and not in the orthographic projection, does not yield a clear perception of three-dimensional structure in the way that other motions do (Ullman, 1979a). The structure-from-motion scheme is an essentially local one, however, since it operates off nuclei of only four points. Even the perspective projection is locally orthographic, so there are no practical difficulties involved in using orthographic deprojection techniques like Ullman's scheme even in the real-life perspective case.

Optical Flow

J. J. Gibson has long believed that "the fundamental visual perception is that of approach to a surface. This percept always has a subjective component as well as an objective component, *i.e.* it specifies the observer's position, movement, and direction as much as it specifies the location, slant and shape of the surface" (1950). Sixteen years later he enunciated similar opinions and illustrated them with Figure 3–55 (1966, fig. 9.3).

The mathematics of this situation began to be studied quite soon, but only for special cases or for particular aspects of the general case (Gibson, Olum, and Rosenblatt, 1955; Lee, 1974; Clocksin, 1978). Nakayama and Loomis (1974) showed how depth contours may be extracted from a representation of the retinal velocity field induced by motion of the observer. Only recently, however, has a general treatment of the problem appeared (Longuet-Higgins and Prazdny, 1980).

The optical flow problem, as I shall employ the term, is the use of the retinal velocity field induced by motion of the observer to infer the three-dimensional structure of the visible surfaces around him. These visible surfaces are assumed to be stationary. The principal difference from Ullman's approach is that the optical flow effects rely on the polar projection, whereas the structure-from-motion approach is inherently orthographic. Thus, the optical flow approach can in principle deal with planar surfaces, on which the structure-from-motion approach necessarily fails.

The input representation

The information, called optical flow, on which our analysis is to operate can be thought of as the instantaneous positional velocity field (Gordon, 1965), which associates with each element on the retina the instantaneous velocity of that element. As usual, these elements are to be thought of as having some physical meaning.

This information is by no means as simple to acquire as optical flow devotees sometimes seem to assume. We have already seen in Section 3.4

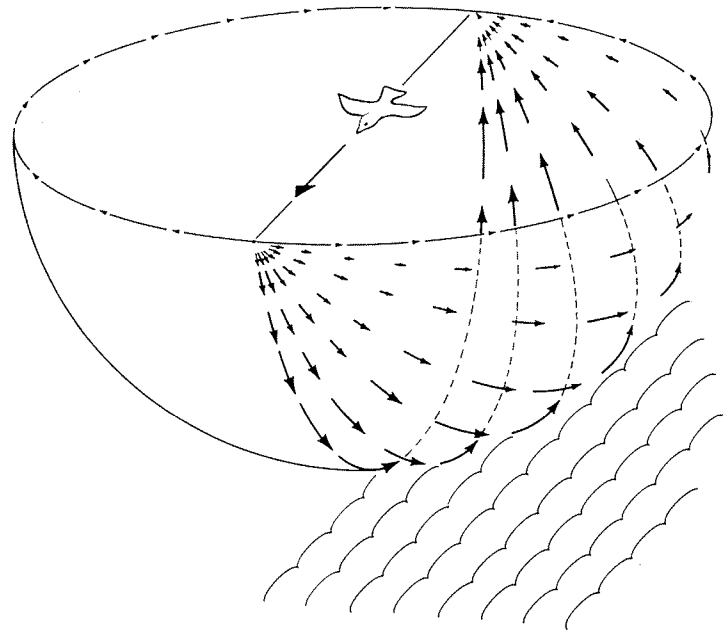


Figure 3–55. Gibson's example of flow induced by motion. The arrows represent angular velocities, which are zero directly ahead and behind. (Reprinted from J. J. Gibson, *The Senses Considered as Perceptual Systems,* Houghton Mifflin, Boston, 1966, fig. 9.3. Copyright © 1966 Houghton Mifflin Company. Used by permission.)

that local measurements alone can give little more than the direction of movement because of the aperture problem. In fact, fully specifying the optical flow is equivalent to solving the simpler of the two correspondence problems in apparent motion, since knowing the flow field enables one to establish the correct correspondences between two frames photographed in sufficiently rapid succession. Hence if optical flow analysis is carried out by our visual systems, it must rely on an input of the same sort that feeds the structure-from-motion computations.

Mathematical results

If an observer is approaching a stationary surface on a linear trajectory, the point of impact is the singularity in the optical flow field, and the time to

impact depends only on the angular velocities in the field (Koenderink and van Doorn, 1976). It is doubtful whether these facts are much used by our visual systems, since Johnston, White, and Cumming (1973) simulated optical expansion during approach to a surface and showed that human observers could reliably locate the focus of expansion only immediately prior to apparent impact with the surface. When teaching a pupil to land an airplane, a flying instructor will spend some time explaining that the current estimated landing point is the focus of expansion. This requires concentration and learning, for it is not a natural reflex. So Gibson's (1958) hypothesis that the center of optical expansion plays a major role in the control of locomotion is probably false for humans, although it may be more relevant to birds.

An authoritative account of the mathematics of optical flow has appeared only recently (Longuet-Higgins and Prazdny, 1980; Prazdny, 1979). It showed that from a monocular view of a rigid, textured, curved surface, it is possible in principle to determine the gradient of the surface at any point, the motion of the eye relative to that surface from the velocity field of the changing retinal image, and the field's first and second spatial derivatives. The relevant equations are redundant, thus providing a test of the rigidity assumption.

There is an interesting contrast between this result and Ullman's structure-from-motion theorem. In Ullman's scheme, four points are sufficient provided that the observer waits long enough to obtain at least three distinct views of them. Longuet-Higgins and Prazdny's scheme makes a slightly different trade-off; only two frames are required, so the time needed to acquire the measurements can be shorter. (Two frames suffice here because shape recovery is based on the perspective, not orthographic projection.) On the other hand, the local spatial neighborhoods involved in the computation are not just points, as in Ullman's scheme; they have to be large enough to give reliable estimates of the first and second spatial derivatives of the velocity field.

This analysis is another example of how computational theory can help empirical investigation. By solving the mathematics of the problem—and this was surely long overdue—Longuet-Higgins and Prazdny have provided a framework within which to inquire whether we humans actually do make use of optical flow, as Gibson suggested, and if we do, how. It is already clear that there are some ways in which we might have made use of it but actually do not. Attributing importance to the focus of expansion of retinal flow is one thing we could do but apparently do not. Another example is Ullman's conveyor belt demonstration, illustrated in Figure 3–54. We do not see regions 1 and 3 as having a different geometry from region 3, whereas most optical flow theories would say that we should.

Nevertheless, we could still use optical flow in some form, perhaps only weakly and more in peripheral than in central vision. That is, after all, where we might expect precision of measurement to be too low for a system based on Ullman's structure-from-motion scheme, yet it is also where we would expect to find the most evident optical flow. It remains to be seen whether optical flow is used in human vision.

3.6 SHAPE CONTOURS

As we discussed in Chapter 2, when we inquired into the physical basis for the primal sketch, there are four basic ways in which contours can arise in an image. They are (1) discontinuities in distance from the viewer, (2) discontinuities in surface orientation, (3) changes in surface reflectance, and (4) illumination effects like shadows, light sources, and highlights. Earlier in this chapter, we saw how different aspects of the primal sketch can be used as the input representation for processes based on stereopsis or on motion that are capable of finding boundaries from the differences between two or more images of a scene. We turn now to the more difficult case of a single, monocular image and ask how its contours can convey unambiguous information about shape. The mystery that needs explaining is that contours in an image are two-dimensional, yet we often see them in three dimensions. The question is how and why we make this three-dimensional interpretation.

I call the contours that we shall examine *shape contours,* because they are all two-dimensional contours that yield information about three-dimensional shape. I shall not discuss at all how to find them in an image—we spent long enough on that task in Chapter 2. Nevertheless, it is worth pointing out that although the physical origins of contours can be divided into the four categories mentioned, these origins give rise to a wide range of detectable changes in the image and hence a wide variety of ways in which a particular type of contour may be defined in the image.

For example, consider the possible effects of a discontinuity in depth. This can cause a simple intensity change—in fact, since our visual systems incorporate a predisposition for seeing brighter things as nearer, we would expect this brightness versus depth relation to be generally true of the visual world. If the surface characteristics are the same on both sides of a depth change, then a density- or size-induced texture boundary will be formed. If the two surfaces are not the same object, their textures will usually be very different and so many criteria will yield the boundary.

If the discontinuity is a change in surface orientation, intensity is likely

to change and so is any density measure that the surface reflectance function may happen to support. Any clear orientation organization on the surface will also probably be shifted, and perhaps also some length measures.

If the surface reflectance is organized in any of a number of ways—for example, if it contains parallel lines—then it can convey valuable shape information to the viewer. And so forth.

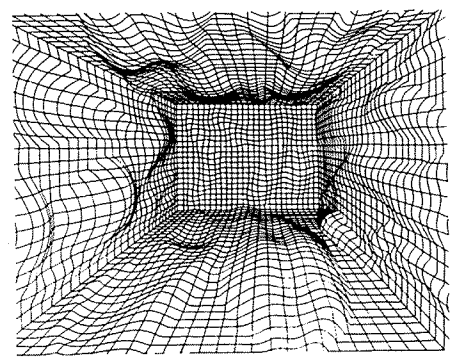
The main point, then, is that contours can be defined on a surface in many ways, and they should be detected in the initial analysis and representation of the image. Some of these contours are more likely to have been caused by one kind of change than another—a discontinuity in orientation, for example, is more likely to be due to a change in surface orientation than to a change in depth—but the rules are not hard and fast. The important fact is that very many such contours can and do tell us about three-dimensional shape, and when one reflects upon it, this is actually quite an amazing fact. Such shape contours form the focus of our interest in this section.

Some Examples

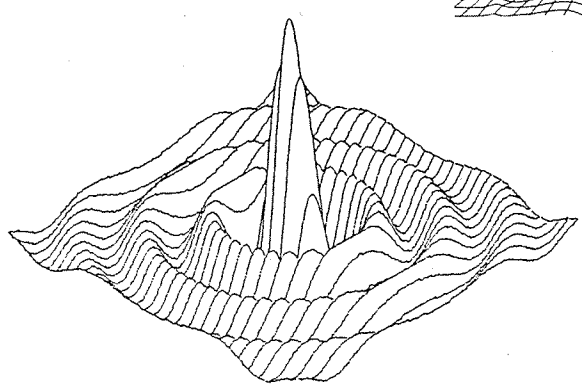
The power and vividness with which contours can depict shape is not in doubt. Figure 3–56 shows some examples, and I think the reader will agree that for sheer three-dimensional realism, Figures 3–56(b) and (c) approach the effects achieved by means of stereopsis or motion. Very much more in doubt than in these other cases is precisely how these examples create this realism. Contours in an image can arise from several distinct physical causes. Some, as in Figure 3–56(a), are occluding contours—contours that arise at a discontinuity in depth, here at the edge of the viewed objects. Other contours arise from changes in surface orientation, texture boundaries, changes in reflectance and pigmentation, or from shadows falling on a surface. Most vivid and puzzling are the contours in Figures 3–56(b) and 3–56(c). To what do these correspond in nature? After all, we rarely come upon objects created by deforming a rectangular wire grid, as in Figure 3–56(b). Why then are we so good at seeing the shape of the wire room depicted there? Is it the same reason why we can see Figure 3–56(c) so well? Is there just one basic trick involved here or the happy coincidence of several that conspire and are jointly responsible for the vividness of the percept?

These, then, are the questions that we shall be studying here. Unfortunately, because we do not yet know whether one phenomenon or several are operating in cases like Figures 3–56(b) and (c), we are not in so strong

(a)



(b)



(c)

Figure 3–56. Examples of two-dimensional contours in an image that impart three-dimensional information to the viewer. (a) *Rites of Spring* by Picasso, an example of shape information from silhouettes. (b) A "wire room." (c) A portrayal of the curve sin x. (b) and (c) are especially vivid. (Part (a) Copyright © SPADEM, Paris/VAGA, New York 1981. Part (b) courtesy of the Carpenter Center for the Visual Arts, Harvard University.)

a position as we were with stereopsis and motion. Psychophysics has not yet told us what the modules are, so we are still stuck in something of the linguist's predicament of not yet having a clear decomposition of language into relatively independent modules.

Nevertheless, some progress has been made. It is convenient to divide our discussion into three categories: (1) contours that occur at discontinuities in the distance of the surface from the viewer (occluding contours), (2) contours that follow discontinuities in surface orientation, and (3) contours that lie physically on the surface. This third type of contour can be due to surface markings or to shadow lines, for example. The important point is that they lie along the surface, and therefore I call them surface contours. Remember that contours in each category can be detected in several ways in an image. In all cases, our principal question is, Why and how can such contours in a single two-dimensional image convey to us unambiguous and often quite detailed information about three-dimensional shape?

Occluding Contours

An occluding contour is simply a contour that marks a discontinuity in depth, and it usually corresponds to the silhouette of an object as seen in two-dimensional projection. I became interested in occluding contours from the observation—which is almost a paradox—that when we look at the silhouettes in Picasso's *Rites of Spring* (reproduced here in Figure 3–56a), we perceive them in terms of very particular three-dimensional shapes, some familiar, some less so. This is quite remarkable, because the silhouettes could, in theory, have been generated by an infinite variety of three-dimensional shapes, which, from other viewpoints, would have no discernible similarities to the shapes that we perceive. It takes only a little imagination and moderate mischief to concoct a quite bizarre three-dimensional shape to demonstrate this point. We might, for example, arrange spikes and protuberances in a highly baroque style that happen to combine unexpectedly to produce the silhouette of a man or a goat when viewed from one special direction.

Yet we never think of such things when we are faced with these silhouettes. One can perhaps attribute part of the phenomenon to a familiarity with the depicted shapes, but not all of it, because we can use a silhouette to convey an unfamiliar shape, and because even with considerable effort it is difficult to imagine the more bizarre three-dimensional surfaces that could have given rise to the silhouettes in Picasso's painting. The paradox, then, is that the bounding contours in *Rites of Spring* apparently tell us more than they should about the shapes of the figures. For example, neighboring points on the bounding contours here could arise from widely separated points on the original surface, but our perceptual interpretation usually ignores this possibility.

This situation is so reminiscent of ignoring the many possible snowstorm interpretations of random-dot stereograms or the two-cylinder, random-dot moviegrams that one is almost forced to draw the obvious conclusion: Somewhere buried in the perceptual machinery that can interpret silhouettes as three-dimensional shapes, there must lie some source of additional information that constrains us to see the silhouettes as we do. Probably, but perhaps slightly less certainly than in the analyses of motion and stereopsis, these constraints are general rather than particular and do not require a priori knowledge of the viewed shapes.

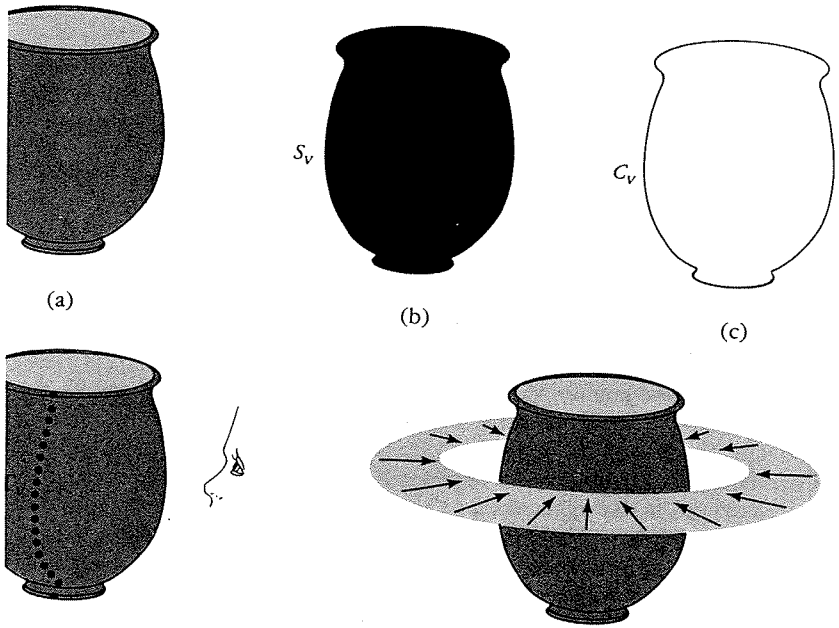
If these constraints are general, then there must be some a priori assumptions in the way we interpret silhouettes that allow us to infer a shape from an outline. These assumptions must pertain to the nature of the viewed shape. Moreover, if a surface violates these implicit assumptions, then we should see it wrongly. Our perceptions should deceive us in the sense that the shape we assign to the contours will differ from the shape that actually caused them. One common instance is the shadowgraph, where the appropriate arrangement of the hands can, to the surprise and delight of a child, produce the shadow of an objectively quite different three-dimensional shape, like a duck, rabbit, or ostrich.

Constraining assumptions

The question we have to ask is, What assumptions are reasonable to make—that we unconsciously employ—when we interpret silhouettes like those of Figure 3–56(a) or Figure 3–57(b) as three-dimensional shapes?

Three seem to be important (Marr, 1977a). The first is that *each line of sight from the viewer to the object should graze the object's surface at exactly one point.* In other words, each point on the silhouette (Figure 3–57b) should correspond to one point on the viewed surface (Figure 3–57a). The reason for assuming this is that even if this correspondence did not exist, we could not possibly tell that it did not, and it would usually happen only as a result of an accidental alignment of two parts of the object along the line of sight.

This assumption allows us to speak of a particular curve on the object's surface called the *contour generator,* illustrated in Figure 3–57(d). It is the set of points on the surface that projects to the boundary of the silhouette in the image, and I shall use the letter Γ to denote it.



:57. Four structures of importance in studying the a priori conditions that we bring to ıe analysis of an occluding contour. (a) A three-dimensional surface, Σ. (b) Its silhouette from viewpoint V. (c) The contour C_V of S_V. (d) The set of points Γ_V that project onto the e) A condition for the theorem discussed in the text. In particular, the meaning of "all wing positions in any one plane" is shown.

The second assumption says that, except possibly in a very few instances, points that appear to be close together in the image actually are close together on the object's surface. The illustration in Figure 3–58(a) helps to explain this assumption. Think of a and b as being two hills, with the contour generators that give rise to a and b following the skyline on the top of each hill. If the dashed portion of b happens to be invisible, then at point P the visible contour generator leaps from one hill to the next—it is discontinuous. The sharp concavity at P, in fact, hints of this discontinuity, and so we half expect it. In the body of a and b, however, we do not expect it to happen, and in fact we assume it does not. This is our second assumption, and it says that *nearby points on the contour in an image arise from nearby points on the contour generator on the viewed object.*

The last assumption is a little more sophisticated, for it pertains to the

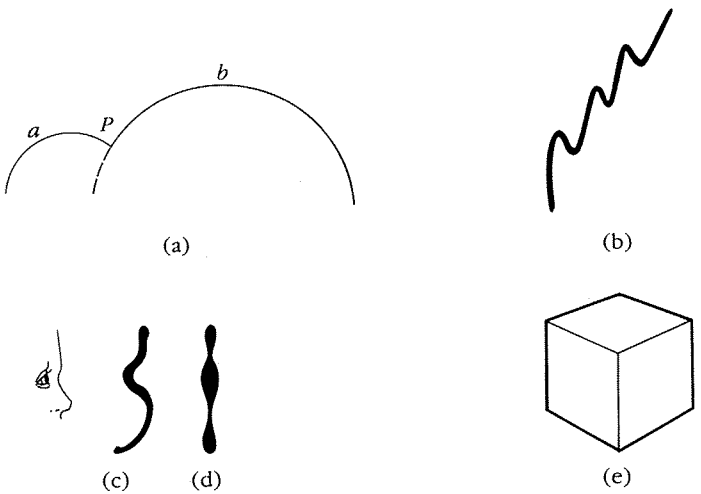


Figure 3–58. (a) The second assumption, that nearby points on a contour arise from nearby points on the contour generator, essentially says that there are no points like P on the contour. If the dotted portion of b were invisible, the contour generator would leap from a to b, causing a discontinuity at P. (b) A typical piece of contour. The only features we can hope to make use of are its convexities and concavities, that is, its points of inflection, and these must be properties of the surface and not of the imaging process. For example, if a viewer is close to a snake (c), the convexities and concavities in the image (d) arise not because of properties of the snake, but because of variations in its distance away. (e) If the occluding contour shown with thick lines is present on its own, one perceives a hexagon. The interior lines change it into a cube, since they suggest that the occluding contour is not planar.

type of clue that an image contour might give about shape. Suppose, for example, that we have been presented with a piece of contour like that shown in Figure 3–58(b). The previous two assumptions allow us to think of this contour as coming from a contour generator on the surface, and we can safely assume that adjacent points on the contour come from adjacent points on the contour generator. Because the imaging process is what it is, we cannot rely on any measurements that we make on the contour in the image, and so the only remaining straightforward feature is that sometimes the contour bends one way and sometimes the other. In other words, there is a qualitative distinction between convex and concave segments, which, provided that the surface is sufficiently smooth, rests in turn on the notion of an inflection point. In general, of course, points of inflection in a contour need have no significance for the surface. The contour generator could

weave around in an arbitrary and complex way, or it could move directly toward and then away from the viewer. The latter case might, under the perspective projection, give rise to convexities and concavities rather in the way illustrated by Figures 3–58(c) and (d). So our next question has to be, How exactly should we formulate an assumption saying that points of inflection in a contour are significant, that they somehow reflect real properties of the viewed surface and not artifacts of the imaging process?

Our previous two assumptions allow us to think of the contour generator as a piece of wire bent in three-dimensional space. If inflection points on the contour are to reflect genuine inflections on this piece of wire, however, two mathematical conditions must be satisfied:

1. The transformation due to the imaging process that produces the contour from the wire must be linear. This rules out the perspective transformation and restricts the validity of our theory to distant views—the object must be small relative to its distance from the viewer.

2. The curve on which the transformation acts must lie in a plane. In other words, the convex–concave distinction in the image can be meaningful only for distant views and only if the bent wire that is the contour generator lies in a plane. This gives us our third assumption, that *the contour generator is planar.*

This third assumption is a strong one that sharply delimits the class of surfaces whose shapes can be interpreted by silhouette. However, it seems unavoidable if we wish to distinguish convex and concave segments in the interpretation process. Fortunately, however, the results of using this assumption are very robust—if the contour generator is not quite planar but nearly so, then the surfaces are usually only a little misbehaved. And interestingly, the planar condition is actually embodied in much modern design. All of the outlines drawn in mechanical engineering diagrams satisfy the condition, so it has its uses even outside the study of vision. If the condition is violated, we do seem to get the shape wrong. The occluding contour in Figure 3–58(e), for example, is marked with thick lines and, if shown on its own, gives the appearance of a two-dimensional hexagon. With the additional information provided by the interior lines, however, it takes on a quite different interpretation. As a cube, the occluding contour is no longer planar.

Implications of the assumptions

In order to see what these assumptions really mean, we have to understand how they constrain the geometry of the surfaces being viewed. Clearly,

some surfaces will satisfy the assumptions and some will not. What about a surface makes it satisfy them? To answer this question we should reformulate our assumptions as restrictions on the geometry of the viewed surface and then see what their consequences are. To remind ourselves of these restrictions, I restate them here:

1. Each point on the contour generator projects to a different point on the contour.

2. Nearby points on the contour arise from nearby points on the contour generator.

3. The contour generator lies wholly in a single plane.

We need one more idea before we can formulate the critical result—the idea of a *generalized cone.* This idea was introduced by T. O. Binford (1971) as a way of representing shapes in a computer program, and it is illustrated in Figure 3–59. A generalized cone is the surface created by moving a cross section along an axis. The cross section may vary smoothly in size, getting fatter or thinner, but its shape remains the same. Thus a football is a generalized cone and so is a pyramid or, roughly, a leg or an arm, or a snake, or a tree trunk, or a stalagmite. In fact, we can think of a horse as being composed of eight generalized cones, one for each leg and one each for the head, neck, body, and tail.

