Marr, D. (1982). *From images to surfaces*. In Vision. (pp. 99-111).
San Francisco: Freeman.

CHAPTER 3

From Images to Surfaces

3.1 MODULAR ORGANIZATION OF THE HUMAN VISUAL PROCESSOR

Our overall goal is to understand vision completely, that is, to understand how descriptions of the world may efficiently and reliably be obtained from images of it. The human system is a working example of a machine that can make such descriptions, and as we have seen, one of our aims is to understand it thoroughly, at all levels: What kind of information does the human visual system represent, what kind of computations does it perform to obtain this information, and why? How does it represent this information, and how are the computations performed and with what algorithms? Once these questions have been answered, we can finally ask, How are these specific representations and algorithms implemented in neural machinery?

The study of working visual systems can help us in this endeavor, and nowhere is this clearer than in the study of visual processes. At the level of computational theory, the investigator's first question is, What computational problems are being solved, and what information is needed to solve them?

As usual, the point is best made with an example. Because of how our eyes are positioned and controlled, our brains usually receive similar images of a scene from two nearby points at the same horizontal level. If two objects are separated in depth from the viewer, the relative positions of their images will differ in the two eyes. You can see that this is so by holding your thumb at various distances from your eyes against a background. Closing first one eye and then the other will then convince you that objects in the world have somewhat different positions in the images cast upon each of your retinas. The relative difference in position is called *disparity*; it is usually measured in minutes of arc, and the disparity between the images of your thumb and the background in your two eyes increases as you move your thumb nearer to you. One minute of disparity roughly corresponds to a depth difference of 1 in. for an object 5 ft away.

The brain is capable of measuring disparity and using it to create the sensation of depth. For purposes of demonstration, a stereoscope from a souvenir shop will do: When individual views are seen with just one eye at a time, they look flat. However, if you have good stereo vision and look with both eyes, the situation is quite different. The view is no longer flat: The landscape jumps sharply into relief, and your perceptions are clearly and vividly three-dimensional.

How does stereo vision work? Unfortunately, we cannot even begin to ask the right questions from just the evidence described above. The reason is that from the experience of everyday life or even from the small experiment with the stereoscope, it is not at all clear how separate stereoscopic processing is from the more familiar, monocular analysis of each image. If stereo processing were an isolated module, so to speak, then one could tackle it on its own. But it may not be isolated—for example, stereo vision could involve a complicated and gradually increasing interaction between the individual processings of each eye and a comparison of the results between the two eyes. This is not as absurd as it seems. It does not take much imagination to see how such a scheme might work. We could start by finding, for example, the images of an oak tree as seen independently by the left and right eyes. Then we could find the trunk in each image and then, perhaps, the lowest branch on the right hand side of the trunk. Pretty soon we would have correspondences between the small details of the left and right images whose disparity could be measured accurately. And because the match has been obtained in this general-to-specific way, there is never any real problem in deciding what should match what.

This type of approach, incidentally, is typical of the so-called top–down school of thought, which was prevalent in machine vision in the 1960s and early 1970s, and our present approach was developed largely in reaction to it. Our general view is that although some top–down information is sometimes used and necessary (see Figure 3–1 and Marr, 1976, fig. 14), it is of only secondary importance in early visual processing. The evidence for this comes from psychophysics and for some reason was willfully ignored by the computer vision community. The argument suggested by this evidence is a simple one. If, using the human visual processor, we can experimentally isolate a process and show that it can still work well, then it cannot require complex interactions with other parts of vision and can therefore be understood relatively well on its own.

