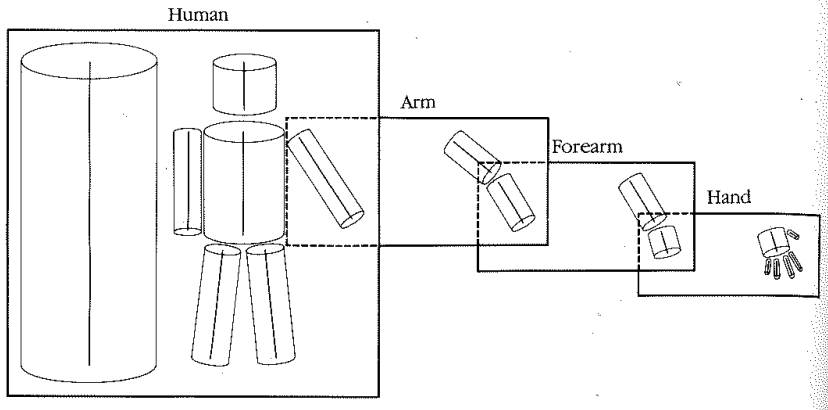

Figure 5–3. This diagram illustrates the organization of shape information in a 3-D model description. Each box corresponds to a 3-D model, with its model axis on the left side of the box and the arrangement of its component axes on the right. In addition, some component axes have 3-D models associated with them, as indicated by the way the boxes overlap. The relative arrangement of each model's component axes, however, is shown improperly, since it should be in an object-centered system rather than the viewer-centered projection used here (a more correct 3-D model is given by the table shown in Figure 5–5c). The important characteristics of this type of organization are: (1) Each 3-D model is a self-contained unit of shape information and has a limited complexity; (2) information appears in shape contexts appropriate for recognition (the disposition of a finger is most stable when specified relative to the hand that contains it); and (3) the representation can be manipulated flexibly. This approach limits the representation's scope, however, since it is only useful for shapes that have well-defined 3-D model decompositions. (Reprinted by permission from D. Marr and H. K. Nishihara, "Representation and recognition of the spatial organization of three-dimensional shapes," *Proc. R. Soc. Lond. B 200,* 269–294.)

more global ones. Without this modularization, the importance of the relative spatial arrangement of two adjacent fingers would be indistinguishable from that of the relation between a finger and the nose. Modularity also allows the representation to be used more flexibly in response to the needs of the moment. For example, it is easy to construct a 3-D model description of just the arm of a human shape that could later be included in a new 3-D model description of the whole human shape. Conversely, a rough but usable description of the human shape need not include an elaborate arm description. Finally, this form of modular organization allows one to trade off scope against detail. This simplifies the computational processes that derive and use the representation, because even though a complete 3-D model description may be very elaborate, only one 3-D model has to be dealt with at any time, and individual 3-D models have a limited and manageable complexity.

Coordinate System of the 3-D Model

There are two kinds of object-centered coordinate systems that the 3-D model representation might use. In one, all the component axes of a description, from torso to eyelash, are specified in a common frame based on the axis of the whole shape. The other uses a distributed coordinate system, in which each 3-D model has its own coordinate system. The latter is preferable for two main reasons. First, the spatial relations specified in a 3-D model description are always local to one of its models and should be given in a frame of reference determined by that model for the same reasons that we prefer an object-centered system over a viewer-centered one. To do otherwise would cause information about the relative dispositions of a model's components to depend on the orientation of the model axis relative to the whole shape. For example, the description of the shape of a horse's leg would depend on the angle that the leg makes with the torso. Second, in addition to this stability and uniqueness consideration, the representation's accessibility and modularity is improved if each 3-D model maintains its own coordinate system, because it can then be dealt with as a completely self-contained unit of shape description.

The coordinate system for specifying the relative arrangement of a 3-D model's component axes can be defined by its model axis or by one of its component axes. We shall refer to the axis chosen for this purpose as the model's *principal axis.* For the examples given here, the principal axis will be the component axis that meets or comes close to the largest number of other component axes in the 3-D model (for example, the torso of an animal shape). The location of the principal axis must also be spec-