We are now ready for the basic result, and I hope the reader finds it as surprising as I did:

If the surface is smooth (for our purposes, if it is twice differentiable with a continuous second derivative) and if restrictions 1 through 3 hold for all distant viewing positions in any one plane, as illustrated in Figure 3–57(e), then the viewed surface is a generalized cone. The converse is also true; if the surface is a generalized cone, then restrictions 1 through 3 will be observed.

This means that if the convexities and concavities of a bounding contour in an image are actual properties of a surface, then that surface is a generalized cone or is composed of several such cones. In brief, the theorem says that a natural link exists between generalized cones and the imaging process itself. The combination of these two must mean, I think, that generalized cones will play an intimate role in the development of vision theory.

Stated baldly, this result means that, in general, shape cannot be derived from occluding contours alone unless that shape is made from generalized cones and is viewed from a position from where its axis is not

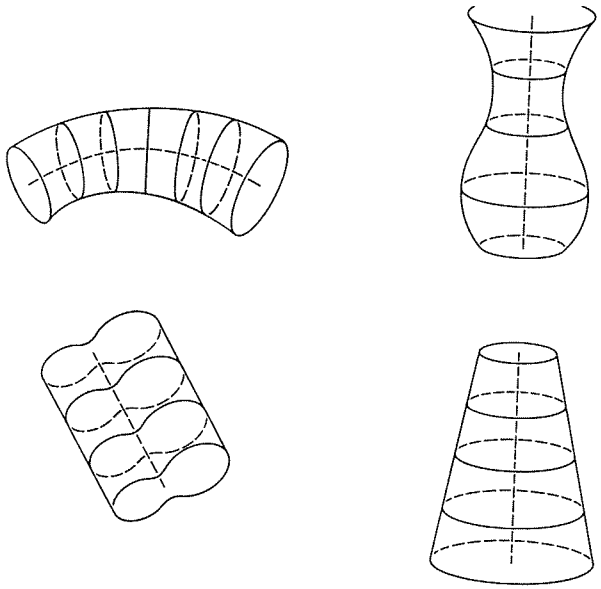


Figure 3–59. The definition of a generalized cone. As used in this book, the term *generalized cone* refers to the surface created by moving a cross section along a given smooth axis. The cross section may vary smoothly in size, but its shape remains constant. We here show several examples. In each, the cross section is shown at several positions along the trajectory that spins out the construction.

foreshortened (foreshortening would occur in Figure 3–57(e) if the vantage point was from above or below). If there is no foreshortening, however, and even if the viewed shape is constructed of several different generalized cones like the silhouette of a man or of a horse, then the shape can be at least partially reconstructed. Perhaps the most important thing, as we shall see later in the book, is that the *axes* of the cones can be recovered from the image, because this helps to establish an object-centered coordinate system in the viewed shape. I shall say more about this in Chapter 5 and will briefly illustrate an algorithm for decomposing silhouettes into their constituent generalized cones. (See Marr, 1977a, for the theorems behind the algorithm.)

For now, however, it is enough to note that the use of occluding contours requires the three restrictions that we formulated, and they hold if and only if the viewed shapes are generalized cones. The principal implication of these restrictions is that the surface goes in and out where the contour goes in and out. Not much more can be said from the occluding contours alone.

Surface Orientation Discontinuities

Surface orientation contours mark the loci of discontinuities in surface orientation. For example, they follow the creases on a surface, like the interior lines of Figure 3–58(e) or the longitudinal peaks and troughs of Figure 3–60. With regard to recovering the geometry of the surface, the most important question about such a contour is whether it corresponds to a convexity or a concavity on the surface. In Figure 3–58(e), all the interior contours represent convexities, whereas in Figure 3–60 convexities and concavities alternate, sometimes in an interestingly confusing way.

Unfortunately, it is often difficult to distinguish convexities and concavities from purely local cues in a monocular image. We have a predisposition to see such contours as convex (see Figure 3–61b), but even examples that are loaded one way can be made to alternate (compare Figures 3–61a and c).

There are certain things to be said about combinations of such contours—for example, Waltz-like (1975) constraints, of the form illustrated in Figure 1–3, apply, which specify that one cannot have two concave and

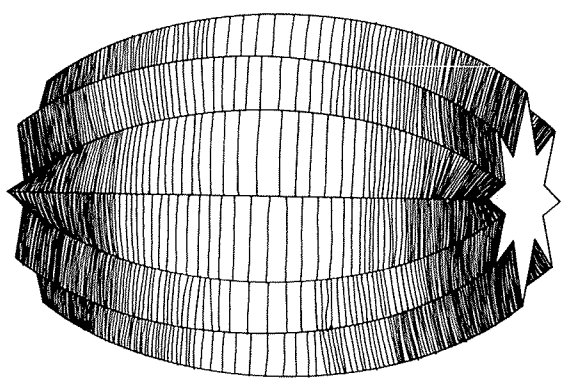


Figure 3–60. A sketch of a generalized cone showing its silhouette (the circumscribing contour) and fluting (the contours spanning its length). The fluting marks lines of discontinuity in surface orientation.

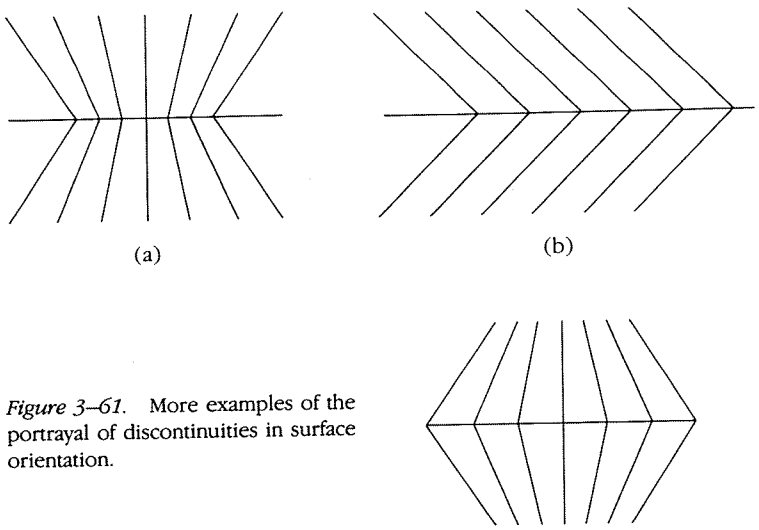


Figure 3–61. More examples of the portrayal of discontinuities in surface orientation.

one convex contour meeting at a point. However, these are not properties purely of the isolated contours, and I shall discuss that class of more complicated phenomena in Chapter 4. The only knowledge currently available for helping to distinguish isolated convex and concave contours is due to Horn (1977). He showed that, at least for the visual world of matte white prisms, the intensity profiles across different types of edge are characteristically different. If the intensity profile across the edge is a step change or very sharp peak, the edge is probably convex. If the intensity profile is roof shaped, the edge is probably concave. However, there is no evidence yet that the human visual system uses these cues in classifying edges.

Surface Contours

Surface contours arise for various reasons in the image of smooth surfaces, and they yield information about the three-dimensional shape of the surface, in the manner illustrated by Figure 3–62. The question of interest, of course, is how this is done, and it has recently been explored in some detail by Stevens (1979). The underlying observation is that we do not

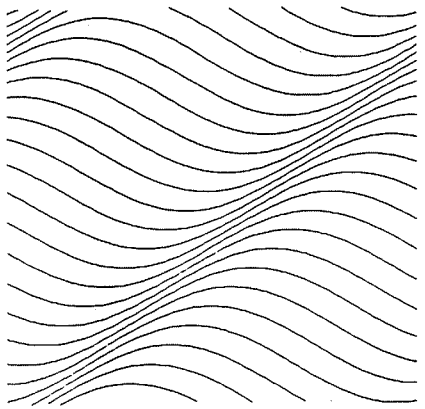


Figure 3–62. The undulating surface is suggested by a family of sinusoids. The curves are naturally interpreted as surface contours, that is, the images of markings on a physical surface. What constraints can be brought to bear in making this three-dimensional interpretation? (Reprinted by permission from K. Stevens, "Surface perception from local analysis of texture and contour," Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1979.)

perceive Figure 3–62 as purely two-dimensional; there is no doubt that what we see is a smooth, undulating surface. As we have seen many times now, this means that we are bringing some a priori assumptions to bear on our analysis of such images.

Once again, the fundamental computational questions are What are these assumptions, Why do we use them, and How do they enable us to recover three-dimensional surface orientation information from a single two-dimensional image? In this discussion of Stevens' work, I shall maintain the distinction between an image contour and its corresponding contour generator on the surface, which we met first in our analysis of occluding contours, illustrated in Figure 3–57. The difference here is that the contour generators are no longer restricted to just the silhouette boundaries of an object but may arise within the silhouette because of internal surface markings or various kinds of illumination effects. For example, the contours of Figure 3–62, are naturally interpreted as the image of markings on the surface, and we shall call these markings the contour generators of the image contours. These contours may, of course, be quite abstract objects, perhaps created by rows of dots, but we take the machinery and represen-

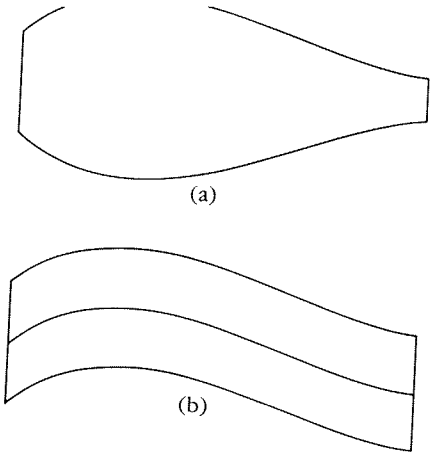


Figure 3–63. The curves in (a) are interpreted as occluding contours, and the underlying surface is seen as a generalized cone—in this case, a vaselike object. Such contours were studied in Section 3.5 and are further considered in this discussion. Those in (b) are interpreted as surface contours, and the surface appears like a gently curved flag or ruled sheet of paper. (Reprinted by permission from K. Stevens, "Surface perception from local analysis of texture and contour," Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1979.)

tation abilities of the full primal sketch for granted here. We shall call such contours *surface contours*. Note that occluding contours are almost never surface contours (see Figure 3–63).

The puzzle and difficulty of surface contours

What makes the issue of surface contours so extremely difficult to analyze satisfactorily is that there is no obvious physical source of surface contour regularity that our perceptual machinery can use to such advantage. The world really seems to have less structure than diagrams like Figure 3–62, and I remain deeply puzzled about why we can interpret such figures so vividly.

Stevens (1979), in a useful first approach to these issues, divided the problem into two halves; inferring the shape of the contour generator in three-dimensional space and then determining how the surface itself lies in relation to the contour generator. The first step is that of discovering the

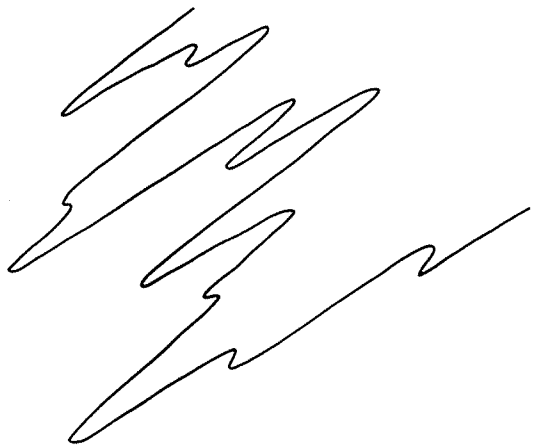


Figure 3–64. This curve appears to have a specific three-dimensional shape, as if planar and foreshortened by the slant of the plane relative to the viewer. Why and how is this interpretation derived? (Reprinted by permission from K. Stevens, "Surface perception from local analysis of texture and contour," Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1979.)

shape of a piece of wire bent in three-dimensional space so that it lies along the contour generator and has the correct appearance in the image. The second step can then be thought of as gluing a ribbon along the wire so that it follows faithfully the strip of surface that lies directly under the contour generator.

Determining the shape of the contour generator

When we observe a single contour, the curve appears to have a specific three-dimensional shape and to lie in a plane. The impression gained from Figure 3–64, for example, is of a planar curve whose plane has a definite, if somewhat weakly specified, slant and tilt. The assumption that the contour generator is planar greatly simplifies the problem, but it is difficult to be confident of such an assumption, although shadow boundaries cast by straight edges and certain types of surface reflectance organizations will often produce planar contour generators on a surface.

There are other assumptions that one might make. Stevens (1979) pointed out that much can be done if symmetry, even of only a rough or

skewed kind, is detected in the figure (see also Marr, 1977a). Witkin (1978) suggested that it is sometimes useful to assume that the real-life contour generator has the minimum possible curvature, the visible curvature of the image contour being derived in part from the imaging process. But these ideas are still ad hoc and disorganized.

The effects of more than one contour

The weakness of our perception of single contours like that of Figure 3–64 is probably related to the unsatisfactory lack of any realistic interpretive assumptions that one might bring to bear upon such perceptions. If there are several contours, however, the vividness of our perception is much enhanced, as in Figure 3–62. Except in very rare and accidental situations, if surface contours are parallel in the image, their contour generators are parallel on the surface.

That the contour generators are parallel so that one can be shifted across the surface onto its neighbor, leads to quite a powerful idea about how to recover surface orientation from surface contours. Parallel contour generators essentially mean that we can locally ignore the curvature of the surface in the direction of the shift. Technically, the surface is then developable. This means that the surface can be thought of locally as a cylinder, which is a surface with two principal curvatures, one of which is zero—the surface is flat in that direction.

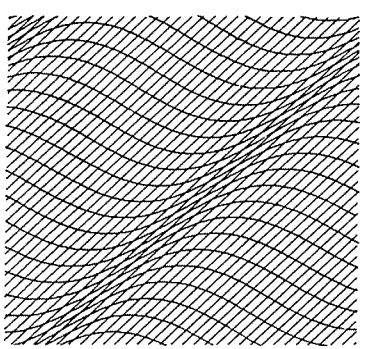


Figure 3–65. The wavy lines represent visible contours in the image, and the straight lines, which have zero curvature, make explicit the parallel relationships between adjacent wavy lines. Such a surface is locally a cylinder, because one of its curvatures (and hence its Gaussian curvature) is zero. (Reprinted by permission from K. Stevens, "Surface perception from local analysis of texture and contour," Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1979.)

The idea is illustrated by Figures 3–65 to 3–67. Figure 3–65 shows a surface in which two types of contours are visible—the wavy ones, which are the family of parallel contour generators that we suppose are in fact present in the image, and the orthogonal set of straight lines, which have zero curvature and represent the correspondence between the locally par-

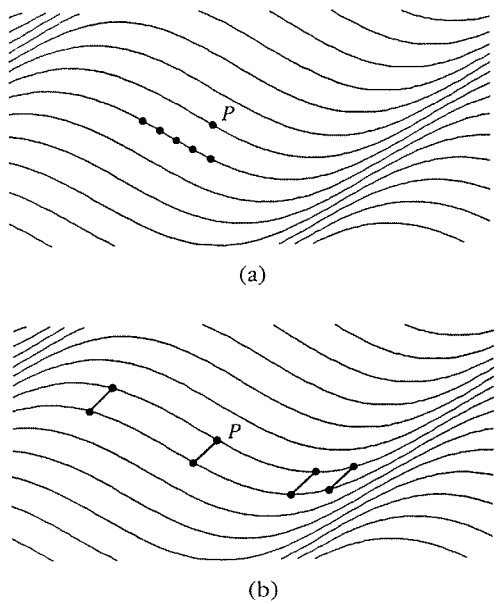


Figure 3–66. Usually, of course, the correspondences between adjacent parallel surface contours will not be explicit in the image, as they were in Figure 3–65. However, the correspondence can be found, even in the less straightforward cases. For example, if the surface contours are straight for a portion of their length, as in (a), the tangent to a point *P* on one contour may be parallel to various tangents on the adjacent contour; however, only one choice would result in a correspondence line that is parallel to the other correspondence lines between curved portions of adjacent contours, as in (b). (Reprinted by permission from K. Stevens, "Surface perception from local analysis of texture and contour," Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1979.)

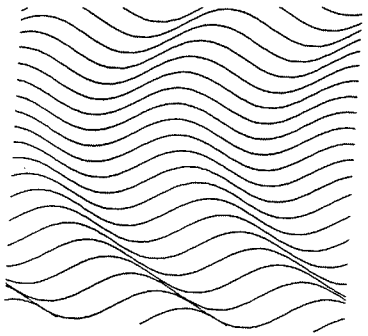


Figure 3–67. Although, strictly speaking, the assumptions and techniques illustrated in Figures 3–65 and 3–66 require that the surface be cylindrical, in practice they can be used assuming that they hold only locally, since the parallel correspondence need be established only between adjacent contours. Hence the local cylinder restriction allows us to interpret surfaces whose global structure is not cylindrical. (Reprinted by permission from K. Stevens, "Surface perception from local analysis of texture and contour," Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1979.)

allel contour generators. In identifying the correspondence with straight lines, we are assuming that the surface is locally of a peculiarly simple sort, with one of its curvatures being zero. Once both the wavy lines and the correspondence lines are available, surface orientation is quite well constrained, because we know that in three dimensions these two types of lines are perpendicular.

Usually, of course, the correspondence contours will not be visible in the image, but Figure 3–66 illustrates how they may be recovered, even in apparently ambiguous situations (some details are given in the legend). Finally, one can extend the idea to quite general surfaces, as explained by Figure 3–67, because the fundamental assumption on which the interpretation is based must hold only locally—in this case, between adjacent surface contours. Figure 3–67 shows an example of how this basic requirement—that one of the curvatures vanishes—holds only locally and approximately. The structure of the surface depicted there can be recovered by using methods based on these ideas, even though globally it is certainly not a developable surface.

Stevens pointed out one other interesting fact, namely, that if a highlight appears along a continuous curve on a surface, then the curve is planar (assuming that the light source and vantage points are distant from the surface). This contour is like one of our correspondence contours, along which one of the principal curvatures of the surface is zero. In this case the surface normal coincides with the normal to the plane containing the gloss contour, just as in Figure 3–65 the surface normal lies perpendicular to both the straight (correspondence) and wavy lines. So the conditions that Stevens suggested for the recovery of surface orientation from surface contours do actually occur in real life.

In summary, then, the recovery of surface orientation from surface contours remains an intriguing and unsolved problem. On the other hand, Stevens' main suggestions—the planarity of the contour generator and of the locally developable assumption—seem to be powerful ingredients for achieving the recovery, and I shall be surprised if they are not used by us in practice in some form.

3.7 SURFACE TEXTURE

The notion that surface texture may provide important information about the geometry of visible surfaces has attracted considerable attention in the last 30 years. Perhaps the main impetus for this interest was the hypothesis formulated by Gibson (1950), which states that texture is a mathematically and psychologically sufficient stimulus for surface perception. By this he meant that there is sufficient information in the monocular image of a textured surface to specify uniquely the distance to points on the surface and to specify the local surface orientation. Furthermore, he claimed that the human visual system can and does use this information to derive such surface information.

In an ideal world, where the surfaces are smooth and regularly and clearly marked and exhibit sufficient density of detail so that gradients in an image can be measured quite precisely, Gibson's claim would have much to recommend it. Unfortunately, however, the world is a much rougher place, in which uniformity and regularity are the exception or only an approximation rather than the rule, so my own view is that we should be surprised when something can be done rather than when it cannot. In addition, as Stevens (1979) has pointed out, much of the rather simple mathematics associated with these questions has had a somewhat flawed presentation in the past. We shall therefore be wise to take a critical and skeptical attitude to the supposed power of texture perception except when it can be demonstrated beyond doubt that the human visual system is using it.

The Isolation of Texture Elements

The first problem, and one that has hardly been addressed at all, is how to extract from an image the uniform texture elements on which subsequent analysis must rest. A full answer to this would include a complete understanding of the full primal sketch and of the selection by similarity, whose business it is to classify items by origin and whose importance we have already encountered (for example, Figure 2–3). Let us, however, take this for granted, and assume that the world's surfaces are covered with regular and sufficient markings, and that we are capable of discovering them from our early representations of the image.

Surface parameters

As we have already seen several times, there are two ways in which a surface may be specified relative to the viewer: We can either specify the distance to local pieces of it, or we can specify the surface orientation relative to the viewer. Surface orientation itself is naturally split into two components, which we have called slant and tilt. Slant is the angle by which the surface dips away from the frontal plane, and tilt is the direction in which the dip takes place.

Mathematically, of course, distance and surface orientation are almost equivalent, being related by an integration (see Chapter 4). For the nervous system, the question is a quite different one—Which of these quantities, distance, slant, or tilt, is actually extracted directly from measurements of variations in texture? In his recent study of this question, Stevens (1979) concluded as follows:

1. Tilt is probably extracted explicitly.
2. Probably distance is also extracted explicitly.
3. Slant is probably inferred by differentiating estimates of scaled distance made in accordance with point 2.
4. In particular, measurements of texture gradients, which are closely associated mathematically with slant, are probably not made or used, perhaps because of the inaccuracies inherent in the measuring process.

We look now at the reasons for his conclusions.

Possible measurements

Stevens observes that even very different looking textures pose the same computational problems, and that one must be careful not to postulate more mechanisms than the problems require. Figure 3–68 shows an

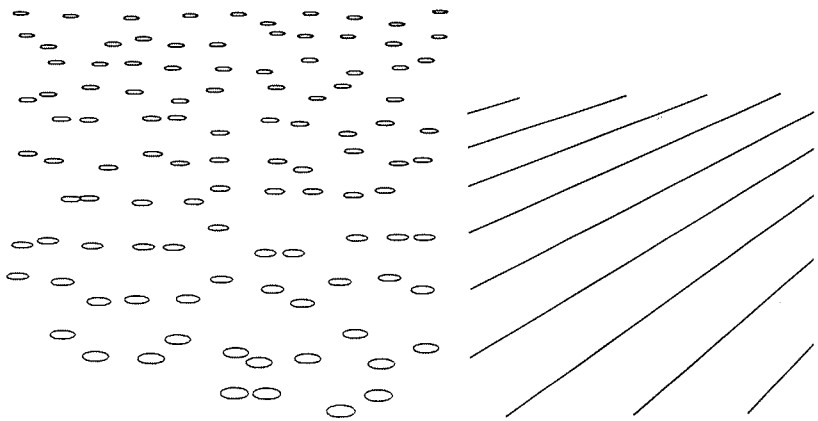
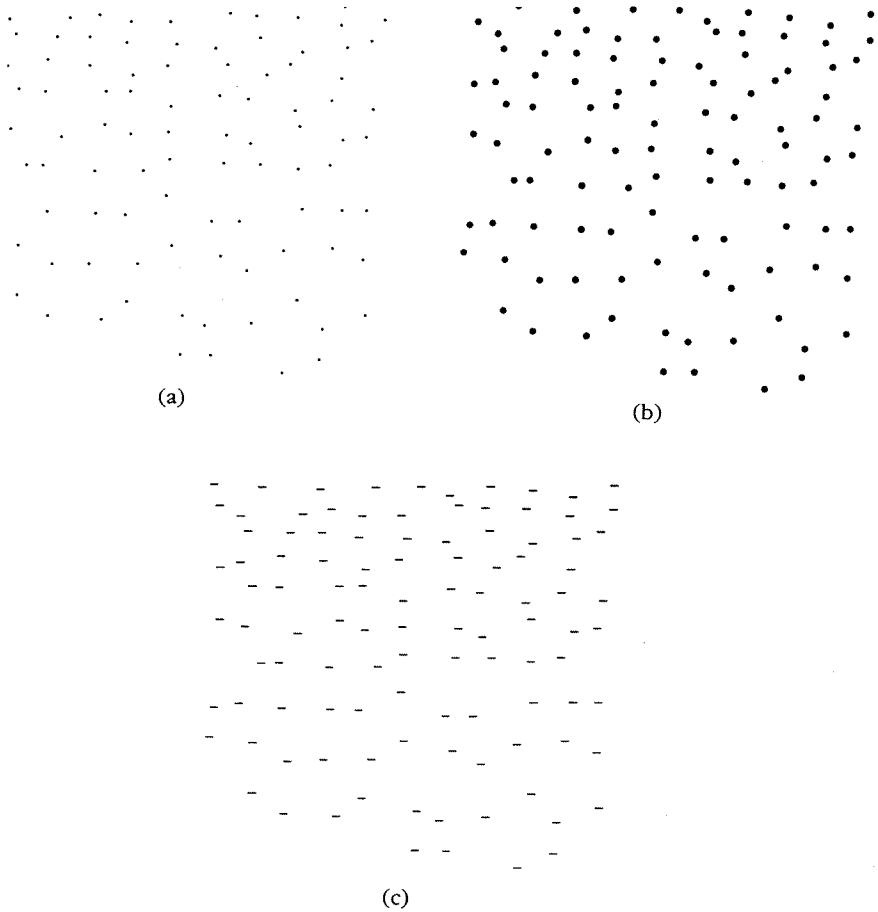


Figure 3–68. These two types of texture, although they look very different, in fact pose computational problem. In (a), the ellipses vary in width, eccentricity, and density exactly were produced by the perspective projection of equal-sized circles lying in a plane that sl from the viewer. A number of measurements could be made from such an image and use determine the geometry of the plane, and a large part of our discussion will concern whic measurements are likely to be used. In (b), the converging lines suggest a slanted surf with parallel, equally spaced straight lines. Although it has been suggested that different p are required to interpret (b) than are required for (a), this is not necessarily true, since ments of spacing, separation, and so forth can be made in both. In fact, the apparent supe the converging contours in (b) over the more random textures of (a) could be due sol greater precision in image measurements that is allowed by patterns like (b). There is n computational reason to invoke separate mechanisms.

example of this; although the two patterns look very different, similar measurements of spacing and size can be made in both. Our first question is, Which of the many possible measurements are in fact yielding the perceptual clues that give us the impression of a slanted surface? In Figure 3–68(a) are they the sizes of the ellipses, their distances apart, their density, or their density gradients?