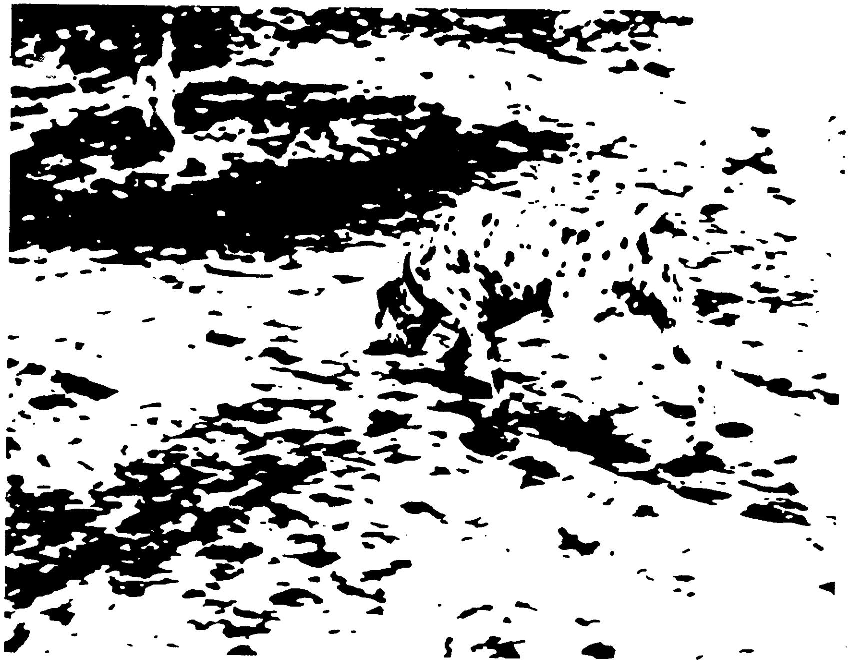

Figure 3–1. The interpretation of some images involves more complex factors as well as more straightforward visual skills. This image devised by R. C. James may be one example. Such images are not considered here.

One way of isolating a visual process is to provide images in which, as much as possible, all kinds of information except one have been removed and then to see whether we can make use of just that one kind. Bela Julesz did this for stereopsis by inventing the computer-generated random-dot stereogram, which we met in Figure 1–1. Both the left and right images shown there are computer-generated assemblies of black and white squares that are identical except for a centrally located, square-shaped region shifted horizontally in one image relative to the other. That is, it has a different disparity. The stereo pair contains no information whatever about visible surfaces except for this disparity.

When the pair is viewed stereoscopically and fused, one vividly and unmistakably perceives a square floating in space above the plane of the background. This proves two things: (1) Disparity alone can cause the sensation of depth, and (2) if there is any top–down component to the processing (and, in fact, we think that there probably is a little), it must be of a very limited kind, because neither image contains any recognizable large-scale monocular organization.

This observation—which is qualitative rather than quantitative, not at all technical, and, like many of Julesz's demonstrations, absolutely and strikingly convincing to behold—is fundamental to our approach, for it enables us to begin separating the visual process into pieces that can be understood individually. Computer scientists call the separate pieces of a process its *modules,* and the idea that a large computation can be split up and implemented as a collection of parts that are as nearly independent of one another as the overall task allows, is so important that I was moved to elevate it to a principle, the *principle of modular design.* This principle is important because if a process is not designed in this way, a small change in one place has consequences in many other places. As a result, the process as a whole is extremely difficult to debug or to improve, whether by a human designer or in the course of natural evolution, because a small change to improve one part has to be accompanied by many simultaneous, compensatory changes elsewhere. The principle of modular design does not forbid weak interactions between different modules in a task, but it does insist that the overall organization must, to a first approximation, be modular.

From a theoretical point of view, observations like Bela Julesz's are extremely valuable because they enable us to formulate clear computational questions that we know must have answers because the human visual system can carry out the task in question. It was Julesz's findings that allowed us to formulate our theory of human stereopsis (Marr and Poggio, 1979). The analogous findings of Miles (1931) and of Wallach and O'Connell (1953) allowed Ullman (1979b) to develop his theory of structure from motion. Some other experiments by Julesz (1971, chap. 4), together with Braddick's (1974) identification of a short-range, short-term process in apparent motion, contributed to the formulation of our theory of directional selectivity.

