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 diffi-

culties 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 wellnigh 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 coun-

terpart 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.

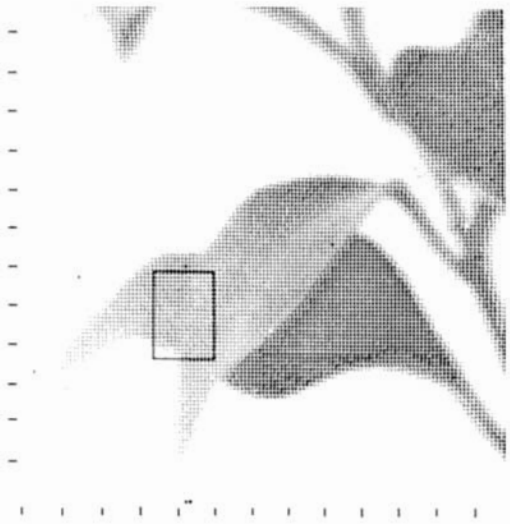
4.3 REFORMULATING THE PROBLEM

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

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.



(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)

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

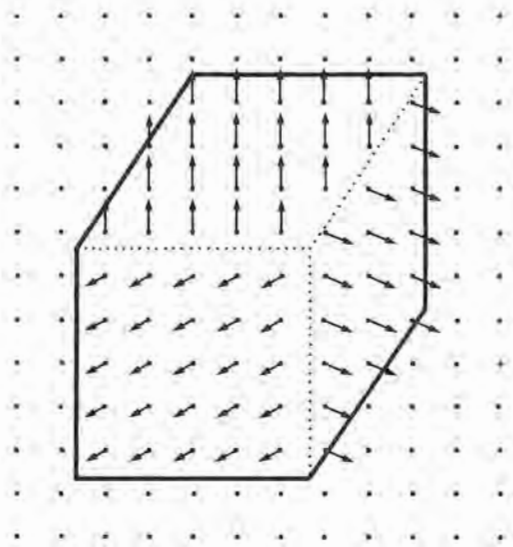


Figure 4–2. Another example of a 2½-dimensional sketch, this time of a cube. The surface orientation is again represented by arrows, as explained in the text and in the legend to Figure 3–12. Occluding contours are shown with full lines, and surface orientation discontinuities with dotted lines. Depth 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