In Figure 3–69, all the information that appeared in Figure 3–68(a) except the density gradient has been removed, and three types of tokens have been used to mark the positions of the ellipses. In all cases, although the density gradients are plainly visible and their directions clearly delineated, there is little or no impression of slant.

Surface tilt, on the other hand, does seem to be obtained quite directly from an image, although it is worth noting that it can be done in two ways



3–69. One of the possible measurements for inferring surface slant in Figure 3–68(a) is ıdient of the density of the ellipses. Texture gradient measures, in fact, have several mathe- lly attractive properties. In this figure, however, the exact gradient present in Figure 3–68(a) ɜn reproduced using three different types of local texture element. In every case the density ıt is obvious, but the impression of a slanted surface is absent, even under the best viewing ons. An impression of slant can sometimes be obtained using very high density gradients, : values involved are not physically plausible. Examples like these call into question the of whether our own visual systems actually use texture gradient measures to infer the slant ıtured surface. (Reprinted by permission from K. Stevens, "Surface perception from local : of texture and contour," Ph.D. thesis, Department of Electrical Engineering and Computer ·, Massachusetts Institute of Technology, 1979.)

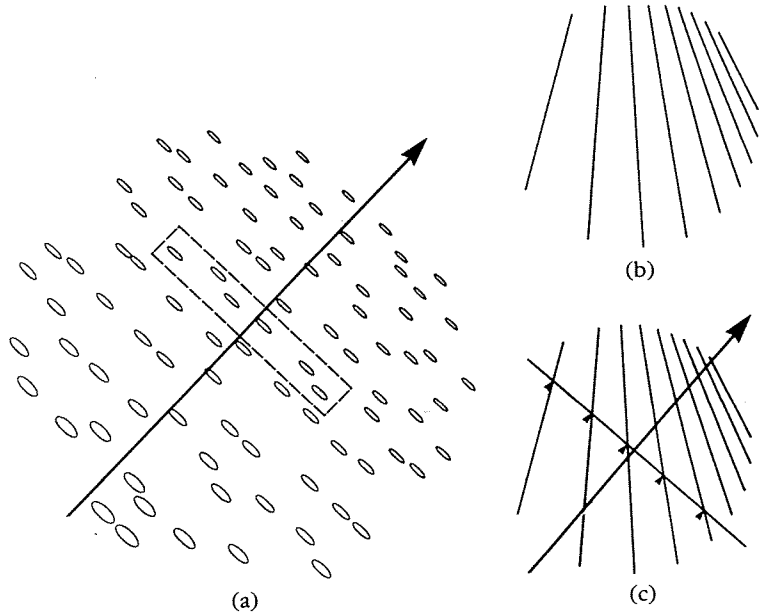


Figure 3–70. The tilt of a surface is the direction in which it is slanted away from the viewer. If the surface bears a uniform texture, the projection of the axis of tilt in the image indicates the direction in which the local density of the texture varies most, or equivalently, it is perpendicular to the direction in which the texture elements are most uniformly distributed. Either technique can be used to recover the tilt axis, as illustrated in this figure. Interestingly, however, the tilt axis in situations like (b) can probably be recovered most accurately by using the second method, that is, searching for the line that is intersected by the perspective lines at equal intervals. This method is illustrated in (c). (Reprinted by permission from K. Stevens, "Surface perception from local analysis of texture and contour," Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1979.)

(see Figure 3–70). We can detect either the direction in which the local density of the texture varies most or, equivalently, the line perpendicular to the direction in which the texture is most uniformly distributed. Interestingly, in cases like Figure 3–70(b), the second method probably provides the more accurate measurement. It is necessary only to search for the direction shown along line l in Figure 3–70(c), which the lines of perspective intersect at equal intervals. It is also known that the human visual system can detect equal intervals to within only a few percent.

Estimating scaled distance directly

Stevens' final demonstration appears in Figure 3–71, and it provides his reason for believing that we directly measure the size of the texture element from which we infer distance and then obtain an internal estimate of slant by a process akin to differentiation (see Chapter 4).

When viewed as a lighted display in a darkened room, Figure 3–71(a) gives the appearance of a slanted plane scattered with uniform-sized spheres. One possibility is that a texture gradient measure is being used to infer slant—for example, the gradient in the width of the circles. Figure 3–71(b), however, also appears strikingly three-dimensional under the same viewing conditions, yet there is no gradient here. The larger circles appear to be nearby, and the smaller ones further away. Both cases are explained by assuming that the circles correspond to uniform-sized spheres and that the different sizes in the image arise because of their different distances away, according to the simple geometrical rule that measured diameter varies as $1/r$. Therefore, the human visual system may not measure slant directly, preferring instead to estimate relative depth from size and perhaps brightness changes and then to infer slant from this.

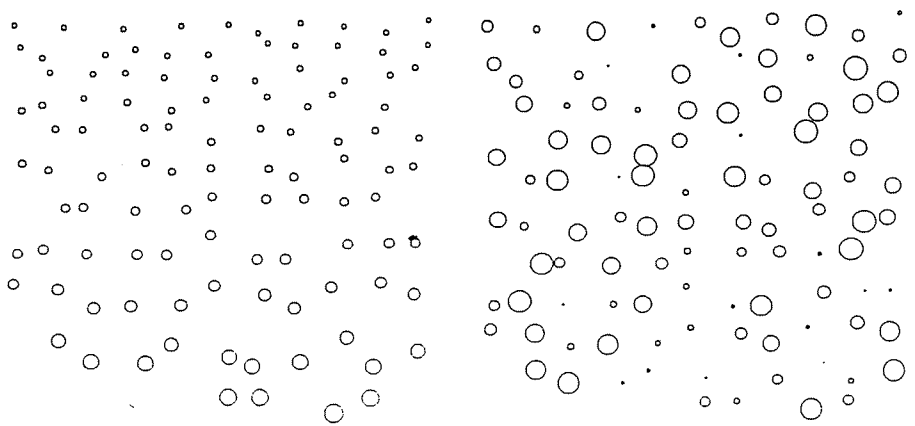


Figure 3–71. Are texture gradients used in texture vision? The visible gradient in (a) might be responsible for the apparent slant, but under suitable viewing conditions (b) appears just as three-dimensional. It could therefore be that the size or brightness of the circles is actually being used to determine slant. (Reprinted by permission from K. Stevens, "Surface perception from local analysis of texture and contour," Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1979.)

Summary

The analysis of texture is another topic that lies in a somewhat unsatisfactory state. The mathematics is easy, but the psychophysics is not, nor is it at all obvious to what extent the vagaries of the natural world allow the visual system to make use of the possible mathematical relations. In addition, unhappily little is yet known about the later stages of the full primal sketch, where the basic texture elements are actually found. Once more is known of this matter, however, empirical studies can be conducted on a variety of natural images. Probably only then shall we ever actually understand why the human visual system handles texture information in the rather peculiar and limited way in which it appears to operate.

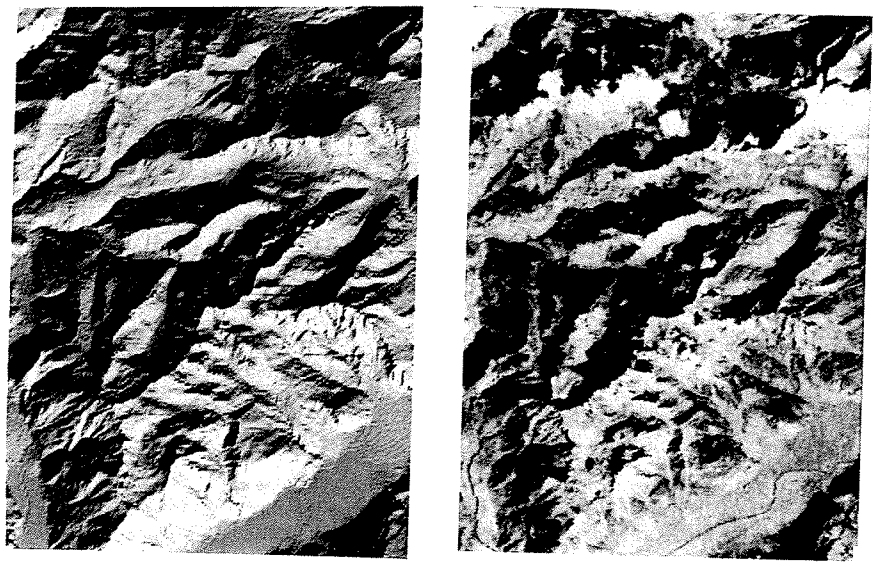
3.8 SHADING AND PHOTOMETRIC STEREO

The importance of makeup in the theater, and the widespread use of makeup in everyday life suggest that the human visual system incorporates some processes for inferring shape from shading. It seems likely, however, that the power of these processes is only slight, perhaps deriving from the combination of shading cues and information from occluding contours. On its own, shading acts as only a weak determiner of shape, and one of the most interesting problems in the theory of human early vision, along with color, is exactly what and how much information we are able to recover from shading.

From a purely theoretical point of view, the shape-from-shading problem was one of the very first to receive a careful analysis, and in his doctoral thesis (summarized as Horn, 1975), B. K. P. Horn showed how the differential equations relating image intensity to surface orientation could be solved provided that the illumination was simple and the surface reflectance known and uniform.

Since then Horn (1977) has reformulated his work in terms of the gradient space, which makes it much simpler to understand. The main use of his work has been in the development of methods for analyzing hill shading. Suppose, for example, one knows the terrain in a part of the Swiss Alps; the question is, How would it appear at 10 AM on a sunny summer's day? Or at 4 PM? Figure 3–72 shows that Horn's methods can answer these questions. By comparing the predicted image with an actual satellite photograph one is able to extract information about the reflectance properties of the land surface without being confused by the shading due to the particular terrain and illumination characteristics.

A mathematical understanding of the shape-from-shading problem is



(a) (b)

e 3–72. Comparison of the predicted and actual appearance of a portion of the Swiss Alps. omputed by Horn's methods from a knowledge of the terrain map and the reflectance map at time of day. (b) is a photograph taken from a LANDSAT satellite.

probably a prerequisite for any serious study of the human capacity for recovering shape from shading, so I have outlined the important ideas here. The interested reader should consult Horn (1977) for more details, as my account will not be very technical.

Gradient Space

The first thing necessary when discussing shape from shading is a sensible way of talking about surface orientation. For this, we borrow the representation popularized in a slightly different context by Huffman (1971) and Mackworth (1973).

Suppose we have a surface of some kind, as illustrated in Figure 3–73(a). Provided the surface is smooth, a given point on the surface will have a local tangent plane—that is, there will be a plane that is locally tangential to the surface at that point—and a local surface normal, which

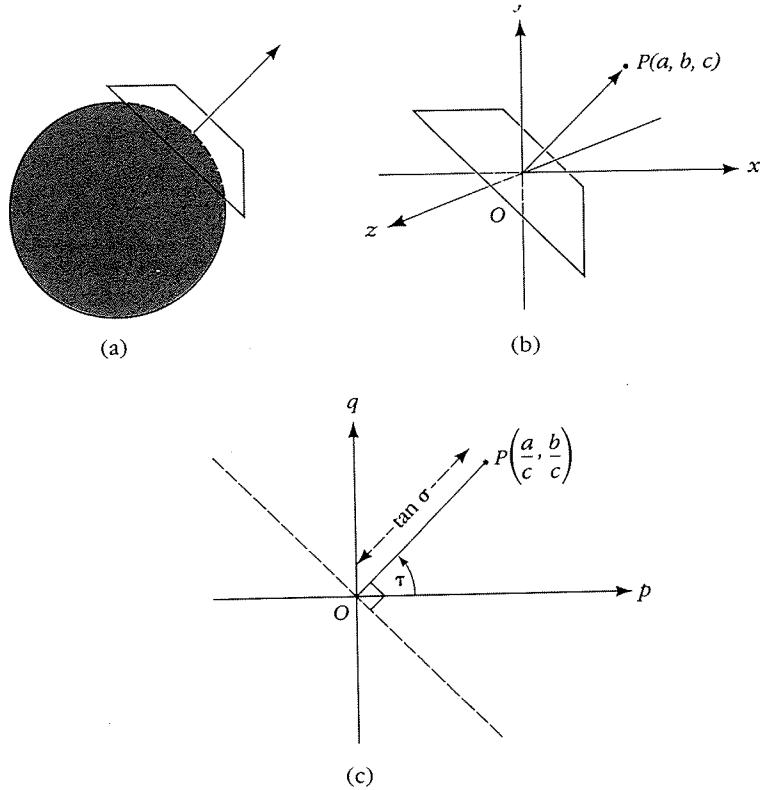


Figure 3–73. An explanation of gradient space. The local normal to the surface (a) can be represented as a vector (a,b,c), as in (b). Since we are interested only in the vector's direction, this can be reduced to $(a/c, b/c, 1)$, which can be represented as the two-dimensional vector $(a/c, b/c)$, as in (c). The quantity a/c is usually denoted by p, and b/c by q.

is the outgoing normal to the tangent plane at that point. Now take the same tangent plane, move it to the origin of the coordinate system, and draw in its normal OP, as in Figure 3–73(b). Suppose the coordinates of P happen to be (a, b, c). It clearly does not matter how long OP is, since only its direction matters, so we could just as well use the point P' at $(a/c, b/c, 1)$. But now we can represent P' by just two numbers, $(a/c, b/c)$—that is, by just the two-dimensional point P in Figure 3–73(c). This is the *gradient space* representation of surface orientation.

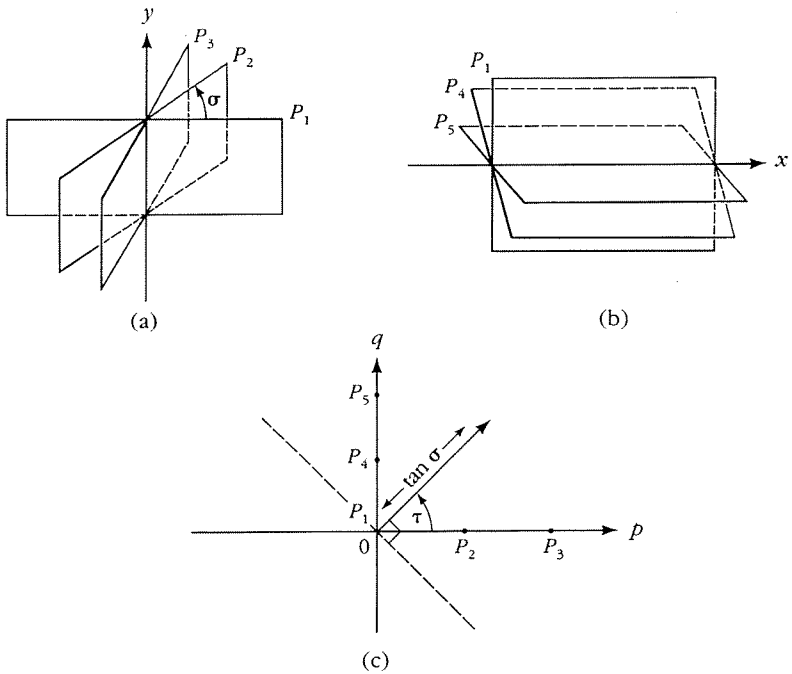


Figure 3–74. Understanding gradient space. The orientation of the frontal plane P_1 corresponds to the origin in *(p,q)* space. As the plane is rotated about the vertical (a), the corresponding point in gradient space moves along the p-axis (P_2, P_3) as shown in (c). If the plane rotates about the horizontal x-axis, as in (b), its representation in gradient space moves out along the q-axis (P_4, P_5). Similar arguments apply to rotations about intermediate axes. The depicted angle τ is called the tilt, and the angle σ the slant of the plane.

Gradient space is an elegant way of representing surface orientation. A few examples will help to make its properties clear. For a frontal plane, with the surface normal coming directly toward the viewer, $a = b = 0$ and the point P is at the origin O in Figure 3–73(c). Now imagine rotating the plane clockwise about the vertical axis as in Figure 3–74(a). Then P moves gradually to the right along the p-axis, as shown in Figure 3–74(c), and the distance from O equals the tangent of the angle of slant. If instead we rotate the plane about the horizontal axis, as in Figure 3–74(b), P moves along the q-axis, as shown in Figure 3–74(c), again by an amount equal to the tangent of the angle of slant. If we rotate about some intermediate axis, shown dotted in Figure 3–74(c), then P moves out at right angles to it along the direction τ from the p axis, as shown in Figure 3–74(b). This angle τ is what in the psychophysical literature is referred to as the tilt of the plane, and the angle between it and the frontal plane is usually called its slant, and sometimes its dip. I shall use the letter σ to denote slant. The distance between the point P and the origin is tan σ.

The reader might like to take a few moments to play with a piece of paper and understand gradient space fully, because it is an important and useful idea. In particular, he might prove to himself that the length of OP is equal to tan σ.

Surface Illumination, Surface Reflectance, and Image Intensity

The study of shape from shading is concerned with finding ways of deducing surface orientation from image intensity values. The problem is complicated because intensity values do not depend on surface orientation alone; they depend on how the surface is illuminated and on the surface reflectance function. In the real world, the prevailing illumination is often complex, especially indoors. Outside is more straightforward—the sun is nearly a distant point source, and the ground illumination that is produced by thick cloud cover is nearly uniform, so these two situations are quite simple. A partly cloudy day can sometimes be treated as a combination of the two. But at ground level, the situation is often made very complex by secondary illumination effects—light bouncing off one surface onto another and thence into our eyes. These effects are almost impossible to treat analytically.

Just like the echo effects in acoustics, secondary illumination becomes especially important for indoor scenes, where light from a ceiling fixture can reach the coffee table top either directly or after reflecting off the ceiling or walls. The ceiling will help to illuminate the walls, and these in turn will reflect light back, helping to illuminate the ceiling—a condition called mutual illumination. The combined complexity introduced by all these effects makes the analysis of shape from shading extremely difficult, and no real progress has yet been made with the problem except in the very simple illumination condition of one distant point source. Horn, however, has effectively solved this situation, and we shall shortly look at how he did it.

The second factor that profoundly influences the shape-from-shading problem is the surface reflectance function. The fraction of light reflected

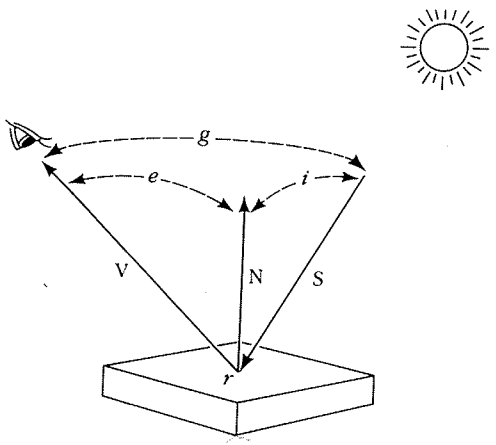


Figure 3–75. The definitions of the angles i, e, and g.

toward the viewer from a source depends on the microstructure of the reflecting surface, and this is usually described as a function of the three angles shown in Figure 3–75—the *angle of incidence* i between the source and the surface normal, the *angle of emittance* e between the line of sight to the viewer and the surface normal, and the *phase angle* g between the incident and emitted rays. The *reflectance function* $\phi(i,e,g)$ is the fraction of incident light reflected per unit surface area per unit solid angle in the direction of the viewer. Intuitively, this says that the amount of light incident on a surface patch that will be reflected to a detector depends directly on the area of the illuminated patch, the value of $\phi(i,e,g)$, and the angular size of the detector.

There are many kinds of reflectance function. A perfect Lambertian surface—a pure matte surface—looks equally bright in all directions and has the simple reflectance function $\phi(i,e,g) = \cos i$. The surfaces of rocky, dusty objects that are viewed from great distances have another interesting type of reflectance function; for fixed-phase angle g, ϕ depends only on $\cos i/\cos e$. This relationship applies to the material in the maria of the moon—and for observation from the earth, g is indeed constant. This has greatly helped the study of lunar topography.

A polished metallic surface has a particularly simple reflectance function ϕ; it is 1 when $i = e$ and $g = i + e$, the properties of a pure mirror. If the surface is not quite so polished, then ϕ is smudged a little around this value, often by convolution with a Gaussian. This smudged property is particularly interesting because many everyday surfaces have a reflectance function that combines two components, one matte and one specular. The reflectance function of glossy white paint is made up of such a combination. For example, if s is the fraction of light reflected specularly, its reflectance function might have the form

$$\phi(i,e,g) = \frac{s(n+1)(2\cos i\cos e - \cos g)^n}{2} + (1-s)\cos i$$

where the first term is the specular component, and the second the matte. The number n determines how sharp the specular peak is; a typical value for a glossy paint might be $n = 16$ (see Horn, 1977).

The Reflectance Map

The best way of understanding the shape-from-shading problem is to understand the reflectance map, which is a way of relating image intensities directly to surface orientation.

Suppose we take a particular type of surface with a known reflectance function ϕ. Suppose we take distant source and viewing positions, so that the phase angle g is constant, and suppose that we take just a single source, so that the problem is expressed in its very simplest form. Then each surface orientation will produce a particular intensity in the image, which we can plot in the (p,q) gradient space map. In fact, let us choose to plot our reflectance map in a particularly simple way—let us draw in the contours of constant reflected intensity, normalized to some scale lying between 0 (for darkness) and 1 (the maximum possible intensity of light that could be found in the image). Then if the measured intensity at a given point is, say, 0.8, we know that the surface orientation (p,q) must lie on the 0.8 contour in the reflectance map.