The existence of a modular organization in the human visual processor proves that different types of information can be analyzed in relative isolation. As H. K. Nishihara (1978) put it, information about the geometry and reflectance of visible surfaces is encoded in the image in various ways and can be decoded by processes that are almost independent. When this point was fully appreciated, it led to an explosion of theories about possible decoding processes. This chapter describes the computational theories of those decoding processes that are now quite well understood. These processes are (1) stereopsis, (2) directional selectivity, (3) structure from apparent motion, (4) depth from optical flow, (5) surface orientation from surface contours, (6) surface orientation from surface texture, (7) shape from shading, (8) photometric stereo (the determination of surface orientation and reflectance from scene radiances—the intensity of reflected light—observed by a fixed sensor under varying lighting conditions), and (9) lightness and color as an approximation to reflectance. Of course, other cues are available, like occlusion, but unless I have been able to give a process a reasonably integrated treatment, I have not discussed it here. Not all of the methods described here have biological relevance—photometric stereo certainly has none—but they are all of interest as ways of inferring the geometry and reflectance of visible surfaces from their images.

3.2 PROCESSES, CONSTRAINTS, AND THE AVAILABLE REPRESENTATIONS OF AN IMAGE

Before embarking on a detailed description of the different theories, I should make some remarks about the general nature of these theories and what the reader should look for in them and expect from them.

The first point is to remind the reader that we expect to analyze processes at three levels (remember Figure 1–4)—the levels of computational theory, of algorithm, and of implementation. Of course, the vision problem has not been completely solved yet, so we cannot analyze at all three levels every process within the human visual system. But we can analyze some processes at all three levels, and many of them at one or two—perhaps even most of the processes that discern surfaces from images.

In every case, we start with the first level—the computational theory—because this book is about the computational approach to vision. And at this level the reader should look out for the physical constraints that allow the process to do what it does. The situation is quite like what happened in Chapter 2. There we were dealing with ways of representing the image, and in order to say what would be useful and what would not, we were continually referring to the interaction between the imaging process and the underlying properties of the physical world that gives rise to structure in images. In this chapter, where we deal with processes instead of representations, the situation is entirely analogous but arises in a slightly different way. We have already met an example of this new situation in the theory of how to combine zero-crossings from different-sized filters in order to make the physically meaningful primitives of the raw primal sketch. The critical point was that, in general, there is no reason why the zero-crossings from two channels that do not overlap in the frequency domain should be related. They are related in early vision because intensity changes are caused by markings on a surface, the edges of objects, and so on, and these happen to have the critical property of spatial localization.

This interaction between the imaging process and the underlying properties of the physical world commonly occurs in the study of visual processes, and we shall meet several examples here. Frequently an apparently insoluble problem arises, such as which dots in the left-hand pattern in Figure 1–1 should match which dots in the right-hand pattern. From the image alone one just cannot tell. The critical step in formulating the computational theory of stereopsis is the discovery of additional constraints on the process that are imposed naturally and that limit the result sufficiently to allow a unique solution. Finding such constraints is a true discovery—the knowledge is of permanent value, it can be accumulated and built upon, and it is in a deep sense what makes this field of investigation into a science (Marr, 1977b).

Once we have isolated where the extra information comes from—in what ways, if you like, the information is constrained by the world—we can incorporate it into the design of a process. For combining zero-crossings, for example, this was done by the spatial coincidence *assumption*—that coincident zero-crossings are adequate evidence of a physical edge. Thus, the constraints are used by turning them into an assumption that may or may not be internally verifiable.

This, then, is one aspect of the top-level computational theory of a process, but there is another, almost as important. We saw in Chapter 1 that a process can be viewed as a transformation from one representation to another. Addition, for example, maps a pair of numbers into a number. All the processes that we shall discuss take as their inputs properties of the image and produce as their outputs properties of the surfaces—indicating to us either something about the geometry or the reflectance of the surfaces.

We shall discuss ways of representing the outputs of these processes in the next chapter, but now we are concerned with their inputs. What should serve as the inputs to these processes? We already have four options—the image itself, zero-crossings, the raw primal sketch, and the full primal sketch. Part of the computational theory must indicate which of

these four should be used (or if something else entirely is appropriate) and why, and a portion of the investigation of each process will deal with this question.