Figures 3–76 to 3–79 show some examples. Figure 3–76 is the reflectance map for a pure matte (Lambertian) surface illuminated from a source that is by the viewer. In Figure 3–77, the surface is the same, but the source is in a different direction (actually in direction $p = 0.7$, $q = 0.3$). Notice here there is a shadow line—the line of surface orientations at which the surface becomes self-shadowed from the source. Figure 3–78 shows the peculiar reflectance map characteristic of the maria of the moon, and Figure 3–79 shows the reflectance map for our glossy white paint. The very closely spaced circular contours correspond to values of intensity that change very rapidly with any change in surface orientation, and so they correspond to the specular component. The rest of the map looks more like Figure 3–77 and corresponds to the matte component.

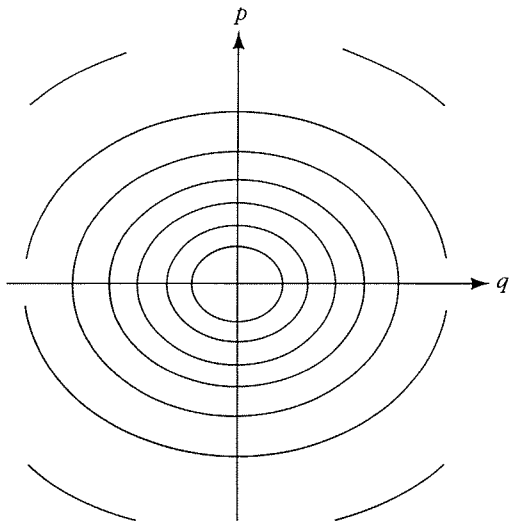


Figure 3–76. Contours of constant cos i. Contour intervals are 0.1 unit wide. This is the reflectance map for objects with Lambertian surfaces when there is a single light source near the viewer.

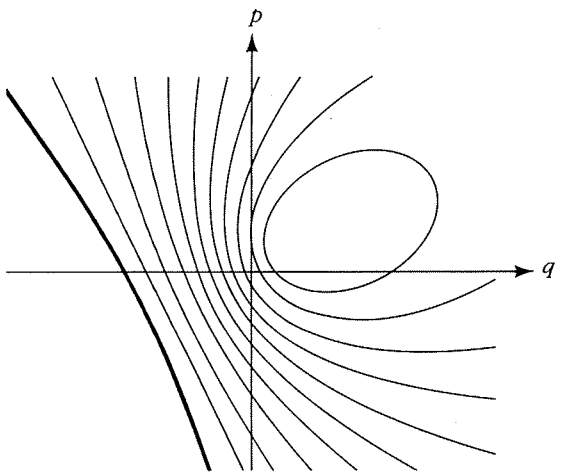


Figure 3–77. Contours of constant cos i. Contour intervals are 0.1 unit wide. The direction to the source is $(p_s, q_s) = (0.7, 0.3)$. This is a typical reflectance map for objects with Lambertian surfaces when the light source is not near the viewer.

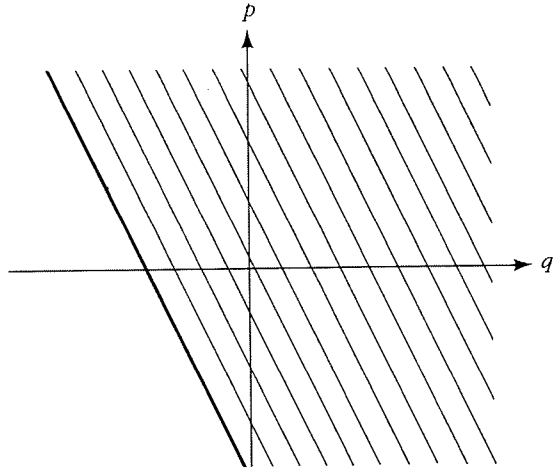


Figure 3–78. Contours of $\phi(i,e,g) = \cos i/\cos e$. Contour intervals are 0.2 unit wide. The reflectance function for the material in the maria of the moon is constant for constant cos i/cos e.

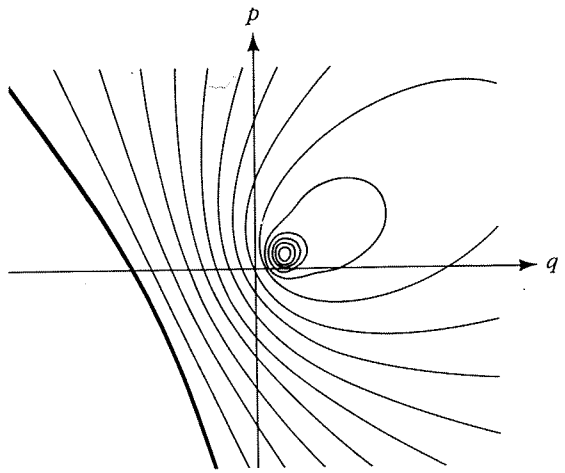


Figure 3–79. Contours for $\phi(i,e,g) = 0.5s\,(n + 1)(2\cos i \cos e - \cos g)^n + (1 - s)\cos i$. This is the reflectance map for a surface with both matte and specular components of reflectance when the surface is illuminated by a single point source. Glossy white paint can produce such a ϕ.

The fundamental problem with recovering shape from shading is that, even with all the simplifying assumptions that enable us to use a reflectance map, it is still very difficult. Knowing the intensity value places one on a particular isoluminance contour in the reflectance map—for example, it might tell us that the surface orientation lies on the 0.8 contour—but it does not tell us where. Unless we have additional information, one position on the contour is just as good as any other.

However, the problem can be solved. The extra condition we need is to assume that the surface is smooth and that surface orientation varies smoothly (that is, is differentiable). Essentially this says that if you are at one point in the image and know the surface orientation there and how it changes locally, then if you move in one direction across the image, you can tell from the new intensity value what the new local orientation is.

This is an amazing fact, because one would not think that smoothness constrains the answer enough. But it does because of a beautiful mathematical trick (Horn, 1977), which I am unfortunately unable to reduce to succinct English. So from a mathematical point of view, the problem is soluble. However, from a biological point of view, this type of solution, even given the major simplifications on which Horn's approach rests, is still far too complicated to be used. To solve the shape-from-shading equations for a general reflectance map requires successive integration along paths in the image whose loci can be determined only as the integration proceeds. Solving these equations in a simpler, more parallel way appears quite hopeless unless we are prepared to introduce other constraints.

A number of other approaches have therefore been tried. Woodham (1977) combined constraints on surface orientation—like minimizing local curvature—and constraints from shading to produce a local iterative approach to determine surface orientation. Brady (1979) suggested restricting the type of surface as well, for example, to generalized cones, and showed how one can then determine the direction of the light source.

However, I think it is fair to say that none of these approaches has yet thrown much light on the use of shading information by the human visual system. The difficulty is probably that we do not use this information very well. The human visual processor seems to use only coarse shading information, often but not always correctly, which is probably why shading is easily overridden by other cues. Situations where the human visual system does not perform well always cause trouble because knowing how the problem ought to be solved mathematically may throw very little light on how we ourselves approach it. Unfortunately, the same may be true for color, as we shall see. Nevertheless, we do make some use of shading, so there is definitely something here to be understood.

Finally, there is a technique for recovering shape from reflectance maps that cannot possibly have any biological significance, but which is so elegant that I cannot resist mentioning it. The idea was introduced by Woodham (1978) and developed by Horn, Woodham, and Silver (1978), and it rests on the following idea. Given an image and a reflectance map for one position of the light source, suppose that we measure image intensity at one particular point. As we have seen, we may then deduce that the corresponding surface orientation lies on a particular contour in gradient space—the 0.8 contour was our example in the previous section—and I have illustrated it in Figure 3–80(c). The problem is that we do not know where along this contour the correct surface orientation (p,q) is.

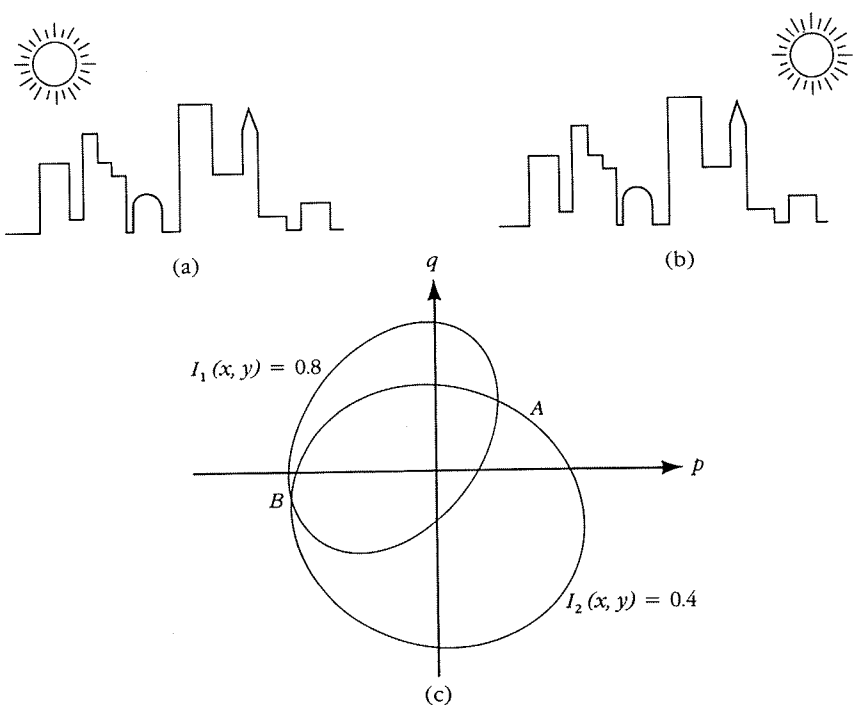


Figure 3–80. The idea behind photometric stereo. Images I_1 and I_2 are taken of the same scene under two different lighting conditions, and so two different reflectance maps are employed. From the first, image intensity measurements may place a particular point in the image on the 0.8 contour (a); from the second, on the 0.4 contour (b). Hence the true surface orientation (p,q) corresponds to either point A or point B in (c), the intersection points of the two contours.

Suppose, however, that we now move the light source—or, in an outdoor scene, we wait until later in the day—and then take a second image from the same viewpoint. The surface geometry relative to the viewer is all the same, but the reflectance map changes. For example, the situation may change to look like Figure 3–80(b), and the intensity measurement at the same point in the image puts us on the 0.4 contour in the reflectance map, as shown in Figure 3–80(c). Then the true surface orientation is narrowed down to just two possibilities—the two points at which the first 0.8 contour and the second 0.4 contour intersect, points A and B in Figure 3–80(c). This essentially solves the problem, since the choice between A and B can usually be made easily by using continuity information or by taking a third picture with yet another lighting position.

This type of scheme may be of practical use, since we can usually construct a reflectance map even for complicated lighting conditions, although we usually have to measure the reflectance map empirically because it is too difficult to compute. Provided that the lighting and surface characteristics are the same everywhere in a scene, the sole determiner of image intensity is surface orientation.

3.9 BRIGHTNESS, LIGHTNESS, AND COLOR

All the processes that we have considered so far have used the image of reflectance and illumination changes on a surface to recover information about the geometry of the surface. Nothing has been said about the nature of the surface itself. Yet the reflectance of a surface—whether it is light or dark, whether it reflects red light well or poorly, and so forth—carries information that often has important biological significance. For example, we can tell just by looking whether a fruit is ripe, whether a branch is strong enough to bear one's weight, whether a leaf is green and supple, whether an insect is likely to be poisonous, and many other things.

The business of recovering surface reflectance, then, is important, and we are actually quite good at it. It is surprising how much perceived color depends upon the reflectance of a surface and how little it depends on the spectral characteristics of the light that enters our eyes. According to Helson (1938), an illuminant may be up to 93% chromatic, but provided it contains at least 7% "daylight", surfaces with uniform spectral reflectance—that reflect equally at all wavelengths—will remain achromatic. The opposite aspect of the problem is by how wide a range of stimuli we can be fooled into saying that brightness differences exist where they objectively do not—from the Hering grid and Benussi ring on the one hand to the phenomenon of subjective contours on the other. Some examples appear in Figure 3–81.

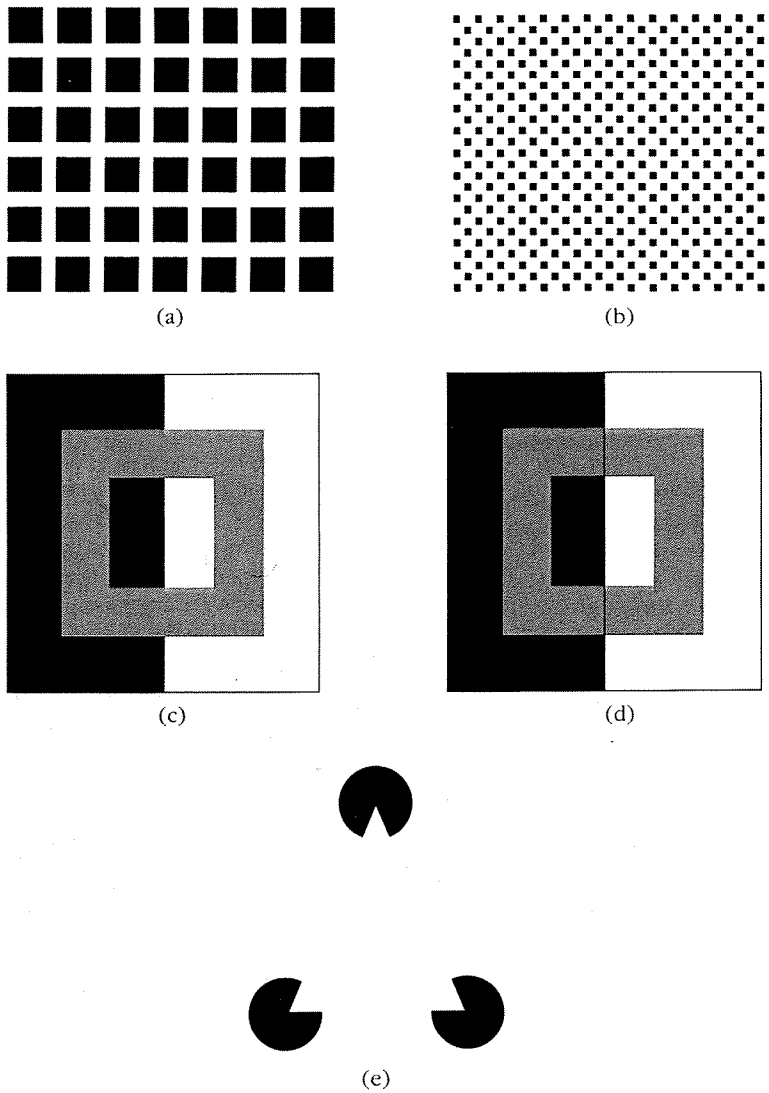


Figure 3–81. Some well-known brightness illusions. (a) The Hering grid. (b) An illusion by Robert Springer that provokes the appearance of faint diagonal lines. (c), (d) The Benussi ring; notice how the simple addition of a contour in (d) can cause the two gray regions to look different. (e) The Kanizsa triangle.

The theory of color vision is in an unsatisfactory and interesting state. On the one hand, we have for a long time had a fairly adequate phenomenological description, due to Helson (1938) and Judd (1940). Their equations can be used to predict the colors that will be perceived by a subject about as accurately as the subject is able to describe them, and they can, without modification, account for Land's (1959a, 1959b) famous two-color projection demonstrations in which images produced with only two colors gave full color percepts (Judd, 1960; Pearson, Rubinstein,, and Spivack, 1969). As Helson and Judd themselves commented, however, there are probably many other equations that describe color perception just as well; in fact, Richards and Parks (1971) proposed a simpler model that is nearly as accurate.

The problem is that these formulations are *descriptions* of color vision, not *theories* of it. The researchers do not say why their equations are good at separating the effects of the illuminant from the effects of surface reflectance. Of course, there may be no real theory of color vision, and these descriptions may be as close as we can get—but I hope not. The only attempt at a true theory of color vision is Land and McCann's (1971) retinex theory. This theory has been criticized for explaining nothing that the Helson–Judd formulation cannot account for, and this is probably true. But that comment misses what, from this book's perspective, is the most important difference between these two theories, namely, that the Helson–Judd formulation is a phenomonological description, whereas the retinex idea is a computational theory that is based on particular assumptions about the physical world. To bring these points out, let us look in more detail at the two formulations.

The Helson–Judd Approach

The basis for Helson and Judd's approach to color vision is the time-honored view that object color depends on the ratios of light reflected from the various parts of the visual field rather than on the absolute amounts. Helson and Judd tried to construct a formula that predicts what color a given piece of paper will appear to have under different illumination conditions and against different backgrounds. Thus they were not so much interested in color constancy as in quantifying the extent to which constancy is violated as the illumination and background are changed.

Their formulation is based on two steps. First, find out what "white" should be for the conditions prevailing in the scene; second, compute what color the paper should have by referring to this estimate of white. The basic idea behind finding the white is (1) to take the standard daylight white, which by a suitable choice of coordinates we can denote as (r_w, g_w); (2) to measure the "average" color of the whole visual field, which we denote by (r_f, g_f); and (3) to assume that the current white (r_n, g_n) lies somewhere between these two. For example, we might write

$$r_n = r_f - k(r_f - r_w)$$

$$g_n = g_f - k(g_f - g_w)$$

according to which the current white lies on the straight line joining daylight white to the average over the current visual field.

This basic idea is then modified by incorporating various empirical observations that Helson and Judd made to produce a complex expression that is no longer linear. In other words, the modifications push the current white off the line joining daylight white to the current average, so as to account for the various odd effects that Helson and Judd found empirically. The most important modification comes about because of a notion they had called *adaptation reflectance,* which is essentially a shade of gray that depends on the scene. Papers that are lighter than this gray take on the hue of the illuminant, whereas darker papers take on the complementary hue. Of course, linear formulas cannot account for this effect. Other modifications arise because adaptation effects increase in power as we move away from white, peculiar effects occur if the blue component of the illuminant is intense, and so forth. The result is a long and complicated formula, adding to the basic equations above a number of second-order, nonlinear terms, each justified by a particular aspect of the experimental findings. The second part of the scheme, assigning color relative to this estimate of white, has a simple formulation. To determine the hue to be associated with the point (r, g), we simply examine the orientation of the line joining it to the current white (r_n, g_n); the length of this line determines the saturation.

The interesting thing about this approach is that these assumptions lead to a successful predictor of perceived color. What is missing is an explanation of why we can make these assumptions and why they lead to valid color perception under such a wide range of circumstances.

Retinex Theory of Lightness and Color

Land and McCann (1971), on the other hand, base their theory firmly on assumptions about the physical world. It applies to the planar world of so-called Mondrians, which, as we saw in Chapter 2, consist of rectangular

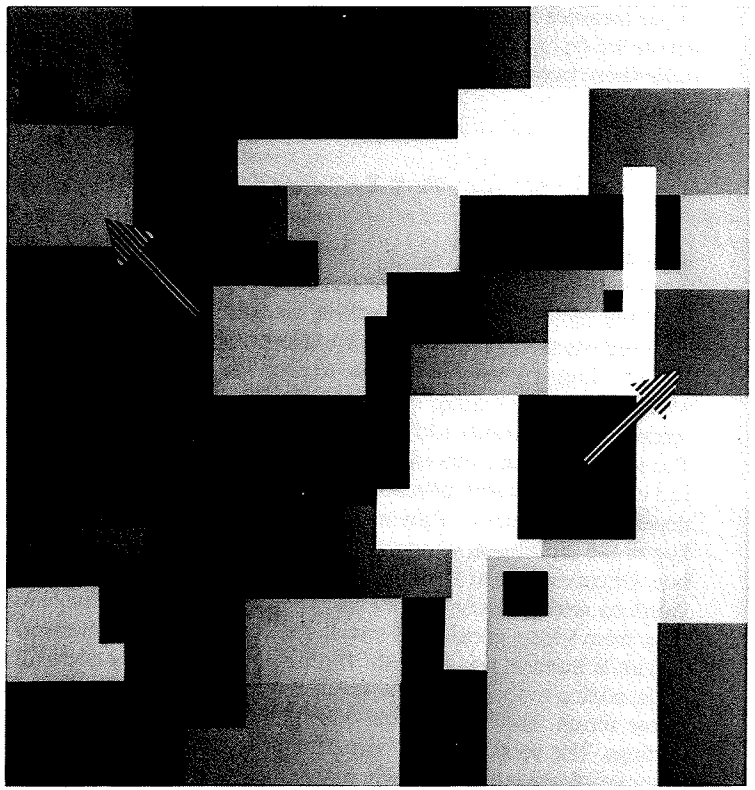


Figure 3–82. The two marked squares have the same luminance, yet one is perceived as being much darker than the other. (Reprinted by permission from E. H. Land and J. J. McCann, "Lightness and retinex theory", *J. Opt. Soc. Am. 61* (1971), 1–11, fig. 3.)

patches affixed to a large board that can be illuminated in various ways (see Figure 2–30). The first part of the theory, concerned with what Land and McCann called lightness, deals with monochromatic images of just this kind. The central problem, as they state, is to separate the effects of surface reflectance from the effects of the illuminant, because as has long been known, what we perceive as the color of a surface is much more closely connected with spectral characteristics of its reflectance function than with the spectral characteristics of the light falling upon our eyes.

How can these effects be separated? What critical characteristics might enable us to separate the effects due to changes in illumination from the effects due to changes in reflectance? Land and McCann proposed the following: Changes due to the illuminant are on the whole gradual, appearing usually as smooth illumination gradients, whereas those due to changes in reflectance tend to be sharp. This dichotomy is certainly true in the Mondrian world that they studied, and hence if we can separate the two types of change, we can separate effects of changes in the illuminant from the effects of changes in reflectance in these images.

An example of what Land and McCann mean appears in Figure 3–82. This shows the image of a monochromatic Mondrian lit from above. The two patches marked with arrows have exactly the same intensity, yet one appears to be darker than the other. If one removes the effects of the illumination gradient, one patch would actually become much darker than the other. The argument is that this computation is essentially what our visual systems do, and it is called the retinex computation.

Algorithms

The retinex computation has been implemented in at least two ways. Land and McCann themselves used the one-dimensional approach illustrated in Figure 3–83(a). If we trace the image intensities along any path from A to B as shown, they will have the form shown in the first graph, portions of slow changes interspersed with large jumps at the reflectance boundaries. By applying a threshold, we can remove the effects of the slow changes, thus arriving at the curve in the second graph, which describes the effects of the reflectance changes only. Since the system is conservative, it does not matter which path from A to B is used—the resulting assignments of reflectance will always be the same. Land and McCann used this technique together with a sufficient number of randomly chosen paths across the image to cover all locations adequately.

Horn (1974) derived a two-dimensional analogue of this algorithm, illustrated in Figure 3–83(b) and consisting essentially of the same three steps. The first step is to take a differencing operator, here having a two-dimensional center–surround form. Then we ignore small values and accept only large ones, which correspond to the reflectance changes. Finally, using only the large changes, we reconstruct the image to get a two-dimensional analogue of the second graph in Figure 3–83(a). For this, Horn suggested an interesting iterative algorithm based on nearest-neighbor interactions in order to implement the equations shown in Figure 3–83(b).

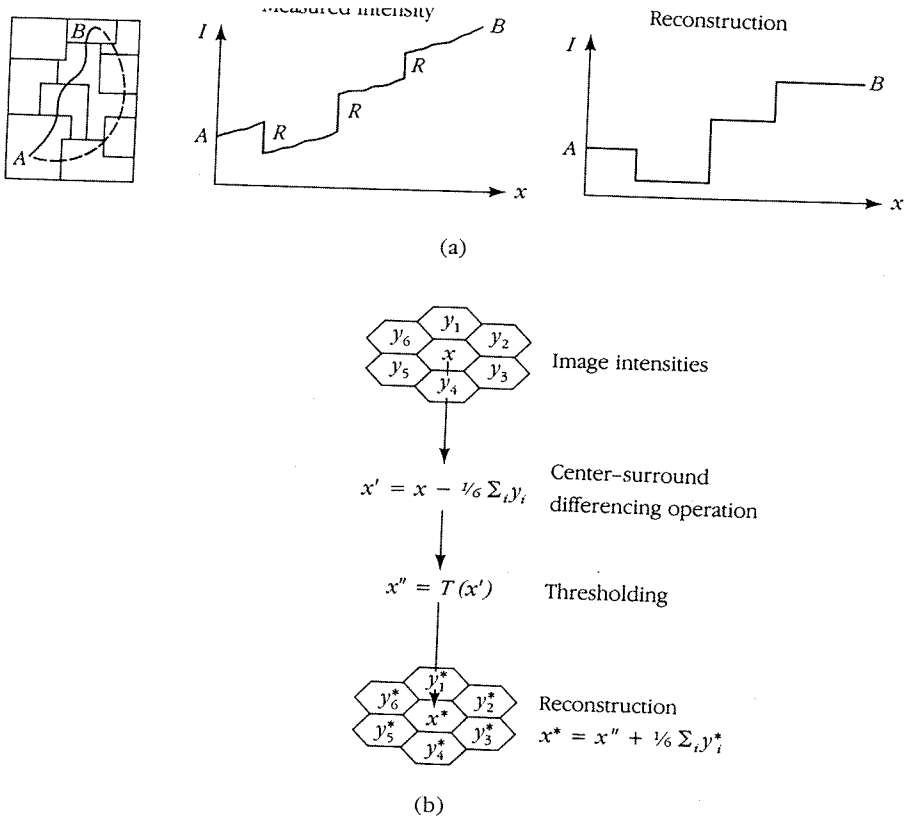


Figure 3–83. Diagrams illustrating retinex algorithms. (a) Land and McCann's one-dimensional algorithm. (b) Horn's two-dimensional version. In both, the idea is to ignore smooth changes in intensity, taking account only of discontinuities. See text for details.

Extension to color vision

The operations diagrammed in Figure 3–83 show the retinex operating monochromatically. In order to apply the operation to color, Land and McCann require that it be performed independently in each of the red, green, and blue channels. What then emerges from each, they hope, are signals that depend not on the illuminant but solely on the surface reflectance. These can be combined to give a percept of color that happily rests solely on properties of surface reflectance and not on the vagaries of its particular, present illuminant. Of course, there is still the need to calibrate the signals in the three channels relative to one another, but Land and McCann suggest that this can be done by calling the brightest point in the scene white.

McCann, McKee, and Taylor (1976) have recently published comparisons between the results predicted by such an algorithm on their Mondrian stimuli and the psychophysical estimates of color made by subjects who viewed them. They found that the agreement between their subjects and their predictions was as good as the agreement among their subjects.

Comments on the retinex theory

To me, the positive aspects of Land and McCann's work seem to be threefold. First, they have attempted to construct a real theory of color vision, as opposed to a description of color perception. Second, they have drawn attention to the importance of boundaries and described one way in which boundary effects may propagate across an image. Such effects had been known for a long time—for example, the Craik–Cornsweet illusion and the Benussi ring—but boundary effects do not appear explicitly in the Helson–Judd equations. Third, Land's earlier work formulated an interesting principle considered important by Judd, namely, that when the colors of the patches of light making up a scene are restricted to a one-dimensional variation of any sort, the observer usually perceives the objects in that scene as essentially without hue.

The case against the retinex theory seems to consist of one major and several minor arguments. The major argument is that there is more to simultaneous contrast than is present in the retinex theory. That is, formulations like Helson and Judd's that are based on the idea of simultaneous contrast may be able to explain Land and McCann's effects, but the gradient-eliminating retinex theory cannot explain all of simultaneous contrast, because these effects occur perfectly well in situations of uniform illumination, where there are no illumination gradients. In addition, Land and McCann apparently did not always pay adequate attention to the effects of simultaneous contrast in their displays. For example, in Figure 3–82, one of the squares has darker neighbors than the other, so one might expect them to appear different just on these grounds. In any event, brightness perception and color perception appear to involve at least some effects that are not predicted by Land and McCann's approach.

One possible explanation is that these extra effects are introduced by aspects of the problem that Land and McCann did not consider. For exam-

ple, their theory applies only to planar surfaces, and these other effects may be introduced only to deal with the added complications of having different surface orientations in different parts of the visual field. This, however, is unlikely. Although there certainly are three-dimensional effects on brightness perception, they are probably not very large. Gilchrist (1977) recently claimed that perceived orientation could affect brightness perception by factors of up to 30%, but, in repeating his experiments, Ikeuchi (1979) was unable to obtain factors much greater than 5%–10%.

The first of the minor arguments against the retinex idea is computational: The theory involves a threshold (the level of gradient at which the cutoff occurs), but it does not say what that threshold should be. It is a matter of unhappy experience that whenever we have to set a threshold in an image-processing task, we usually have problems—which is one reason why the idea of zero-crossings is so attractive. The problem is that if the threshold is too low, it will not remove the illumination gradient; but if it is too high, it will remove valuable shading information. Gradual changes in surface orientation also produce gradual changes in intensity across an image, and these might be too valuable to throw away cavalierly. And gradual changes in surface coloration can also be important. After all, we can see a rainbow, even one that has been enlarged by binoculars.* The color changes are not thresholded out.

The second minor argument arises from neurophysiological observations. According to the retinex theory, the red, green, and blue channels are processed independently, each in the manner of Figure 3–83, and combined only afterward. This, however, is not the observed situation. Neural processing seems to be based on an opponent-color approach—where the output depends on the difference between two color channels—right from the start. Even in the retina, most color-sensitive cells have an opponent-color organization (DeValois, 1965), and DeValois and his associates have found an impressive correlation between the psychophysics of color discrimination and the observed neurophysiological properties of lateral geniculate color-opponent cells.

These findings do not disprove the notion that the retinex function is being computed in the visual pathway. One could argue, as Horn (1974) pointed out, that the retinex can be carried out on any three linear combinations of red, green, and blue just as well as on the original channels themselves, and this adjustment might make the retinex theory compatible with the neurophysiological observations. But this argument is not very convincing, especially since the theory provides no very good reason why one should want to operate on linear combinations rather than on the original signals.

Some Physical Reasons for the Importance of Simultaneous Contrast

It is a widespread and time-honored view, going back at least to Ernst Mach, that object color depends upon the ratios of light reflected from the various parts of the visual field rather than on the absolute amount of light reflected. Of course, this must be because although the illumination of a scene, which greatly influences the spectral distribution in its image, changes drastically from time to time and from place to place, we are relatively immune to the variation. The range of color constancy is, of course, bounded—when we buy clothing we insist on seeing the items in daylight or under tungsten illumination if the store's lighting is fluorescent. But the important point is that although our perceptions may only approximate the objective reflectances, they do this much more accurately than they reflect the spectral qualities of the light falling upon the retina.

Even within a single scene, the intensity of illumination can change drastically, from sunlight to shadow, for example, or from near the lights in a large hall to the dim recesses of the furthest-flung corners. The spectral characteristics can also change, although usually not by so much. The light becomes greener under a tree than in the open; in the mouth of a cave it can turn browner. So even though the main fluctuations in spectral content occur over time, they can still occur in a single scene, and this does not much affect us.

How can we deal with such a wide range of effects? What the simultaneous-contrast phenomena* seem to be drawing attention to is an argument of the following kind. Suppose you pass an embankment where a yellow or blue flower happens to be growing amid a background of green grass and clover. Although the absolute spectral characteristics of the light coming from the flower cannot at all be relied upon as a clue to its surface reflectance characteristics, either in the matter of its lightness or of its spectral properties, nevertheless its characteristics relative to other nearby surfaces probably are reliable. If the flower appears lighter than the grass, this is probably due to a characteristic of the flower and not of the illumination (though the head of the flower could be just catching the sun). If the flower looks bluer than the grass, then it probably really is. If it looks yellower, then, again, it probably really is.

Furthermore, what is so amazing about simultaneous-contrast effects—even as simple as those in Figures 3–81(b) and (c)—is that the visual

*The tendency for color or brightness of one area to affect neighboring areas.

system seems to take them so seriously. That is, we get what looks like wrong answers in situations as simple as the Bernussi ring (Figure 3–81c), where we would think that almost any sensible scheme would give an answer reflecting the objective truth of the situation. I find this so striking that I am tempted to believe that relative observations may be *all* one relies on.

Even so, to make a success of a scheme based only on relative measurements, we have to make a basic distinction between changes in the image due to changes in reflectance (like the difference between a flower and grass) and those due to changes in illumination (like the shadow of a nearby tree). The fact is that shadowed lawn looks darker than unshadowed lawn, and a daisy in the sun looks brighter than a daisy in the shade, but the shadowing does not much affect the color of the lawn or the daisy. The sunlit daisy and the shadowed daisy both look white, and (critically) the shadowed daisy does not look gray.

We naturally consider the sunlit daisy brighter than the shadowed one. This suggests that brightness is a subjective quality related to the intensity of the prevailing illuminations. The reflectance of the surfaces, on the other hand, is more closely related to the qualities of lightness and color. Changes in lightness are ideally pure scalar changes in a surface's reflectance involving no changes in the surface's spectral characteristics (detectable through the three color channels), whereas changes in color refer ideally to changes in the spectral characteristics of a surface and may be described by the two components hue and saturation. Helson (1938) and Judd (1940) defined the terms *brightness, lightness,* and *color* purely psychophysically, but I think that to regard them as perceptual approximations to illumination intensity and to the value and spectral distribution of surface reflectance is consistent with their definitions (see Judd, p. 3).

The computational problem, therefore, is how to formulate in a reasonable way the physical basis for estimating brightness, lightness, and color from an image. The first point to note is that surface orientation can influence brightness (according to our definition) but not usually a surface's lightness or color, because at some orientations a surface will be more directly illuminated than at others. The final solution to the computation of brightness will therefore have to await an estimate of the surface's orientation. As we have noted, however, the effects of 3-D interpretation on perceived brightness are still not fully established.

The major source of brightness changes is shadows, and again, as we saw in Section 2.4, these can be detected autonomously by using the ideas behind the operator $\nabla I/I$. These two phenomena, surface orientation changes and shadows, provide the main sources of discontinuity in brightness, so provided that they are taken adequately into account, we can be

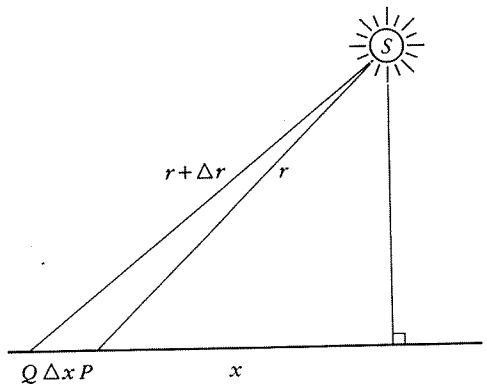


Figure 3–84. Gradients in intensity that are due solely to illumination are usually small and almost linear. S is a source that illuminates the plane containing P and Q. The most significant terms in the difference between the intensities at P and Q depend on Δx, the distance between the two points.

fairly sure that the remaining changes in the illuminant are smooth rather than sharp.

Our next observations are (1) that locally measurable illumination gradients on a flat surface can occur only if the light source is not very far away, (2) that they are small unless the source is very near, and (3) that they are approximately linear except perhaps directly under the source. This can be seen from Figure 3–84. The illumination at P is I/r^2, and at Q nearby it is $I/r^2 - 2x\Delta x/r^3 + O(1/r^4)$. If $\Delta x/x$ is small, the change from P to Q varies approximately with $-2\Delta x/x$. This is essentially linear in Δx, the distance between P and Q, provided that Δx is small compared with x. In other words, illumination gradients are almost always small and linear. This may be one reason why the human visual system is insensitive to small linear changes in intensity (see Brindley, 1970, p. 153).

Hypothesis of the Superficial Origin of Nonlinear Changes in Intensity

These observations suggest that the following approach to the physical basis of color vision may be profitable: *In the absence of sharp changes in brightness, detectable as shadow boundaries or changes in surface orientation, all nonlinear changes in intensities may be assumed to be due to*

properties of the surface—either its orientation or its reflectance. In other words, in the absence of obvious illumination effects like shadows, measurable nonlinear local differences in either image intensity or spectral distributions are due to changes in the lightness or color of the surface. This assumption allows us to ignore small linear changes and provides a basis for the idea that lightness and color may be recovered from measurements of nonlinear local changes in intensity and spectral distribution made, for example, by comparing their values at each point with their values in the surrounding neighborhood.

Implications for Measurements on a Trichromatic Image

According to physiological descriptions, some opponent-color cells in the retina of the monkey have receptive fields with rather mixed properties, like a red center and green surround (Gouras, 1968; de Monasterio and Gouras, 1975). There seem to be no internal reasons for doubting these reports; nevertheless, I find such cells extremely difficult to understand in general and impossible to fit into the $\nabla^2 G$ framework that we developed in Chapter 2.

The reason for the difficulty is that a cell with such a receptive field, illustrated for convenience in Figure 3–85(a), signals a complex mixture of spatial and chromatic information. It signals neither a pure $\nabla^2 G$ function for a single chromatic channel, like the red $\nabla^2 G$ receptive field illustrated in Figure 3–85(b), nor a purely chromatic message about the relative strengths of signals in the two channels at one point in the image, as would the receptive field illustrated in Figure 3–85(c). In fact, Figure 3–85(a) is not even a zero-mean operator—it is not like a second derivative, and its zero-crossings are meaningless. To use it, we have to pay special attention to *changes* in its value—for example, if its green-center, red-surround analogue looks at a lawn, it will fire everywhere over it, slightly harder for the more saturated greens. This seems to me not only poor engineering but also a contradiction to the experience we have about how the nervous system likes to code changes rather than pure values; in other words, it violates Barlow's (1972) second dogma about the economical neural encoding of stimulus information.

In order to make a reasonable concrete suggestion about what these cells are signaling, I would like to combine two pieces of information. The first is that the $\nabla^2 G$ style of analysis requires that the spectral characteristics of the center and of the surround be essentially the same, related to one another by a minus sign. This is necessary for zero-crossings to be useful.

The other piece of information is the idea that lightness and brightness

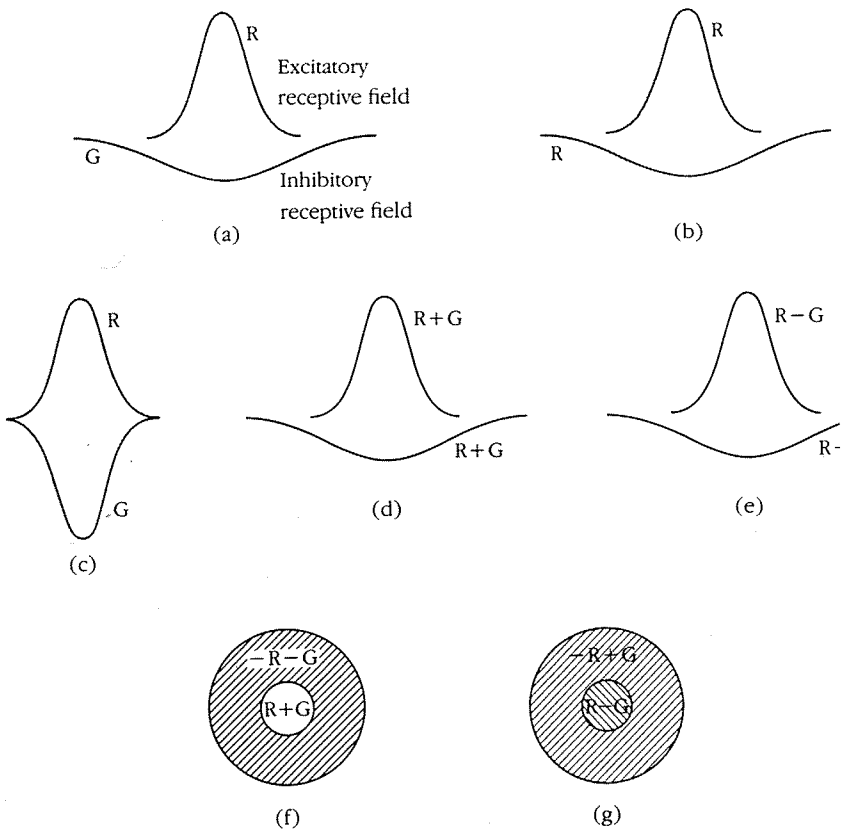


Figure 3–85. Various possible organizations of chromatic receptive fields. The organizatio assumed to be composed spatially of the difference of two Gaussian distributions. (a) So red–green opponent-color receptive field. (b) A red-center, red-surround receptive field. red–green opponent-color receptive field in which the spatial distribution of the two are ide (d) A pure luminance (red + green) receptive field. (e) A pure color-difference receptive (red − green). (f) The two-dimensional receptive field corresponding to (d). (g) The two-d sional receptive field corresponding to (e). R = red; G = green.

should be separated from color. Luminance boundaries correspond effectively to change in the summed contributions of the red and green channels, which we can write $(R + G)$. To detect these boundaries requires a $\nabla^2 G$ operator running on this sum, as illustrated in Figure 3–85(d). To detect changes in color, on the other hand, our hypothesis of the last section tells us to detect *relative* changes in the amounts of red and green

[illegible] can be done by a $\nabla^2 G$ operation on $(R - G)$, the difference between red and green signals, as illustrated in Figure 3–85(e).

Now the first type of cell, whose receptive field is illustrated in Figure 3–85(f), will not be very color selective, since its maximal stimulus will be a white central spot, and it can be turned off by any combination of red and green in the center and in the surround. The only criterion is that the effective luminances should balance.

The second type of cell is quite different, however. Its optimal stimulus would be a red center accompanied by a green surround, and it would therefore look like a color-opponent cell. Such a cell would respond best to changes in color and it should not respond at all to a pure white spot at its center, provided that the red and green are appropriately balanced in the white. Such a cell should respond to color boundaries but not to other boundaries. In order for such a cell to be insensitive to nonwhite lightness boundaries, like the boundary between two reds that differ only in the fraction, not the quality, of the light they reflect, the quantities R and G would have to be in logarithmic units. Such a cell would then act as a pure detector of changes in color. $\nabla^2 G$ operators are also insensitive to linear gradients.

Summary of the Approach

The main ideas of this approach, then, are to separate brightness from lightness and color and then to separate the estimation of lightness (average percent reflectance) from color (spectral distribution). Local changes may be recovered from zero-crossings in the lightness $(R + G)$ image and in the color image based on $(R - G)$ and $(B - G)$ (where B = blue).

The principal neurophysiological consequences are that no receptive fields should mix color and spatial variations in the manner of Figure 3–85(a); instead, receptive fields should exist as shown in the configurations of Figure 3–85(d) (for changes in lightness and brightness) and Figure 3–85(e) (for changes in color). Zero-crossing segment detection can subsequently occur in a similar way on each type of measurement yielding luminance contours from the first type, and color change contours from the second.

3.10 SUMMARY

In this chapter we have seen some of the quite striking variety of ways in which surface information is encoded in images, and we have explored as far as is presently possible how such information may actually be recovered. At the moment, the different processes appear to use slightly different input representations; the simplest, like directional selectivity, is driven by the zero-crossings, and the more elusive, like surface texture, probably involves the most complex aspects of the full primal sketch. I have summarized the discussion in Table 3–2.

Table 3–2. Processes producing surface information from image information and their probable input representations.

Process	Probable input representation
Stereopsis	Mainly ZC with eye movements helped by FPS
Directional selectivity	ZC
Structure from motion	FPS for correspondence; perhaps only RPS for detailed measurements
Optical flow	FPS(?) if process is used at all
Occluding contours	RPS, BC
Other occlusion cues	RPS
Surface orientation contours	RPS, BC
Surface contours	RPS, IC, GT
Surface texture	RPS, GT
Texture contours	BC
Shading	IC, RPS; possibly others

Note: BC = boundary contours created by discrimination processes and curvilinear aggregation of tokens; FPS = full primal sketch = RPS + GT + IC + BC. GT = group tokens, created by grouping processes in the full primal sketch; IC = illumination contours (shadows, highlights, and light sources); RPS = raw primal sketch (edges, blobs, thin bars, discontinuities, and terminations); ZC = zero-crossings, discontinuities, and terminations.

Another interesting aspect of all these processes is that, in addition to using slightly different input representations, they all involve slightly different assumptions about the world in order to work satisfactorily. As we have seen, in each case the surface structure is strictly underdetermined from the information in images alone, and the secret of formulating the processes accurately lies in discovering precisely what additional information can safely be assumed about the world that provides powerful

enough constraints for the process to run—for example, uniqueness and continuity in stereopsis, rigidity in motion, and so forth. Much of the art of formulating these processes lies in the precision and accuracy with which these additional constraints are expressed, and our survey has included some processes that I find satisfactorily formulated and others that remain puzzling and rather ill defined. The constraints assumed by the various processes have been set out roughly in Table 3–3, but the reader should bear in mind that few of these are certain. Thus the table should be regarded more as a guide to current thinking than as a definite statement of what allows these processes to run.

Finally, a few words about research strategy in this area. As we have seen, there are striking differences in the clarity and precision with which we have been able to formulate the different processes. Some are straightforward and clean, like stereopsis, structure from motion, and directional selectivity, whereas others, like visual texture and surface contour analysis, seem to be inherently muddy. That is not because the first kind are intellectually easier—on the whole they are not. For example, the mathematics associated with stereopsis or with structure from motion is not as easy as that associated with visual texture. Rather, the analytical difficulties arise from deciding what can be safely assumed about the world in order to help the processes interpret images of it. Where this can be done cleanly, more or less by inspection of the real world, we have on the whole been able to develop a clean theory. But where it cannot, I think there is no hope of understanding the processes properly until some other means have been found for determining what is safe to assume about the world and what is not, together with the related question of the reliability of the different kinds of information.

In the end, these are empirical questions, not so much about our visual systems (although the answers will be reflected in their structural design), as about the statistical structure of the visual world. I think that one will have to accept this, taking more of an engineering point of view when trying to answer them. As our knowledge of how to implement these early processes improves, we shall have to build fast machines that can run these processes in real time and acquire in that rather direct way a more detailed knowledge of which tricks pay off in practice and which do not. Studying vision is a mixture of studying processes and studying the world from this rather specialized point of view—something that natural evolution has been doing for a long time.

The first step is to build a unified system that employs all the processes that we currently understand, but much remains to be done before even this limited goal should be attempted. For one thing, processes like the construction of the raw primal sketch require a great deal of computational

Table 3–3. Guide to additional assumptions implicit in processes deriving surface information from images.

Process or representation	Implicit assumptions
Raw primal sketch	Spatial coincidence
Full primal sketch	Various assumptions about spatial organization of reflectance functions
Stereopsis	Uniqueness; continuity
Directional selectivity	Continuity of direction of flow
Structure from motion	Rigidity
Optical flow	Rigidity
Occluding contours	Smooth, planar contour generator
Surface contours	Surface locally cylindrical; planar contour generators
Surface texture	Uniform distribution and size of surface elements
Brightness and color	Only local comparisons reliable
Fluorescence	Uniform light source

power. Even the fastest general purpose machines are several orders of magnitude too slow for real-time vision, and although the emerging very large scale integration (VLSI) technologies will eventually provide the necessary power, the sensors and technology are not yet available and will not be for several years. And then, of course, there is the question of what one would do with the output of a machine that could run a set of processes like the ones described in this chapter. It is to this question that we now turn our attention.