Ultimately, of course, psychophysics tells us which input representation is used—if the process is in fact incorporated in the human visual system. There is, however, one useful point to bear in mind (Marr, 1974b): Essentially, since the constraints allow the processes to work, and since the constraints are imposed by the real world, by and large the primitives that the processes operate on should correspond to physical items that have identifiable physical properties and occupy a definite location on a surface in the world. Thus one should not try to carry out stereo matching between gray-level intensity arrays, precisely because a pixel corresponds only implicitly and not explicitly to a location on a visible surface.

This point is important. For example, failure to recognize it held Wallach and O'Connell (1953) up for years by their own admission. They could not understand why the shadow of a bent wire should be different from the shadow of a smooth solid object. If a wire is rotated, its shadow moves, and one instantly perceives the wire's three-dimensional shape; if a solid object is rotated, its shadow moves but one cannot perceive its shape. The reason is that the shadow of the wire produces an outline that is effectively in one-to-one correspondence with fixed points on the wire, each having a definite physical location that changes from frame to frame, admittedly, but that always corresponds to the same piece of wire. For the rotating object this is just not true. From moment to moment, the points on the silhouette correspond to quite different points on the object's surface. The image primitives are no longer effectively tied to a constant physical entity. Hence the shape recovery process fails.

On the other hand, the more complex the derivation of a representation from an image, the longer the derivation is liable to take. In real life, time is often of the essence; especially in the analysis of motion, an answer is required as soon as possible—before the image has become out-of-date or before the mover has eaten the viewer. In general, therefore, evolution is prejudiced toward getting things started as soon as possible.

Hence, although processes that operate on the information in an image could use any of a wide variety of input representations in principle, in practice they are likely to use the earliest representations that they possibly can. The range that we have discussed includes the gray-level image, zero-crossings, the raw primal sketch, and the full primal sketch. The earlier ones are not yet "physical," and so a bit unsafe, which might cause us to make mistakes. But for some purposes this possible error is worth the extra speed, for example, in the control of eye movements in response to a sudden change in an image and perhaps also for looming detectors in the theory of directional selectivity (see Section 3.4). Furthermore, just because a boundary is physical does not always make it safe to use. The edges of a uniform cylindrical lamppost give rise to perfectly good edges in the images seen by the left and right eyes, but these edges correspond to different lines on the physical surface. This gives the stereopsis process trouble when, having matched the images, it tries to calculate how far away the lamppost is.

So our rule, then, that the inputs to a process should consist of elements with close physical correlates, is only a general one. It is clearly inappropriate for some things, like shape from shading or photometric stereo, but probably rather important for things like the correspondence process in apparent motion (Ullman, 1978) or the analysis of shape from surface contours or texture. The rule has its attendant dangers, though, and for some processes it is obeyed only marginally—for example, I think that both stereopsis and directional selectivity can use zero-crossings directly. However, the important point is that the rule is sufficiently strong and apparently valid and that violations cannot be allowed to go unnoticed. They have to be defended.

So much, then, for the level of computational theory. The second of the three levels of understanding a process is the level of the algorithm. At this level we formulate a particular procedure for implementing a computational theory. There are two principles that guide the design of algorithms, and they probably ought to be satisfied by any serious candidate for an early visual process in the human visual system. One principle says, roughly, that the algorithm has to be robust: the other, that it must behave smoothly. They are as follows (Marr, 1976):

1. *Principle of graceful degradation.* This principle is designed to ensure that, wherever possible, degrading the data will not prevent the delivery of at least some of the answer. It amounts to a condition on the continuity of the relation between different stages in the processing. For example, it should be required that a rough two-dimensional description of the kind that a vision system might compute out of a drawing enable the system to compute a rough three-dimensional description of what the drawing represents.

2. *Principle of least commitment.* This principle requires not doing something that may later have to be undone, and I believe that it applies to all situations in which performance is fluent. It states that algorithms that are constructed according to a hypothesize-and-test strategy should be avoided because there is probably a better method. My experience has been that if the principle of least commitment has to be disobeyed, one is either doing something wrong or something very difficult.