CHAPTER 4

The Immediate Representation of Visible Surfaces

4.1 INTRODUCTION

In this chapter, we shall discuss the issues and problems surrounding the idea of the 2½-D sketch, whose acquaintance we have already made in Section 3.3. The central point is a simple one—that the 2½-D sketch provides a viewer-centered representation of the visible surfaces in which the results of all the processes described in Chapter 3 can be announced and combined. The construction of the 2½-D sketch is a pivotal point for the theory, marking the last step before a surface's interpretation and the end, perhaps, of pure perception.

The idea that such a representation might exist and that its construction can be regarded as the goal of early visual processing will probably strike the reader as unsurprising, especially since this book is written within precisely such a framework. But when we started out we had no such framework, and in trying to find a way of understanding what vision was, we were confused, having to grapple with almost philosophical difficulties concerning what perception was for. The reader who cares to examine Marr (1976) closely, for example, will find no explicit statement of what the primal sketch was for. He will find it more or less defined, justified on general grounds, and closely tied to physical reality. But the idea that the purpose of early vision is to recover explicit information about the visible surfaces was only implicit.

In fact, at that point, much of computer vision was in considerable disarray, because, with the exception of Horn's (1975) work, the idea that the main point of vision was to tell the shapes of things had not yet been taken seriously. And although perceptual psychologists like Gibson had the notion that surfaces are important, the idea of an internal representation obtained by certain processes was foreign to their thinking. In retrospect, our lines of thought and the kind of questions we asked at that time were rather muddled; inquiry had to do with feature-based recognition, how to separate figure from ground, how to extract and interpret a "form" or "figure," how much analysis could be done in a data-driven or bottom up way, and how much needed top–down influences. In addition, we had no coherent framework that allowed us to see how processes like stereopsis, shading, or motion perception could combine with one another and with the rest of vision to create what we call seeing.

All this type of thinking was dramatically swept away by the idea of the 2½-D sketch, which simultaneously resolved these and many other issues. It told us what the goals of early vision were, it related them to the notion of an internal representation of objective physical reality that *preceded* the decomposition of the scene into "objects" and all the concomitant difficulties associated with object recognition. At the same time, it hinted at the limits of what one might call pure perception—the recovery of surface information by purely data-driven processes without the need for particular hypotheses about the nature, use, or function of the objects being viewed. And finally, it provided the cornerstone for an overall formulation of the entire vision problem—the framework that this book has been written to explain and that has since enabled us to structure our research in a rational and strategic way.

For all these reasons, the emergence during the autumn of 1976 of the idea of the 2½-D sketch, which first appeared in Marr and Nishihara (1978, fig. 2) and was developed at length a little later (Marr, 1978, sec. 3), was for me the most exhilarating moment of the whole investigation. Its first positive consequence was the theory of stereo vision (Marr and Poggio, 1979) which was formulated during the first half of 1977. The reformulation of early visual processing was begun later that year, and of course, the 2½-D sketch ultimately led to the overall framework that we now have (Marr, 1978).

4.2 IMAGE SEGMENTATION

Perhaps the best way to introduce the whole question of the 2½-D sketch is to describe in some detail the impasse that it was intended to resolve. The neurophysiologists' and psychologists' belief that figure and ground constituted one of the fundamental problems in vision was reflected in the attempts of workers in computer vision to implement a process called *segmentation*. The purpose of this process was very much like the idea of separating figure from ground, the idea being to divide the image into regions that were meaningful either for the purpose at hand (which for computer vision might be assembling a water pump) or for their correspondence to physical objects or their parts.

Despite considerable efforts over a long period, the theory and practice of segmentation remained primitive for two reasons. First, it was well-nigh impossible to formulate precisely in terms of the image or even of the physical world what the exact goals of segmentation were. What, for example, is an object, and what makes it so special that it should be recoverable as a region in an image? Is a nose an object? Is a head one? Is it still one if it is attached to a body? What about a man on horseback?

These questions show that the difficulties in trying to formulate what should be recovered as a region from an image are so great as to amount almost to philosophical problems. There really is no answer to them—all these things can be an object if you want to think of them that way, or they can be a part of a larger object (a fact that is captured quite precisely in Chapter 5). Furthermore, however these questions were answered in a given situation did not help much with other situations. People soon found the structure of images to be so complicated that it was usually quite impossible to recover the desired region by using only grouping criteria based on local similarity or other purely visual cues that act on the image intensities or on something like the raw primal sketch. Regions that have "semantic" importance do not always have any particular visual distinction. Most images are too complex, and even the very simplest, smallest images like one depicting just two leaves (Marr, 1976, fig. 13) often do not contain enough information in the pure intensity arrays to segment them into different objects.

Despite the lack of any precise formulation of what it meant, the notion of segmentation continued to be investigated with increasingly complex techniques. It had been a long-standing view that visual perception was analogous to problem solving and should therefore involve the testing and modifying of hypotheses about the viewed object. This idea was common in computer vision (for example, see Minsky, 1975), and it had its counterpart in the psychology of vision (as exemplified by Gregory, 1970). The critical difference between this idea and the use of constraints as described in Chapters 2 and 3 is that, in the problem-solving approach, the additional knowledge or hypothesis that is brought to bear is not general but particular and true only of the scene in question and others like it. Instead of using things like rigidity, we make inferences such as: A black blob at desk level has a high probability of being a telephone.

Naturally, because of their specificity, any very general vision system must command a very large number of such hypotheses and be able to find and deploy just the one or two demanded by the particular situation. This prospect casts a whole complexion on the vision problem, in which the main questions to be addressed concern how to manage vast amounts of information in an efficient way. That is why so much effort was expended on the design of efficient program control structures* for deploying visual knowledge. Incidentally, for this type of reason people in other branches of artificial intelligence believe the problem of control to be an important one.

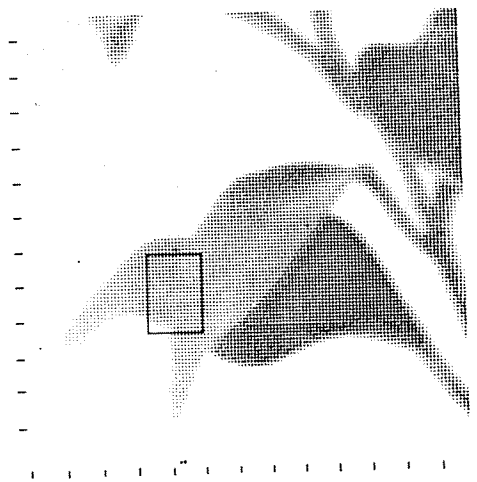
The main thrust of the then-current ideas was, therefore, to invoke specialized knowledge about the nature of the scene being viewed to aid segmentation of the image into regions that corresponded roughly to the objects expected in the scene. Tenenbaum and Barrow (1976), for example, applied knowledge about several different types of scene to the segmentation of images of landscapes, an office, a room, and a compressor. Freuder (1974) used a similar approach to identify a hammer in a simple scene. If this approach had been correct, then a central problem for vision would have been arranging for the availability of the right piece of specialized knowledge at the appropriate time during segmentation. Freuder's work, for example, was almost entirely devoted to the design of what was called a heterarchical control system that made this possible. A little while later, the constraint relaxation technique of Rosenfeld, Hummel, and Zucker (1976) attracted considerable attention for just this reason—it appeared to be a technique whereby constraints drawn from disparate sources could be applied to the segmentation problem while making the control processes required to manage the information only slightly more complex. Our own work on cooperative algorithms was also slightly colored by thoughts that they could perhaps be used to combine constraints from disparate sources, and this provided one of the motivations for trying to develop precise methods of analyzing the convergence of such algorithms (Marr, Palm, and Poggio, 1978).

*The interaction among subprocesses in a computer program.

What was wrong with the idea of segmentation? The most obvious flaw seemed to be that "objects" and "desirable regions" were almost never visually primitive constructions and hence could not be recovered from the primal sketch or other similar early representations without additional specialized knowledge. Edges that ought to be significant are either absent from an image or almost so (see, for example, Figure 4–1), and the strongest changes in an image are often changes in illumination and have nothing to do with meaningful relations in a scene. Given a representation like the primal sketch and the many possible boundary-defining processes that are naturally associated with it, which of all the possible boundaries should one attend to, and why? In order to answer these questions, it was necessary to discover precisely what information we should try to recover from an image and then to design a representation for expressing it.

In order to find the answer, it was necessary to go back to first principles, to return to the physics of the situation. As we have seen several times, the principal factors that determine the intensity values in an image are (1) the illumination, (2) the surface geometry, (3) the surface reflectance, and (4) the vantage point. At some stage, the effects of these different factors are separated.

The main argument was, therefore, as follows: Most early visual processes extract information about the visible surfaces directly, without particular regard to whether they happen to be part of a horse, or a man, or a tree. It is these surfaces—their shape and disposition relative to the viewer—and their intrinsic reflectances that need to be made explicit at this point in the processing, because the photons are reflected from these surfaces to form the image, and they are therefore what the photons are carrying information about. In other words, the representation of the visible surfaces should be carried out before knowing whether the surface belongs to a horse, man, or tree. As for the question of what additional knowledge should be brought to bear, general knowledge must be enough—general knowledge embedded in the early visual processes as



(a)

X =	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
Y																
58	171	169	167	167	166	165	166	164	167	171	171	174	174	175	173	171
57	168	168	168	167	166	167	167	165	169	168	174	176	175	175	175	172
56	168	167	167	165	166	166	167	167	168	170	178	177	176	174	174	173
55	168	168	165	169	167	168	167	165	168	175	177	177	175	175	172	171
54	169	170	167	169	169	168	163	166	172	169	174	173	175	178	173	173
53	171	169	170	168	169	168	169	168	168	170	175	173	175	177	178	176
52	172	171	170	168	169	169	167	168	173	172	173	177	174	175	178	176
51	172	174	171	170	166	168	167	168	172	172	172	177	179	172	175	175
50	171	167	176	169	170	169	168	169	171	172	174	174	173	173	174	178
49	174	172	173	173	173	174	171	171	172	174	172	172	172	169	173	173
48	173	173	173	176	178	172	171	174	174	173	175	175	175	173	173	171
47	173	175	178	173	173	171	171	175	175	177	178	175	174	173	175	178
46	178	175	174	169	173	175	177	175	177	177	174	175	176	177	177	174
45	173	175	173	174	172	173	174	175	174	171	173	174	175	174	172	171
44	177	174	175	175	172	171	172	176	172	173	172	172	173	170	170	175
43	173	171	174	168	176	172	173	173	173	174	171	174	175	173	174	174
42	175	173	171	172	170	171	176	175	178	172	174	175	175	175	175	172
41	181	179	177	172	170	170	169	179	175	174	175	174	172	175	174	175
40	188	184	179	178	176	176	176	174	172	178	172	174	173	172	174	173
39	195	191	188	186	185	183	180	177	178	175	174	176	175	174	176	176
38	200	199	197	193	190	187	185	180	176	175	180	177	175	175	176	177
37	202	202	199	202	199	194	187	180	175	179	177	176	174	175	176	173

(b)

Figure 4–1. (opposite) This image of two leaves is interesting because there is not a sufficient intensity change everywhere along the edge inside the marked box to allow its complete recovery from intensity values alone, yet we have no trouble perceiving the leaves correctly. The table shows the actual intensity values within the box. However, the surface is clearly discontinuous within the box. Consistency-maintaining processes operating in the 2½-dimensional sketch may be partially responsible for this.

general constraints, together with the geometrical consequences of the fact that the surfaces coexist in three-dimensional space.

Was there any chance that such an idea might work? In order to explore it, we needed to look at three questions. First, what might it mean to represent the visible surfaces? In order to answer this, we needed to preview the general classification of shape representations, which we shall spend more time on in the next chapter. Second, we needed to look at the information provided by psychophysics, both about the early processes that we studied in the last chapter and about whether there is any evidence that such processes are combined before the visible shapes are interpreted as objects. Third, we needed to look at the computational aspects of the problem. In what form do these early processes deliver information about the visible surfaces, and how might one combine all the different resources?

Part of our task in formulating the problem of intermediate vision is to examine ways of representing and reasoning about surfaces. We start our inquiry by discussing the general nature of shape representations. What kinds are there, and how may one decide among them? Although formulating a completely general classification of shape representations is difficult, we had already set out the basic design choices that have to be made when a representation is formulated. Three characteristics of a shape representation are largely responsible for determining the information that the representation makes explicit. The first is the type of coordinate system the representation uses—whether it is defined relative to the viewer or to the object being viewed; the second concerns the nature of the shape primitives used by the representation, that is, the elements whose positions the coordinate system is used to define. Are they two- or three-dimensional, in what sizes do they come, and how detailed are they? And the third characteristic is concerned with the organization a representation imposes on the information in a description—is it, for example, flat like an image intensity array, or does it have a hierarchical structure, like the full primal sketch of Chapter 2?

The first question about the coordinate system and the second about the shape primitives both have fairly straightforward answers. The coordinate system must be viewer centered, and the shape primitives must be two-dimensional and specify where the local pieces of surface are pointing. Briefly stated, the reason for this is that the information delivered by all the early visual processes of Chapter 3 depends upon aspects of the imaging process—for example, measures of depth, or surface orientation are obtained relative to the viewer, and so fall naturally into a viewer-centered coordinate frame. The second point is that all these processes tell about the visible surfaces, usually only locally, and so it is this information that needs representing, usually only locally. It is worth going into these points more deeply.

4.4 THE INFORMATION TO BE REPRESENTED

Vision, as we have already seen, provides several sources of information about shape. The most direct are stereopsis and motion, but surface contours in a single image are nearly as effective, and we have seen several examples of other, less effective cues. It often happens that some parts of a scene are open to inspection by some of these techniques and other parts by others. Yet different as the techniques are, they have two important characteristics in common: They rely on information from the image rather than on a priori knowledge about the shapes of the viewed objects, and the information they specify concerns the depth or surface orientation at arbitrary points in an image, rather than the depth or orientation associated with particular objects.

When viewing a stereo pair of a complex surface, like a crumpled newspaper or the "leaves" cube of Ittelson (1960), which is a box with leaves attached to the sides and pointing nearly at the viewer, we can easily state the surface orientation of any piece of the surface and whether one piece is nearer to or further from the viewer than its neighbors. Nevertheless, memory for the shape of the surface is poor, despite the vividness of its orientation during perception. Furthermore, if the surface contains elements lying nearly parallel to the line of sight, their apparent orientation when viewed monocularly can differ from the apparent surface orientation when viewed binocularly.

The reader can check this in a room with a textured ceiling: If you look at it with one eye through a narrow tube, any portion you see through the tube will soon come to be oriented apparently at a right angle to your line of sight. This impression persists despite the certainty of one's knowledge that it is false.

From these observations, we may draw some simple inferences:

1. There is at least one internal representation of the depth, surface orientation, or both associated with each surface point in a scene.

2. Because surface orientation can be associated with unfamiliar shapes, its representation probably precedes the decomposition of the scene into objects.

Table 4–1. Forms in which early visual processes would deliver information about surface geometry changes most naturally.

Process	Natural output form
Stereopsis	Disparity, hence δr, Δr, and s
Directional selectivity	Δr
Structure from motion	r, δr, Δr, and s
Optical flow	? r and s
Occluding contours	Δr
Other occlusion cues	Δr
Surface orientation contours	Δs
Surface contours	s
Surface texture	Probably r
Texture contours	Δr and s
Shading	δs and Δs

Note: r = relative depth (in orthographic projection); δr = continuous or small local changes in r;Δr = discontinuities in r; s = local surface orientation; δs = continuous or small local change in s;Δs = discontinuities in s.

3. Because the apparent orientation of a surface element can change, depending on whether it is viewed binocularly or monocularly, the representation of surface orientation is probably driven almost entirely by perceptual processes and is influenced only slightly by specific knowledge of what the surface orientation actually is. Our ability to perceive the surface much better than we can memorize it may also be connected with this point.

4. In addition, it seems likely that the different sources of information can influence the same representation of surface orientation.

In order to make the most efficient use of these different and often complementary sources of information, they need to be combined in some way. The computational question is, How best to do this? The natural answer is to seek some representation of the visual scene that makes explicit just the information that these processes can deliver.

Fortunately, the physical interpretation of the representation that we seek is clear. All these processes deliver information about the depth or orientation associated with surfaces in an image, and these are well-defined physical quantities. We therefore seek a way of making this information explicit, of maintaining it in a consistent state and perhaps also of incorporating into the representation any physical constraints that hold for the values which depth and surface orientation take over the kinds of surface that occur in the real world.

Table 4–1 lists the types of information that the different early processes can extract from images. The interesting point here is that although processes like stereopsis and motion are in principle capable of delivering depth information directly, they are in practice more likely to deliver information about local *changes* in depth, for example, by measuring local changes in disparity. Surface contours and shading provide more direct information about surface orientation. In addition, occlusion and brightness and size clues can deliver information about discontinuities in depth. The main function of the representation we seek is therefore not only to make explicit information about depth, local surface orientation, and discontinuities in these quantities but also to create and maintain a global representation of depth that is consistent with the local cues that these sources provide. We call such a representation the 2½-D sketch, and the next section describes a particular candidate for it.

4.5 GENERAL FORM OF THE 2½-D SKETCH

In order to provide an example of a representation as a basis for a more thorough discussion about the details of its composition, I will describe first the original proposal for a viewer-centered representation (this is the force of the word *sketch*) that uses surface primitives of one (small) size. It includes a representation of contours of surface discontinuity, and it has enough internal computational structure to maintain its descriptions of depth, surface orientation, and surface discontinuity in a consistent state.

Depth may be represented by a scalar quantity r, the distance from the viewer of a point on a surface. Surface discontinuities may be represented by oriented line elements. As we have seen, surface orientation may be represented as a vector (p,q) in two-dimensional space, which is equivalent to covering the image with needles. The length of each needle defines the slant (or dip) of the surface at the point, so that zero length corresponds to a surface that is perpendicular to the vector from the viewer to that point, and the length of the needle increases as the surface slants away from the viewer. The orientation of the needle defines the tilt, that is, the direction of the surface's slant. Figure 4–2 illustrates this representation; it is like having a gradient space at each point in the visual field.

In principle, the relation between depth and surface orientation is

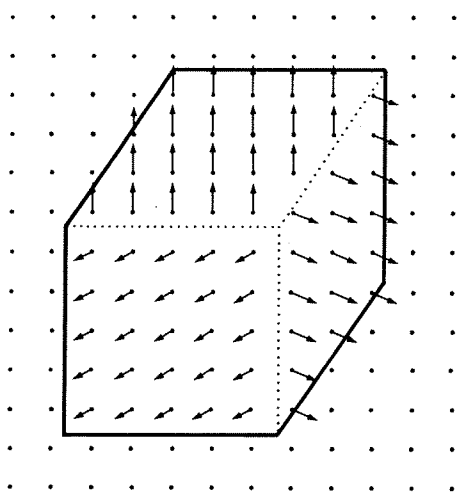


Figure 4–2. Another example of a 2½-dimensional sketch, this time of a cube. The surface orientation is again represented by arrows, as explained in the text and in the legend to Figure 3–12. Occluding contours are shown with full lines, and surface orientation discontinuities with dotted lines. Depth is not shown in the figure, though it is thought that rough depth is available in the representation.

straightforward—one is simply the integral of the other, taken over regions bounded by surface discontinuities. It is therefore possible to devise a representation with intrinsic computational facilities that can maintain the two variables of depth and surface orientation in a consistent state. But note that in any such scheme surface discontinuities acquire a special status (as curves across which integration stops). Furthermore, if the representation is an active one, maintaining consistency largely through local operations, curves that mark surface discontinuities (for example, contours that arise from occluding contours in the image) must be filled in completely, so that the integration cannot leak across any point along an object boundary. It is interesting that subjective contours have this property and that they are closely related to subjective changes in brightness often associated with changes in perceived depth. If the human visual processor contains a representation that resembles the 2½-D sketch, it would be interesting to ask whether subjective contours occur within it.

In summary, then, the argument is that the 2½-D sketch is useful because it makes explicit information about the image in a form that is closely matched to what early visual processes can deliver. We can then

formulate the goals of early visual processing as being primarily the construction of this representation. For example, specific goals would be to discover the surface orientations in a scene, which contours in the primal sketch correspond to surface discontinuities and should therefore be represented in the 2½-D sketch, and which contours are missing in the primal sketch and need to be inserted into the 2½-D sketch so that it is consistent with the structure of three-dimensional space. This formulation avoids all the difficulties associated with the terms *figure* and *ground, region* and *object*—the difficulties inherent in the image segmentation approach; for the gray-level intensity array, the primal sketch, the various modules of early visual processing, and finally the 2½-D sketch itself deal only with discovering the properties of surfaces in an image.

This outline raises many questions of detail, and we shall examine some of them in the next few sections. The reader, however, should be warned not to expect very precise answers. Our knowledge from here on is much less detailed than it has been up to this point. Unfortunately, I cannot provide much more than a framework within which to ask questions. Nevertheless, this has its value, even though denying the satisfaction of permanent answers. Thus, it is worth setting this description out with a little more precision than our discussion of the 2½-D sketch has had hitherto.

4.6 POSSIBLE FORMS FOR THE REPRESENTATION

There has not yet been any determined psychophysical assault on the 2½-D sketch, so we know very little about it or even whether it in fact exists in the sense suggested by our approach to vision. The main questions, however, are not difficult to formulate: What precisely is represented and how? What precisely is the coordinate system?—even saying that it must be viewer centered leaves one with several options. And perhaps most difficult, what kinds of internal computations are carried out within the representation either to maintain its own internal consistency or to keep it consistent with what is allowed by the three-dimensional world?

The first question is, Exactly what kind of surface information is made explicit? Are both depth r and surface orientation s represented, for example, or is only r actually carried in the representation, surface orientation being computed on demand by local differentiation? Or alternatively, is only surface orientation carried explicitly, depth being obtained somehow by local integration?—a more difficult possibility to accept but definitely different from the first alternative.

The best argument for the explicit representation of some function like distance from the viewer comes from the theory of stereopsis. The maximum range of disparities that are simultaneously perceivable without

(a)

(b)

Figure 4–3. A selection of large-disparity stereograms. The reader can test for himself what is the largest disparity for which he can simultaneously fuse both foreground and background. When viewed from 20 cm, these stereograms have disparities of (a) 2°, (b) 2.25°, (c) 2.5°, and (d) 2.75°.

diplopia is the same under four rather different conditions. First, in stabilized-image conditions,* Fender and Julesz (1967) obtained a figure of about 2° for a random-dot stereogram. Second, in the absence of any stabilization—that is, under normal viewing conditions—about the same range is obtained. When the complex stereograms given by Julesz (1971; for example, fig. 4.5–3) are viewed from about 20 cm, they give rise to

*Images are held fixed on the retinas so that eye movements have no effect.

(c)

(d)

Figure 4–3 (continued).

disparities of about the same order; if one views them from much closer, one cannot "see" all of them at once. Third, it seems at present unlikely that the maximum range of simultaneously perceivable disparities is much affected by their distribution. The reader can see for himself from Figure 4–3 that the figure of about 2°, which holds for stabilized-image conditions and for freely viewed stereograms with continuously varying disparities, also applies to stereograms with a single disparity. And fourth, if you experiment informally, using your fingers and real-world surfaces, you will arrive at a similar figure.

These examples suggest that the figure of about 2° for the maximal

range of simultaneously perceivable disparities has a rather general validity (provided there is enough surface at the extreme disparities) and that the figure is independent of eye movements. It is difficult to see how a memory buffer that stored only surface orientation could impose such a restriction, so I would conclude that depth is held in some form, perhaps only roughly, and that the amount that is being held corresponds to 2°–2¼° of disparity.

The second set of arguments concerning why depth should be represented explicitly in some form has to do with the importance of discontinuities in depth. Several early visual processes can yield information about such discontinuities, some of them in only a qualitative way. The most striking are probably occlusion cues, certain texture boundaries, disparity boundaries, and also directional selectivity (see Table 4–1). The perceptual vividness of subjective contours testifies to their importance. And subjectively, if two surfaces lie at very different depths, we seem to be very aware of this fact, even if they have the same surface orientations.

Both kinds of arguments suggest that some form of depth representation exists, and one interesting question is whether the range of simultaneously perceivable depths from apparent motion is commensurate with what we can see stereoscopically. But neither argument forcefully requires that depth information be held very accurately, as it would have to be if it formed the primary representation. Very locally we can easily say from motion or stereopsis information whether one point is in front of another. But if we try to compare the distances to two surfaces that lie in different parts of the visual field, we do very poorly and can do this much less accurately than we can compare their orientations.

This casts doubt, then, on the idea that depth is the basic represented variable, that it is stored accurately over a particular range of values, and that it is differentiated on demand to give surface orientation. There are better arguments against this possibility, too, which come from the fact that many of the processes listed in Table 4–1 yield information about surface orientation directly rather than via information about depth. The most obvious are surface contours, shading, and contours that deal with discontinuities in surface orientation. But in fact, stereopsis and structure from motion are both best suited to delivering information about how things are changing locally rather than about absolute depth—stereopsis because the brain rarely seems to know the actual absolute angle of convergence of the two eyes, dealing instead only with variations in it, and structure from motion because the analysis is local and orthographic, thus yielding only local changes in depth. There is therefore a strong sense in which both processes are very well suited to delivering surface orientation information, and it is probably more accurate to think of them in this way than as if they were primarily concerned with distance from the viewer.

Finally, we can judge surface orientation very accurately, to within a degree or two over the entire range of possible orientations (Stevens, 1979, app. B). This is not on its own conclusive evidence that we represent it explicitly, but taken in conjunction with our poor depth-judging abilities, I think that it is a significant fact which would require explanation if we did not represent it.

My conclusion from these arguments is that we likely represent both quantities s and r internally, but that although we may represent s quite accurately, we represent r only roughly. We may also have facilities for representing local differences in depth more accurately, which would be in addition to our representation of surface orientation.

4.7 POSSIBLE COORDINATE SYSTEMS

Perhaps we should next address the question of a coordinate system. We have already observed that it must be centered on the viewer, but this still leaves several possibilities. The first and most conspicuous point is that all the processes we have discussed are naturally retinocentric, as illustrated in Figure 4–4(a). Relative depth and surface orientation are obtained along and relative to the line of sight, not any external frame. So at least initially,

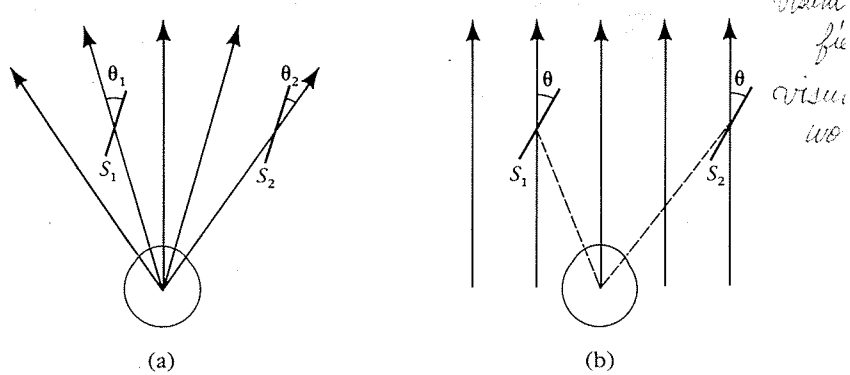


Figure 4–4. In retinocentric polar coordinates, the natural angle to measure a surface's orientation is that formed between the surface and the line of sight. Hence, as in (a), two parallel surfaces S_1 and S_2 are associated with different angles θ_1 and θ_2, respectively, which here have opposite signs. A much more convenient representation is to refer all angles to the direction straight ahead, as illustrated in (b). It is then easy to tell whether two surfaces are parallel and whether they are flat, convex, or concave.

- are almost forced to expect a retinocentric frame within which to express the results of each process.

On the other hand, it must be remembered that coordinates referring to the line of sight are not very useful to the viewer. Decisions about whether two surfaces have the same orientation or whether a surface is flat are not easily made from specifications in such a frame. One must continually allow for the angle of the line of sight, as illustrated in Figure 4–4(a)—a difficulty that is compounded by the effects of eye movements.

The second point, which follows from the first, is that although most early visual processes that deliver surface orientation information do so relative to the line of sight, each process may do so in its own way. In stereopsis, as we saw, there is a natural preference for specifying the components of surface orientation in the vertical and horizontal directions separately, simply because the horizontal positioning of the two eyes distinguishes these two directions. Surface contour and texture information prefer a slant-and-tilt representation of the sort discussed in Sections 3.6 and 3.7. Structure-from-motion information is probably like surface contour information in this respect.

To summarize, then, there are several different ways of representing surface orientation in a retinocentric coordinate frame, and the different early visual processes may use slightly different ones in which to express their own first guesses at what the surface orientation actually is.

The third point is that we have a fovea. Different parts of the visual field are analyzed at very different resolutions for a given direction of gaze. An important consequence of this is that the amount of memory or buffer space necessary to record the results of early visual processes varies widely in the visual field, being much greater for the fovea than for the periphery. This provides another reason for expecting a retinocentric frame, because if one used a frame that had already allowed for eye movements, it would have to have foveal resolution everywhere. Such luxurious memory capacity would be wasteful, unnecessary, and in violation of our own experience as perceivers, because if things were really like this, we should be able to build up a perceptual impression of the world that was everywhere as detailed as it is at the center of the gaze.

The final general point involves the question of consistency. We have already observed that the early visual processes can run independently to a large extent, and that some parts of the visual field will be accessible to some processes, and other parts to other processes. Therefore, the question of maintaining consistency among the different types of information will arise, as well as the question of assigning priorities that accurately reflect the reliabilities of the different processes, that is, assigning priorities so that the best source is believed when different sources are in conflict. This question of consistency should clearly be resolved as early as possible, because until it is, all the information cannot be reduced to just one representation.

These four observations lead to two conclusions. First, information from the different sources is probably checked for consistency and combined in some kind of retinocentric frame. This is because the information is all delivered in this form and because such a representation, containing among other things an enlarged foveal capacity, best matches the capabilities of the preceding processes.

Second, some conversion of the coordinate frame probably takes place at this point in order to express information from the different processes in a standard form and probably also to allow for the angle of gaze. An example of a suitable conversion is illustrated in Figure 4–4(b), where all angles are referred to the direction straight ahead instead of to the local line of sight. Such a conversion would (1) facilitate the computation of predicates like flat, convex, or concave; (2) allow easy comparison of the orientation of surfaces in different parts of the visual field; and (3) prepare the way for the business of allowing for eye movements.

4.8 INTERPOLATION, CONTINUATION, AND DISCONTINUITIES

The issues I wish to discuss next are based on three different types of psychophysical observation. The first is the observation, first studied in detail by White (1962), that one "sees" even a low density (2%–3%) random-dot stereogram as portraying a continuous surface, not as a set of isolated dots. The reader may confirm this for himself by looking at the 5% stereogram in Figure 3–8. The impression of a solid surface is strong. We are aware that the dots all lie at the same depth—they are clearly markings on an otherwise transparent sheet, which is flat and whose surface orientation is clearly apparent. This phenomenon is not altogether surprising in view of the theory of stereopsis described in Section 3.3, because the zero-crossings at which disparity is assigned do not cover the image—most of its area has no zero-crossings at all (examine Figure 3–14, for example)—so the notion that some kind of filling-in has to be carried out is to be expected. Notice, incidentally, that in the cooperative stereo algorithm of Figure 3–7, the filling-in process is incorporated into the algorithm, and this indeed was one of its initial attractions for us.

Eric Grimson (1979) has studied the filling-in or interpolation problem from a psychophysical and a computational point of view and has found that the visual system is very conservative in the amount of filling-in

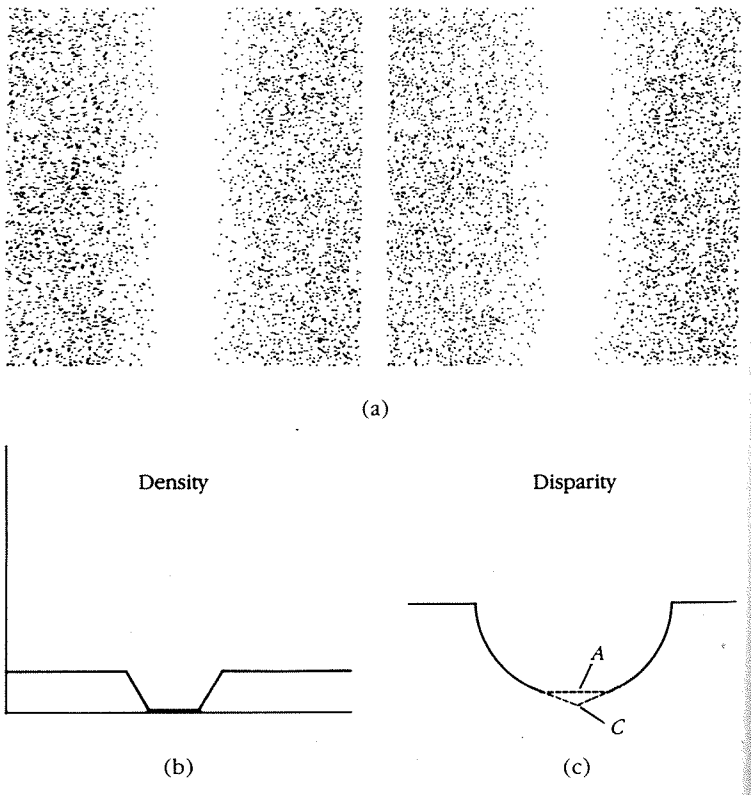


Figure 4–5. The stereogram (a) has the density distribution given in (b) and the disparity distribution indicated by the solid lines in (c). Such a stereogram can be used to explore psychophysically whether and how we interpolate across the gap. The dotted lines in (c) illustrate two interpolation possibilities.

it allows without additional evidence. He created various stereograms like the ones depicted in Figure 4–5, in which the density and disparity both decrease toward the center, as shown. The question is, How, if at all, does the observer fill in across the region where there are no dots? Two of the three possible candidates are shown in Figure 4–5(c): Candidate *A* fills in straight across with constant disparity; candidate *B* (not shown) produces some smooth interpolation that connects the two surfaces without any discontinuity in surface orientation; and candidate *C* continues the surfaces linearly until they intersect.

What the viewer perceives can be determined by putting a probe spot

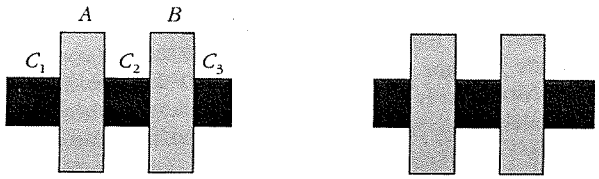


Figure 4–6. In this stereo pair, C_2 is seen at the same depth as C_1 and C_3, despite the fact that there are no disparity cues to the depth of C_2.

in the intermediate region at various disparities and asking the viewer whether it lies above or below the place "where the surface goes." Grimson found that the percept is unfortunately not a vivid one in these circumstances; although the subjects confidently exclude possibilities *A* and *C*, they are vague about the position of *B*. They never report any discontinuities in surface orientation. He concluded that although there seems to be some interpolation, the matter is not straightforward. I shall look at the computational side of the problem a little later.

The second aspect of the problem is what I shall call continuation, which is best illustrated by a stereo pair of Andrew Witkin's, shown in Figure 4–6. This stereogram is perceived as two rectangles *A* and *B* occluding a continuous rectangle containing C_1, C_2, and C_3. The curious thing about this demonstration is that the information about stereo disparity can come from only the vertical lines in the figure. Thus, regions *A*, *B*, C_1, and C_3 contain points at which the disparity is defined, and the fact that we see each as a whole surface is a problem only in interpolation. But for region C_2 there are no such cues. The fact that it is assigned the same depth as C_1 and C_3 must therefore be the result of some continuation process operating "behind" the occluding planes *A* and *B*. It is critically important for the demonstration that lines like the horizontal edges of C_1, C_2, and C_3 be in good alignment. It is as though their accurate alignment in the two-dimensional image allows them to be viewed as evidence of the same surface discontinuity in three dimensions, which then allows surface C_2 to be seen at the same depth as surfaces C_1 and C_3. A similar inference may perhaps be made from some experiments by Naomi Weisstein (1975), who displayed a drifting grating, occluded a central rectangular patch of it, and yet found adaptation effects occurring even within this patch.

These experiments suggest that the viewer-centered representation of surfaces may be capable of representing more than one surface at once. It may also be significant that in suitably constructed random-dot stereograms, like that given in figure 3–19(b), one can simultaneously and vividly see two surfaces. I personally cannot see three at once (compare Julesz, 1971, fig. 5.7–1), although there may be people who can.

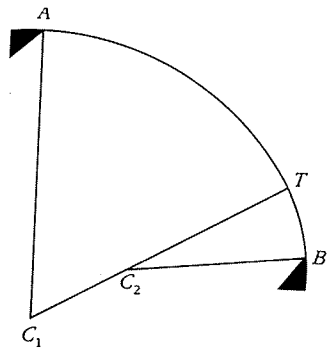


Figure 4–7. The shape of curved subjective contours. They are composed of two circles with centers C_1 and C_2, one emanating smoothly from each initiating point, A and B, that are joined smoothly (at T). Of the infinite number of pairs of circles with these properties, subjective contours follow the pair having minimal curvature.

Finally, there is the question of discontinuities in depth and in surface orientation. We have already mentioned discontinuities in depth in relation to the kind of continuation necessary for the phenomenon of Figure 4–6 and also in relation to the phenomenon of subjective contours. In both cases, continuity and smoothness (minimum curvature) seem to be important criteria. Ullman (1976a) examined phenomenologically the shape of curved subjective contours and found that it could be described accurately by two circles, one emanating from each source point, that are joined smoothly, as illustrated in Figure 4–7. Of the infinite family of pairs of circles with this property, one selects the pair yielding minimal curvature. Ullman also described a local network capable of generating this shape.

Although the shape of these contours is quite well understood, little is known about the conditions that cause their formation except the rather general notions that evidence of occlusion is required, together with rather direct monocular cues as to the exact position of the discontinuity. The Kanisza triangle (Figure 3–81e), the radial sun (Figure 2–25b) and the 5% random-dot stereogram (Figure 3–8), in which the dots themselves contain short, vertical edge segments, all provide both kinds of information in slightly different ways. The topic needs further psychophysical study.

4.9 COMPUTATIONAL ASPECTS OF THE INTERPOLATION PROBLEM

From a computational point of view, two problems need to be understood before planning detailed psychophysical experiments. The first is the notion of discontinuity, and the second, the different possibilities for interpolation.

Discontinuities

Although the distinction between a continuous and a discontinuous change over a continuum is a clear one, where the sample space is discrete the distinction is more elusive. We have already met this problem twice, once in detecting discontinuities in the orientation of zero-crossings where, strictly speaking, they cannot occur, and again in Land and McCann's (1971) lightness algorithm. In both cases one has to set a threshold. In the first case it was based on the point at which a "real" underlying discontinuity can no longer be discriminated from a very high curvature change. This point depends upon the receptive field size associated with the channel, so that what the smaller channels might "see" as smooth, the larger ones might "see" as discontinuous.

In an absolute sense, the resolution of the sample space does impose restrictions on what can be considered a continuous change. For example, in the one-dimensional case, suppose that the underlying representation consists of values specified a distance δ apart. Then by the sampling theorem, the representation cannot contain complete information about frequencies higher than, say, $\pi/\delta = \Omega$. Thus, the representation is effectively band limited by frequency Ω.

Now, although a signal that is band limited by frequency Ω can be represented completely by samples at intervals of δ, there is no guarantee that such a signal can accommodate all sample points at which one places arbitrary numbers. In other words, if the sample values change too fast, the overall signal may exceed the bandwidth of the representation. If this occurs, then the representation is forced to attribute the change to a discontinuity, since it is simply not rich enough to accommodate the changes that are actually occurring. This point is captured precisely by a theorem due to Bernstein, which says that the derivative of a band-limited function cannot get too large compared with the value of the function. If $f(x)$ is a function that is band limited by Ω, and if $f'(x)$ is its derivative, then the theorem states that

$$\sup |f'(x)| \leq \Omega \sup |f(x)|$$

That is, the largest value of $|f'(x)|$ over all x's is not bigger than the largest value of $\Omega|f(x)|$.

This constraint is a fundamental one that applies whenever we try to represent information on a discrete grid, and it is of particular interest here that the human visual system appears unable to represent sine waves in depth whose frequencies exceed 3–4 cycles per degree at the fovea

(Tyler, 1973). For example, the constraint may help to explain why subjective contours that do not appear or that are not very strong when we look at them directly appear much more vivid if we look at them indirectly. Presumably, the resolution of the representation also decreases with eccentricity, so that what can be represented foveally as a very steep gradient must, when presented more eccentrically, be represented as a discontinuity.

As we saw in Section 3.3, stereopsis can sometimes provide clear evidence for a surface discontinuity; if, for example, the horizontal rate of change of disparity, which we shall call d', reaches 1 in either eye, there is a discontinuity in depth as seen from the other eye. But in sparsely featured images, there is often not enough information to decide even this. Perceptually one may be left with a vague feeling that the disparity does change but no exact impression of where. In a sparse random-dot stereogram, if two squares happen to line up along a disparity boundary, vivid subjective contours are formed and the boundary is clearly delineated; however, if the squares in the stereogram are replaced by blurred dots, for example, the perception of the discontinuity is much less vivid.

Although these observations are little more than suggestions, they do hint that the interpolation process is conservative and that the visual system is reluctant to insert contours of discontinuity in either depth or surface orientation unless the image itself provides reasonable evidence of their positions. A contour may not be evident all along its length, but it is unlikely that direct visual evidence of it will be lacking everywhere along it. Eric Grimson (1979) enshrined this view in a dictum, which states that *places of no information are actually places of information.* In other words, one cannot hide discontinuities, and conversely, if the image provides no evidence at all about the presence of a discontinuity, not even an edge fragment anywhere along where one is expected, then such a discontinuity may not be assumed. Hence, in contrived situations where direct evidence is deliberately removed, as in Figure 4–5, we neither insert contours nor interpolate the surfaces in a definite way, and we are thus left with a vague and unsettled perception.

Interpolation Methods

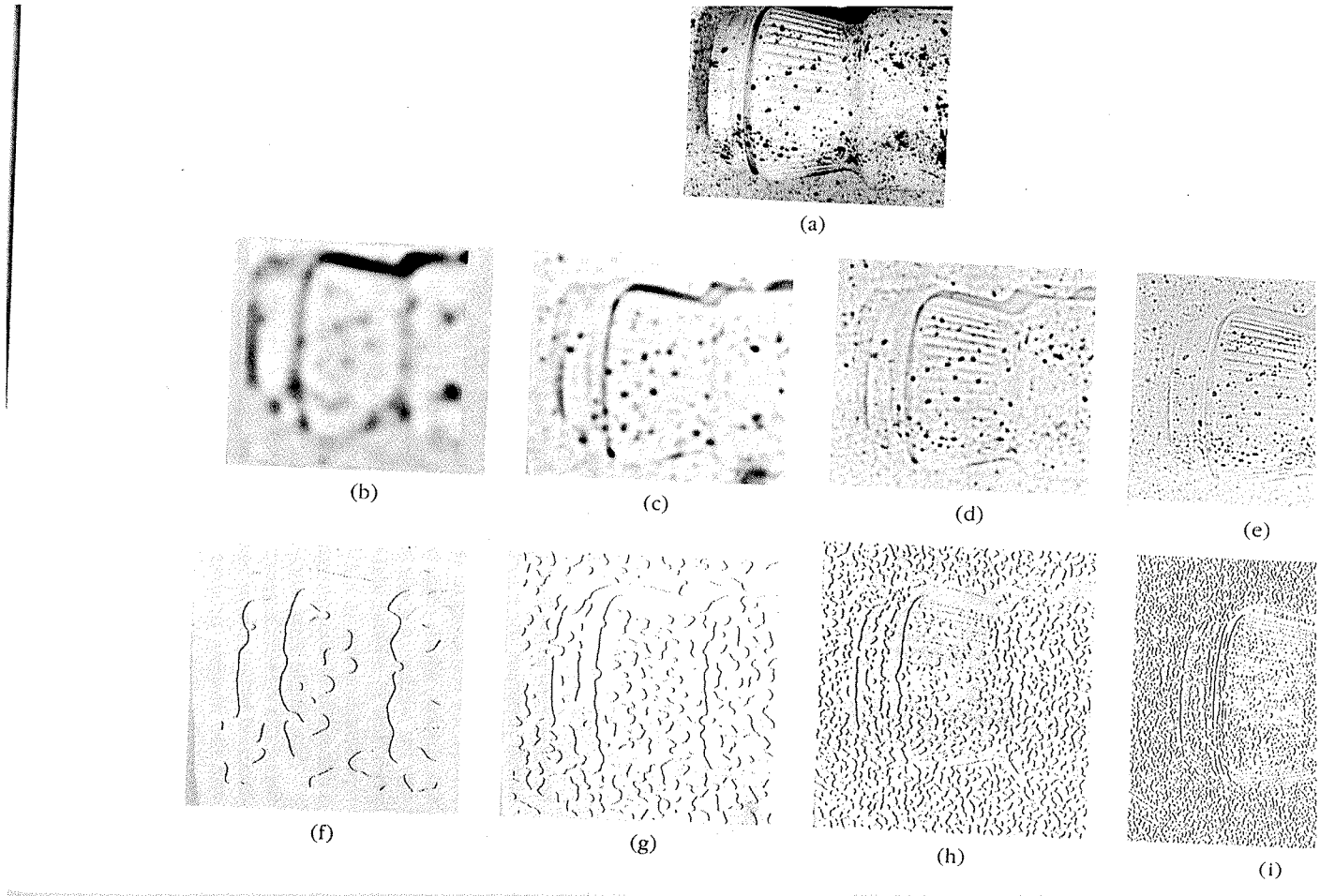
Three principal interpolation methods deserve notice: (1) linear interpolation in depth r, (2) linear interpolation in surface orientation, and (3) "fair surface" interpolation, which is a method used by car makers to give car bodies a smooth shape. Very roughly, method 1 is similar to the inverse transform we met in Horn's (1974) algorithm for the retinex. It tries to minimize the value of the Laplacian operator ∇^2 on the surface. Method 2 approximately minimizes the first curvature of the surface in any given concave or convex region. (This follows from the facts that the first curvature $J = -\text{div } n$, where div n is the divergence of n, n is the surface normal, and that locally averaging n almost minimizes div n.) The objection to both methods 1 and 2, implemented on a grid, is that convergence rates are slow—quadratic, in fact, with the distance between fixed points of the computation. I have already stated my reservations about the use of iterative methods in perceptual computations (see Sections 3.2 and 3.5).

The third possibility, favored over the other two by Grimson, involves the notion of a fair surface, which is a surface whose first and second derivatives vary continuously but which allows discontinuities in the third and higher derivatives. One-shot methods are available for filling-in between neighboring triplets of points and knitting them together along the seams, so as to preserve smoothness in arbitrarily high-level derivatives. Choice of the second derivative as the cutoff point rests on the empirical observation of car designers that customers notice discontinuities in the first and second derivatives of a surface but not in the third. Figure 4–8 illustrates the result of applying a filling-in method of this kind to the output derived from a stereo pair. It gives a smooth and pleasing appearance.