It would be nice to be able to give general rules about processes at the third level of analysis, the level of neural implementation. Unfortunately, only a few process theories have been developed to the point where specific neural implementations have been proposed, and none of these implementations have been confirmed experimentally in every detail so we are not yet in a position to formulate such rules.

However, one suggestion of a rule can be extracted from our experience with cooperative algorithms for stereopsis and locally parallel organization (Marr and Poggio, 1976; Stevens, 1978). It is only a suggestion, however, and I give it with that caution. It is that, if possible, the nervous system avoids iterative methods—that is, pure iteration in which no new information is introduced at each cycle. Instead, it seems to prefer one-shot methods, like Stevens' (1978) one-shot algorithm for finding the local orientation in Glass patterns. The nervous system also seems to prefer methods that run from the coarse to the fine, doing essentially the same thing at each state but being saved from pure iteration by introducing new information at each cycle. Our stereo algorithm has this form, as we shall see in the next section. And it might be a sound design principle, too, since it effortlessly incorporates the principles of graceful degradation and least commitment.

Yet cooperative methods (a type of nonlinear, iterative algorithm) look very plausible from some points of view. They are very robust, for example, and often have a structure that is readily translatable into the inhibitory and excitatory connections of a plausible neural network. Why, then, are they not used?

One possible explanation may be that cooperative methods take too long and demand too much of the neural hardware to be implemented in any direct way. The problem with iteration is that it demands the circulation of numbers around some kind of loop, which could be carried out by some system of recurrent collaterals or closed loops of neuronal connections. However, unless the numbers involved can be represented quite accurately as they are circulated, errors characteristically tend to build up rather quickly. To use a neuron to represent a quantity with an accuracy of even as low as 1 in 10, it is necessary to use a time interval that is sufficiently long to hold between 1 and 10 spikes in comfort. This means at least 50 ms per iteration for a medium-sized cell, which means 200 ms for four iterations—the minimum time ever required for our cooperative algorithm to solve a stereogram. And this is too slow.

This argument against purely iterative algorithms is not compelling. It is, however, persuasive enough to make me skeptical of them as candidates for processes used by the human visual processor, and it suggests that one should try very hard when designing ways of implementing a process to use algorithms with a more open and flexible structure.

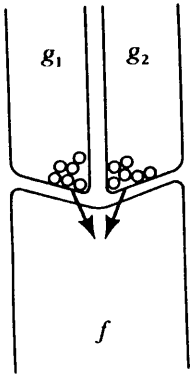

Figure 3–2. The synaptic arrangement considered by Torre and Poggio (1978). Such an arrangement could approximate an AND–NOT gate.