As for the connection between these computational ideas and the truth of how we ourselves find discontinuities or fill in surfaces—to the limited extent that we do these things—these are questions for the future.

4.10 OTHER INTERNAL COMPUTATIONS

The notion of surface continuity may, as we have seen, give rise to various active computations in the 2½-D sketch, including filling-in and the smooth continuation of discontinuities. We would expect other local constraints to be embedded there in a similar way—for example, consistency relations concerning the possible arrangements of surfaces in three-dimensional space, such as the constraints made explicit by Waltz (1975; recall Figure 1–3). Such constraints may eventually form the basis for an understanding of phenomena like the reversal of the Necker cube. From this point of view, it is natural that many illusions concerning the interpretation of three-dimensional structure (the Necker cube, subjective contours, the Muller–Lyer figure, the Poggendorff figure, and so on) should take place after stereoscopic fusion (see Julesz, 1971; Blomfield, 1973). Illusions like the reversing bucket of Figure 5–9 should also have part of their cause here, since the continuity of the bucket's surface plays a critical role in keeping its appearance consistent. The interesting questions here concern how much



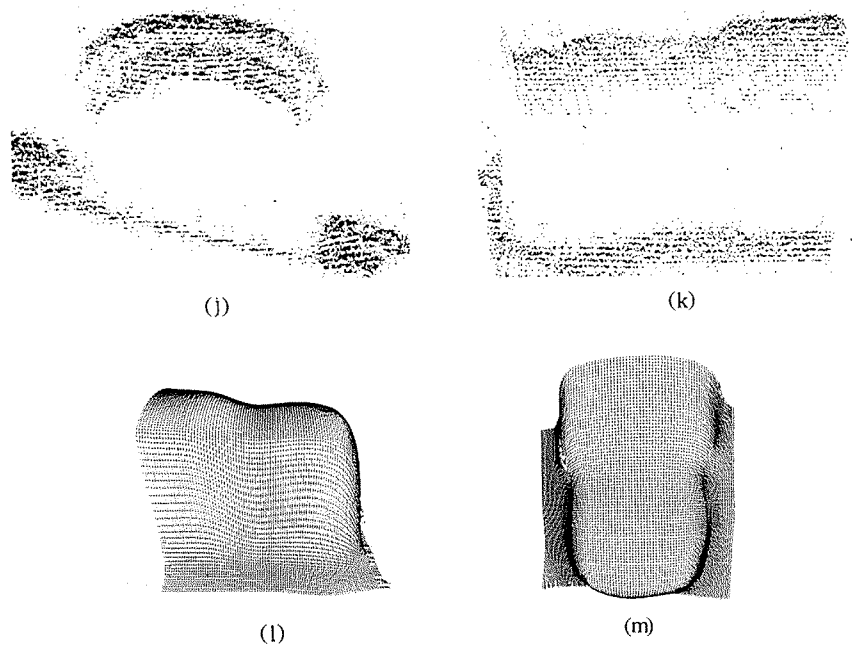


Figure 4–8. (a) Shows one of the images from a stereo pair. (b)–(e) Show its convolution with $\nabla^2 G$ filters of four sizes, and (f)–(i) display the zero-crossings thus obtained. (j) and (k) show two different views of the disparity map obtained after stereo matching, and (l) and (m) show the surface obtained from this information by Eric Grimson's interpolation algorithm.

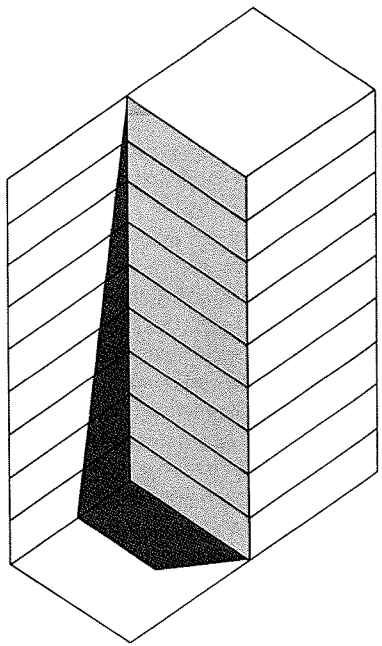


Figure 4–9. The strange reversal of this figure may, like the reversal of the Necker cube, be due to constraints embedded in the 2½-dimensional sketch.

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

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 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.

*A coordinate frame uniquely determined by the shape itself.

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,

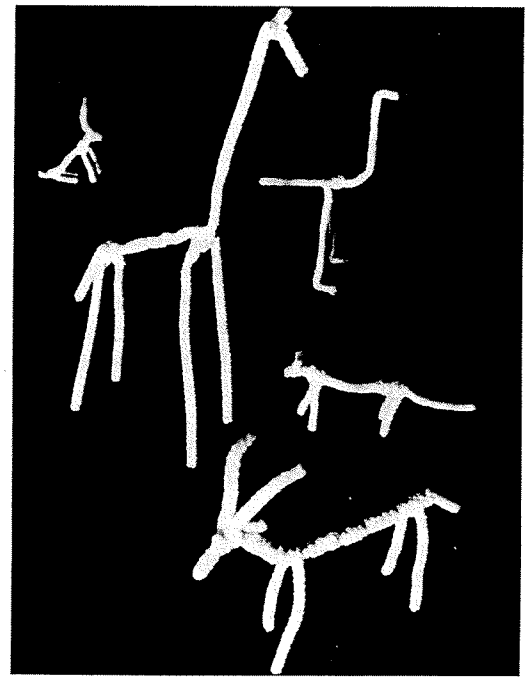


Figure 5–1. These pipe cleaner figures illustrate several of the points developed in this chapter. A shape representation does not have to reproduce a shape's surface in order to describe it adequately for recognition; as we see here, animal shapes can be portrayed quite effectively by the arrangement 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-

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.

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

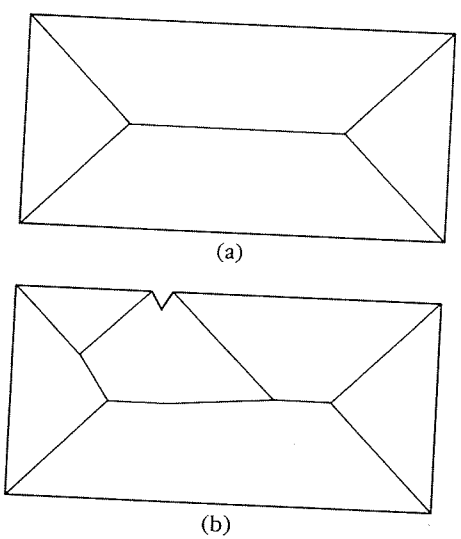


Figure 5–2. Blum's (1973) grassfire technique for recovering an axis from a silhouette. It can be thought of as lighting a fire at the boundary, the axis being defined as where two configurations meet. However, the technique is undesirably sensitive to small perturbations in the contour. (a) Shows the 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

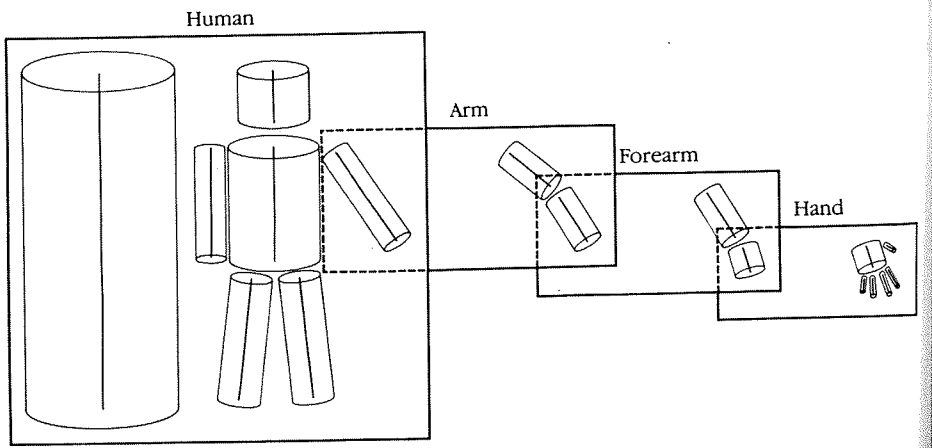


Figure 5–3. This diagram illustrates the organization of shape information in a 3-D model description. Each box corresponds to a 3-D model, with its model axis on the left side of the box and the arrangement of its component axes on the right. In addition, some component axes have 3-D models associated 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-

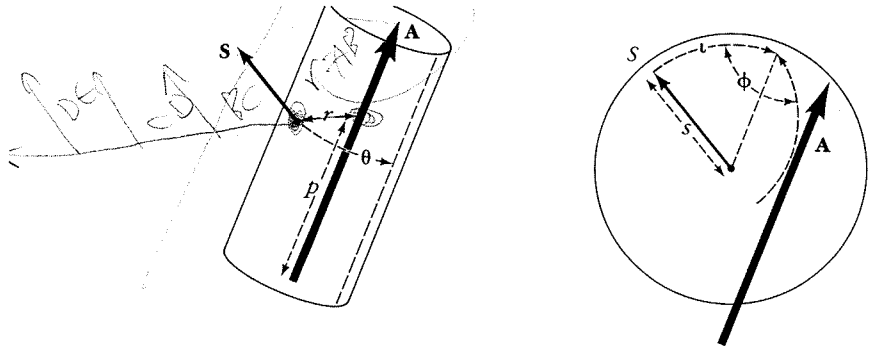
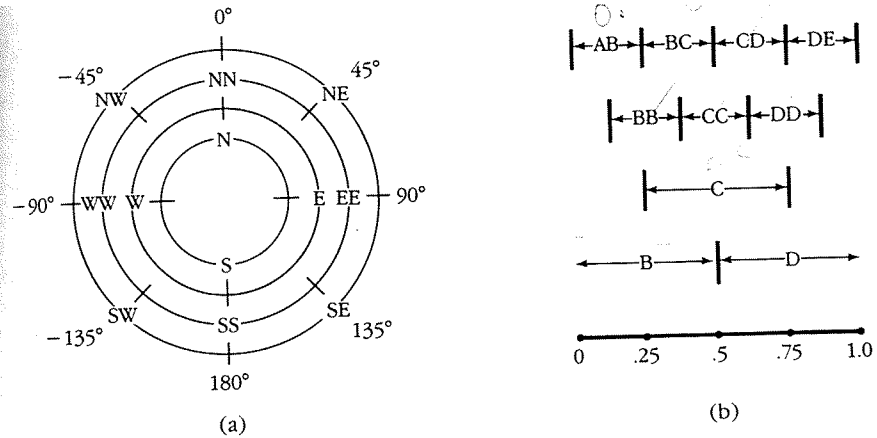


Figure 5–4. The spatial organization of a 3-D model's axes is specified in terms of pairwise relationships between those axes that we call adjunct relations. The disposition in space of one axis **S** is determined relative to another, **A**, by specifying the location of one of its endpoints in a cylindrical coordinate system (p,r,θ) about **A** as shown on the left, and its orientation and length in a spherical coordinate system (ι,ϕ,s) centered on that point and aligned with **A** as shown on the right. (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.)

ified relative to the model axis in order to maintain the connectedness of the distributed coordinate system.

Two three-dimensional vectors are required to specify the position in space of one axis relative to another. One way of doing this is illustrated in Figure 5–4, which represents the position of a vector **S** relative to an axis vector **A** by means of two vectors. The first vector, written in cylindrical coordinates (p, r, Θ), defines the starting point of **S** relative to **A** (Figure 5–4a); the second vector, written in spherical coordinates (ι, ϕ, s), specifies **S** itself (Figure 5–4b). We shall call the combined specification $(p, r, \Theta, \iota, \phi, s)$ an *adjunct relation* for **S** relative to **A**.

Because the precision with which 3-D models can represent a shape varies, it is appropriate to represent the angles and lengths that occur in an adjunct relation in a system that is also capable of variable precision. For instance, one might wish to state that a particular axis, like the arm component of the human 3-D model in Figure 5–3, is connected rather precisely at one end of the torso (that is, the value of p is exactly 0), but with Θ only coarsely specified and with very little restriction on ι. An example of a suitable system incorporating variable precision is illustrated in Figure 5–5.



A	S	p	r	θ	ι	ψ	s
Model	Torso	BC	AB	NN	NN	NN	CC
Torso	Head	DE	AB	NN	NN	NN	BB
Torso	Arm	DE	BB	EE	E	E	DD
Torso	Arm	DE	BB	WW	E	W	DD
Torso	Leg	AB	BB	EE	SS	NN	DE
Torso	Leg	AB	BB	WW	SS	NN	DE

(c)

Figure 5–5. Angle and distance specifications in an adjunct relation must include tolerances so that specificity of these parameters can be made explicit in the representation. One way to do this is shown in the upper diagrams, which associate symbols with (a) angular and (b) linear ranges. An example of adjunct relations for the human 3-D model in Figure 5–3 that are expressed in these symbols is shown in the table (c). **A** and **S** identify the two axes related by the adjunct relation specified along each row of this table. If the mnemonic names listed under **A** and **S** were replaced by internal references to the corresponding 3-D models whenever they exist and left blank otherwise, this table would show essentially all the information carried by a 3-D model. (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.)

5.4 NATURAL EXTENSIONS

These representational ideas, perhaps best epitomized by the hierarchical scheme depicted in Figure 5–3, begin to show how the complexities of shape description may be approached. Perhaps if J. L. Austin had seen such a figure, he would not have thrown up his hands in such despair at the

prospect of formulating rules for representing the shape of his cat (see Section 1.2)! Nevertheless, the ideas are still quite crude, and little work has gone into their development since 1977, mainly because we have been preoccupied with the details of early visual processing. However, questions have frequently arisen about ways of generalizing these ideas, and while answers have not yet been developed in detail, it is worth indicating briefly the most obvious directions in which the representation can be extended.

Perhaps the first point is that one can represent two-dimensional configurations just as easily as three provided, of course, that the patterns are endowed with a natural axis of elongation or of symmetry. Thus we can as easily represent a two-dimensional drawing of a face as the features and details on a real three-dimensional head. A primitive example appears in Figure 5–6. It is particularly interesting to note in this connection that the existence of symmetry in a pattern yields a canonical axis but not a canonical direction along the axis. We still have to decide which end is 0 (down) and which is 1 (up). This choice has to be made when one starts to construct a particular 3-D model, and we seem to make this final choice ourselves using the direction that we are currently taking to be up—usually it is vertically up. If you construct a detailed face description while adhering to this convention and then stand on your head, the details become completely unrecognizable, perhaps because the innate choice mechanism is now using the opposite convention! In addition, face recognition seems to be a rather accurate, specialized, and late-developing process in humans, and interested readers should consult Carey and Diamond (1980) and other works on the subject.

The second point is that the primitives of the 3-D model representation can be extended to include surface primitives, roughly of two kinds. First would be just rough, two-dimensional rectangular surfaces of various sizes, including elliptical shapes and circular ones. Not very many primitives would be needed by the average man, although presumably a sculptor like Henry Moore has a repertoire of hundreds. The second kind of primitive is the notion of something that is not solid but hollow—like a tube or cup, for example. It is not hard to see how such primitives may be organized along much the same lines as the original 3-D model representation, and Figure 5–7 illustrates some preliminary ideas about how such a vocabulary may be deployed to represent various common objects. If we also admit curved axes into the representation, much can be done to represent the more common objects we encounter in everyday life (see Figure 5–7a, and especially Hollerbach, 1975).

The other major directions in which these ideas need to be extended concern not so much the spatial arrangement of a given shape but the spatial configurations formed by several separate objects. These will need

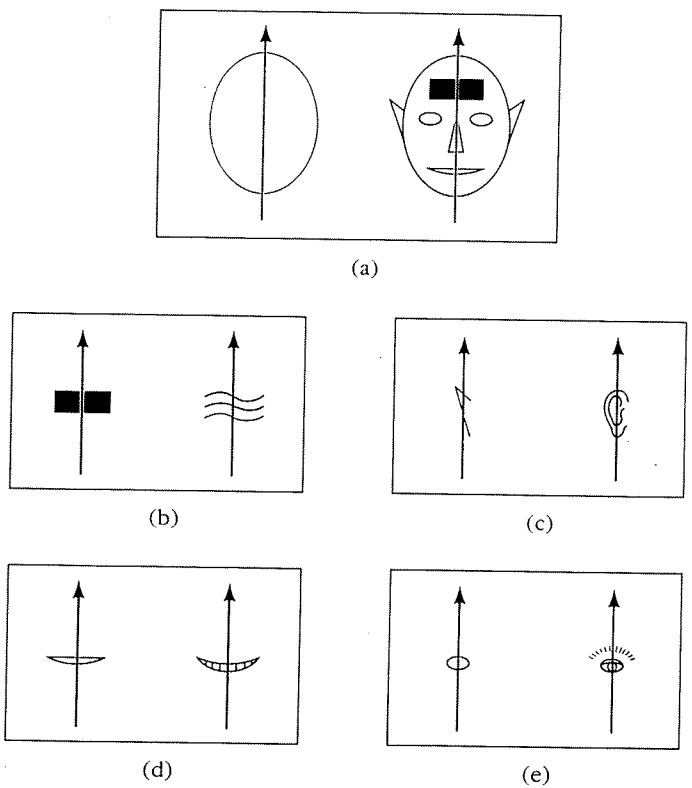


Figure 5–6. The 3-D model for a two-dimensional pattern portraying a face. (a) The overall 3-D model, with the axis determined by symmetry. (b)–(e) Possible 3-D models for the pattern's principal constituents.

at least three types of description. One is the incorporation of their positions in a standard space frame around the viewer in terms of angles and distances from him. Another is the representation of configurations of objects relative to the viewer, for example, the notion that you and two other people happen to create an equilateral triangle. The critical point here is that the position of the viewer is involved and that angular relations—the internal structure of the configuration—are made explicit. Finally, there is the representation of the relative positions of a number of external objects without particular reference to the viewer. For example,

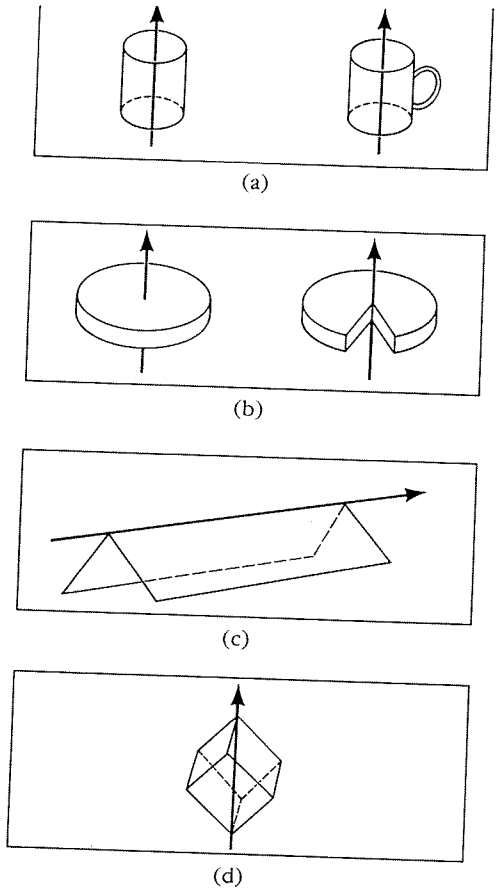


Figure 5–7. Some 3-D models for more complex shapes (a). (b) and (c) may require surface primitives in the representation. (d) Illustrates the representation of a familiar object (a cube) obtained by G. Hinton's unusual choice of axis (a diagonal from one vertex to the opposite one).

three trees might lie in a row, or four buildings form a square. The underlying problems here are exactly the same as those we have already met—how to choose an appropriate canonical coordinate frame within which to make explicit the spatial relations of the configuration.

It is already clear how to approach representational problems of this sort, and for the designers of a vision machine I do not think that these questions will raise any insurmountable difficulties. The major scientific obstacles here, it seems to me, are how to discover what systems and schemes are actually used by humans. I do not expect the answers to be very surprising, but at present I see no empirical way of approaching this type of problem. It seems to be much more difficult to design experiments to answer questions at these rather high levels of analysis than at the lower ones. In fact, perhaps we could say that at these higher levels we are beginning to face all the problems that the linguists have. Designing a successful empirical approach to such questions would represent a major breakthrough.

5.5 DERIVING AND USING THE 3-D MODEL REPRESENTATION

The advantages of modularity, which has been one of our major concerns in the design of the 3-D model representation, will become especially visible as we discuss the processes that derive and use the representation for recognition. In particular, none of the processes have to deal with the internal details of more than one 3-D model at a time even if the complete description of a shape involves many 3-D models. We begin by examining the basic problems associated with identifying a model's coordinate system and its component axes and transforming the viewer-centered axis specifications into specifications in the model's coordinate system. We then treat the task of recognizing this description as a problem of indexing into a catalogue of stored 3-D model descriptions. Finally, we consider the interaction between the process that derives a 3-D model description and the recognition process. The ambiguities introduced by the perspective projection often mean that only coarse specifications of the lengths and orientations of a shape's axes are directly accessible from its image. However, if the recognition process in conjunction with the derivation process, is conservative—so that all the information recognition recovers is reliable—the early stages of the recognition process can make additional constraints available so that a more precise description can be produced.

Deriving a 3-D Model Description

To construct a 3-D model, the model's coordinate system and component axes must be identified from an image, and the arrangement of the component axes in that coordinate system must be specified.

Even if a shape has a canonical coordinate system and a natural decomposition into component axes, there is still the problem of deriving these

features from an image. At present we do not have a complete solution to this problem, but some results have been obtained for shapes that fall within the scope of the 3-D model representation. For example, we saw in Section 3.6 that the image of a generalized cone's axis may be found from the occluding contours in an image provided that the axis is not too foreshortened. An example of the decomposition formed by this method appears in Figure 5–8, and a brief description is given in the legend. Notice that the final decomposition (Figure 5–8f) was derived from the contour (Figure 5–8a) without knowledge of the three-dimensional shape apart from the assumption that it is composed of generalized cones. The method can therefore be used to find the component axes for the 3-D model of a shape that has not been seen before.

This result is somewhat limited, but so is the information it uses, namely, the contours formed by rays that are tangential to the side of a smooth surface. Interestingly, as we saw in Section 3.2, these particular contours are unsuitable for use in either stereopsis or structure-from-motion computations, because they do not correspond to fixed locations on the viewed surface. Creases and folds on a surface also give rise to contours in an image, and these have yet to be studied in detail. Similarly, much work remains in the study of how to use information about shape from shading and texture.

A major difficulty in the analysis of images arises when an important axis is obscured because it is either foreshortened or hidden behind another part of the shape. For example, although the torso-based coordinate system for the overall shape of a horse is easily obtained from a side view, it is difficult to obtain when the horse faces the viewer. There are three ways of dealing with such a situation. The first is to allow for recognition the use of partial descriptions based on the axes visible from the

Figure 5–8. (opposite) The occluding contours of simple shapes composed of generalized cones can be used to locate projections of the natural axes of the cones provided that the axes are not severely foreshortened. One algorithm for doing this is shown in this example from a program written by P. Vatan. The initial outline in (a) was obtained by applying local grouping processes to the primal sketch of an image of a toy donkey. This outline was then smoothed and divided into convex and concave sections to get (b). Next, strong segmentation points like the deep concavity circled in (c) were identified and a set of heuristic rules used to connect them with other points on the contour to get the segmentation shown in (d). The component axes shown in (e) were then derived from these. The thin lines in (f) indicate the position of the head, leg, and tail components along the torso axis, and the snout and ear components along the head axis. (Reprinted by permission from D. Marr and H. K. Nishihara, "Representation and recognition of the spatial organ[illegible] of three-dimensional shapes" *Proc. R. Soc. Lond. B* 200, 269–294.)