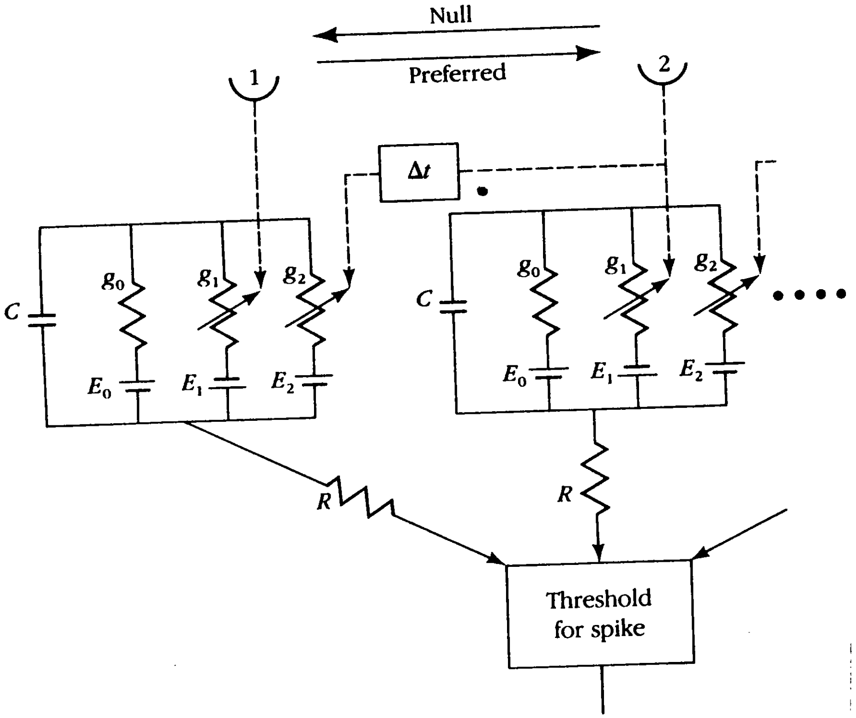

Figure 3–3. The electrical circuit equivalent of the synaptic arrangement shown in Figure 3–2 in the configuration suggested by Torre and Poggio (1978) for implementing directional selectivity. The interaction implemented by the circuit has the form $g_1 - \alpha g_1 g_2$, which approximates a logical AND–NOT gate. A logical AND gate can be implemented by a similar circuit.

One other lesson about neural implementations may perhaps be drawn, this time from the work of Torre and Poggio (1978), who showed how the nonlinear operation AND–NOT could be implemented at the level of synaptic interactions on a dendrite. They showed, using a cable-theoretical analysis, which calculates the time dependent electrical properties of the dendrite from its geometry, that the synaptic arrangement shown in Figure 3–2 has the electrical properties of the circuit shown in Figure 3–3 and the behavior shown in Figure 3–4. It approximately com-

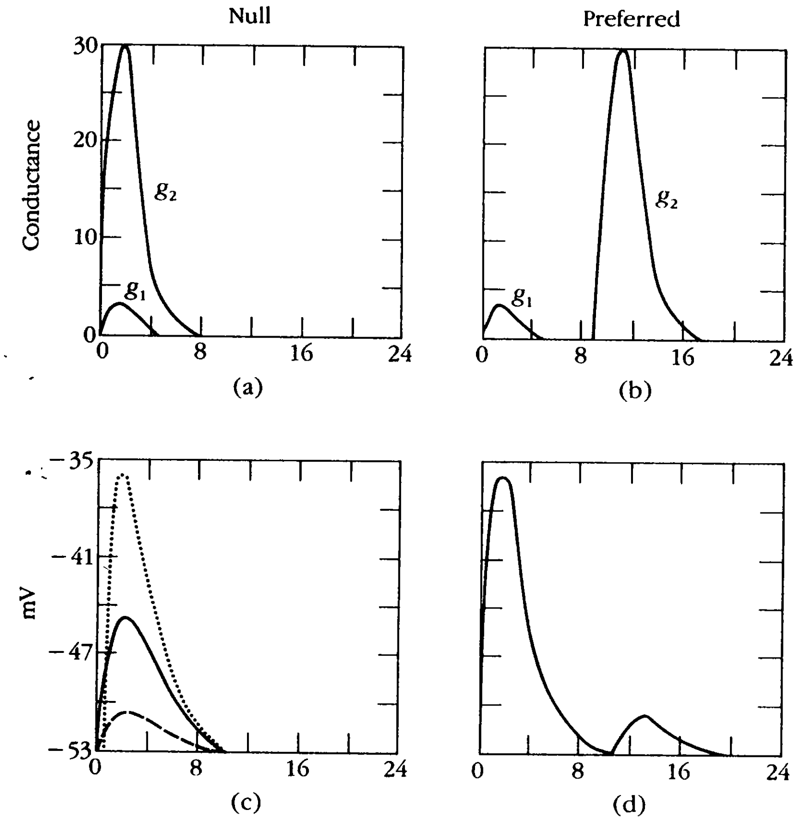

Figure 3–4. The calculated behavior of the circuit in Figure 3–3. For movement in the null direction, the time course of the inputs g_1 and g_2 is shown in (a), and the output of the circuit is the solid line in (c). The dotted and dashed curves show, respectively, the responses with g_1 and g_2 separately. For motion in the opposite direction, the inputs arrive as shown in (b), and the output of the circuit is shown in (d). Notice how attenuated (c) is relative to (d). In this manner, the output of the system can be made directionally selective. The time courses (horizontal axes) are plotted in units of the membrane time constant.

putes $g_1 - \alpha g_1 g_2$, which behaves like AND–NOT, and they suggested that this might be how the ideas of Hassenstein and Reichardt (1956) and of Barlow and Levick (1965) about directional selectivity in the fly and rabbit retinas are implemented (see Section 3.4). Poggio and Torre (1978) extended this idea, showing that a wide range of primitive, nonlinear operations could be implemented using local synaptic mechanisms.

One message of this work is that neurons might do more than we think. Early models, like those of McCulloch and Pitts (1943), tended to see neurons as basically linear devices that could implement nonlinear functions by means of a threshold, which could perhaps be variable if produced by an inhibitory interneuron. This way of thinking led Barlow and Levick to formulate their model of directional selectivity, and I employed it myself when I was interested in the cerebellar cortex (Marr, 1969). We have already seen, however, that local nonlinearities may be important. For example, the scheme for zero-crossing detection in Figure 2–18 is based on the use of many AND gates. The force of Poggio and Torre's work is that such things as AND gates may not require whole cells for their implementation—they can perhaps be executed much more compactly by local synaptic interactions in small pieces of dendrite.

Enough, then, of generalities; let us turn to the processes themselves. I shall start with stereopsis, since it was the first psychological process to be understood and because it led to much of the general knowledge about early vision already incorporated into my account. I have tried not to be too technical in describing the various processes, my aim being to give the reader a general feel for how they all work and to show some examples of them working. For full details, the reader may consult the original articles.

One final point about the organization of the account. Many of these processes divide naturally into two parts, the first concerned with setting up and making a measurement, so to speak, and the second with using the measurement to recover three-dimensional structure. In stereopsis, for example, the first step is the matching process, which establishes the correspondence between the two eyes so that disparities can be measured; the second is the trigonometry that recovers distance and surface orientation from disparity. The first step is the difficult one; the second is easy. In directional selectivity, the first step is to establish the local direction of movement, and the second is to use this sparse local information to help separate figure from ground. Neither step is particularly difficult. In apparent motion, the first step is to establish a correspondence between successive "frames" so that the displacements between frames can be measured; the second step is to use these measurements to recover three-dimensional structure. Here both steps are difficult.

For this reason I have split several of the sections into two parts. Of course, whether a process is indeed implemented by the human visual processor is sometimes unknown, and, even if it were known, whether it is divided as I have described is still an open psychophysical question. In such cases, I have tried to make clear what the current evidence is and what needs to be done to resolve the open questions.

3.3 STEREOPSIS

We saw earlier that the two eyes form slightly different images of the world. The relative difference in the positions of objects in the two images is called disparity, which is caused by the differences in their distance from the viewer. Our brains are capable of measuring this disparity and of using it to estimate the relative distances of the objects from the viewer. I shall use the term *disparity* to mean the angular discrepancy in position of the image of an object in the two eyes; the term *distance* will refer to the objective physical distance from the viewer to the object, usually measured from one of the two eyes; and the term *depth* I shall reserve for the subjective distance to the object as perceived by the viewer.

I shall divide the account into two parts, the first concerned with measuring disparity, and the second with using it. Both parts are separated into the three levels of Figure 1–4. The articles on which this account is based are by Marr (1974b) and Marr and Poggio (1976), which deal with the computational theory; by Marr and Poggio (1979), which deals with the algorithm thought to be used by the human visual system; and by Grimson and Marr (1979) and Grimson (1981), which describe Eric Grimson's computer implementation of the algorithm. Between 1977 and 1979, the additional work done on zero-crossings (Marr, Poggio, and Ullman, 1979; Marr and Hildreth, 1980) allowed certain simplifications in the implementation of the algorithm; most notably, we found from mathematical arguments that we could use circularly symmetric instead of oriented receptive fields for the initial convolutions. This particular detail was arrived at independently on psychophysical grounds by Mayhew and Frisby (1978a).

Measuring Stereo Disparity

Computational theory

Three steps are involved in measuring stereo disparity: (1) A particular location on a surface in the scene must be selected from one image; (2)