General
Introduction

What does it mean, to see? The plain man’s answer (and Aristotle’s, too) would be, to know what is where by looking. In other words, vision is the process of discovering from images what is present in the world, and where it is.

Vision is therefore, first and foremost, an information-processing task; but we cannot think of it just as a process. For if we are capable of knowing what is where in the world, our brains must somehow be capable of representing this information—in all its profusion of color and form, beauty, motion, and detail. The study of vision must therefore include not only the study of how to extract from images the various aspects of the world that are useful to us, but also an inquiry into the nature of the internal representations by which we capture this information and thus make it available as a basis for decisions about our thoughts and actions. This duality—the representation and the processing of information—lies at the heart of most information-processing tasks and will profoundly shape our investigation of the particular problems posed by vision.

The need to understand information-processing tasks and machines has arisen only quite recently. Until people began to dream of and then to build such machines, there was no very pressing need to think deeply
about them. Once people did begin to speculate about such tasks and machines, however, it soon became clear that many aspects of the world around us could benefit from an information-processing point of view. Most of the phenomena that are central to us as human beings—the mysteries of life and evolution, of perception and feeling and thought—are fundamentally phenomena of information processing, and if we are ever to understand them fully, our thinking about them must include this perspective.

The next point—which has to be made rather quickly to those who inhabit a world in which the local utility’s billing computer is still capable of sending a final demand for $1.00—is to emphasize that saying that a job is "only" an information-processing task or that an organism is "only" an information-processing machine is not a limiting or a peremptive description. Even more importantly, I shall in no way use such a description to try to limit the kind of explanations that are necessary. Quite the contrary, in fact. One of the most fascinating features of information-processing machines is that in order to understand them completely, one has to be satisfied with one’s explanations at many different levels.

For example, let us look at the range of perspectives that must be satisfied before one can be said, from a human and scientific point of view, to have understood visual perception. First, and I think foremost, there is the perspective of the plain man. He knows what it is like to see, and unless the bones of one’s arguments and theories roughly correspond to what this person knows to be true at first hand, one will probably be wrong (a point made with force and elegance by Austin, 1962). Second, there is the perspective of the brain scientist, the physiologists, and neurobiologists who know a great deal about how the nervous system is built and how parts of it behave. The issues that concern them—how the cells are connected, why they respond as they do, the neural dogmas of Bawler (1972)—must be resolved and addressed in any full account of perception. And the same argument applies to the perspective of the experimental psychologists.

On the other hand, someone who has bought and played with a small home computer may make quite different demands. "It," he might say, "just simply is an information-processing task, then I should be able to make my computer do it, provided that it has sufficient power, memory, and some way of being connected to a home television camera." The explanation he wants is therefore a rather abstract one, telling him what to program and, if possible, a bias about the best algorithms for doing so. He doesn’t want to know about retinopha, or the lateral geniculate nucleus, or inhibitory interneurons. He wants to know how to program vision.

The fundamental point is that in order to understand a device that performs an information-processing task, one needs many different kinds of explanations. Part I of this book is concerned with this point, and it plays a prominent role because one of the keystones of the book is the realization that we have had to be more careful about what constitutes an explanation than has been necessary in other recent scientific developments, like those in molecular biology. For the subject of vision, there is no single equation or view that explains everything. Each problem has to be addressed from several points of view—as a problem in representing information, as a computational device in deriving that representation, and as a problem in the architecture of a computer capable of carrying out both things quickly and reliably.

If one keeps strongly in mind this necessarily rather broad aspect of the nature of explanation, one can avoid a number of pitfalls. One consequence of an emphasis on information processing might be, for example, to introduce a comparison between the human brain and a computer. In a sense, of course, the brain is a computer, but to say this without qualification is misleading, because the essence of the brain is not simply that it is a computer but that it is a computer which is in the habit of performing some rather particular computations. The term computer usually refers to a machine with a rather standard type of instruction set that usually runs serially but nowadays sometimes in parallel, under the control of programs that have been stored in a memory. In order to understand such a computer, one needs to understand what it is made of, how it is put together, what its instruction set is, how much memory it has and how it is accessed, and how the program may be made to run. But this forms only a small part of understanding a computer that is performing an information-processing task.

This point bears reflection, because it is central to why most analogies between brains and computers are too superficial to be useful. Think, for example, of the international network of airline reservation computers, which performs the task of assigning flights for millions of passengers all over the world. To understand this system it is not enough to know how a modern computer works. One also has to understand the other factors—such as the cost of what aircraft are, and what they do, about geography, time zones, fares, exchange rates, and connections, and something about politics, diets, and the various other aspects of human nature that happen to be relevant to this particular task.

Thus the critical point is that understanding computers is different from understanding computations. To understand a computer, one has to study that computer. To understand an information-processing task, one has to study that information-processing task. To understand fully a particular machine carrying out a particular information-processing task, one has to do both things. Neither alone will suffice.
From a philosophical point of view, the approach that I describe is an extension of what have sometimes been called representational theories of mind. On the whole, it rejects the more recent excursions into the philosophy of perception, with their arguments about sense-data, the molecules of perception, and the validity of what the senses tell us; instead, this approach looks back to an older view, according to which the senses are for the most part concerned with telling one what is there. Modern representational theories conceive of the mind as having access to systems of internal representations; mental states are characterized by asserting what the internal representations currently specify, and mental processes by how such internal representations are obtained and how they interact.

This scheme affords a comfortable framework for our study of visual perception, and I am content to let it form the point of departure for our inquiry. As we shall see, pursuing this approach will lead us away from traditional avenues into what is almost a new intellectual landscape. Some of the things we find will seem strange, and it will be hard to reconcile subjectively some of the ideas and theories that are forced on us with what actually goes on inside ourselves when we open our eyes and look at things. Even the basic notion of what constitutes an explanation will have to be developed and broadened a little, to ensure that we do not leave anything out and that every important perspective on the problem is satisfied or satisfied.

The book itself is divided into three parts. In the first are contained the philosophical preliminaries, a description of the approach, the representational framework that is proposed for the overall process of visual perception, and the way that led to it. I have adopted a fairly personal style in the hope that if the reader understands why particular directions were taken at each point, the reasons for the overall approach will be clearer.

The second part of the book, Chapters 2 to 6, contains the real analysis. It describes informal, but in some detail, how the approach and framework are actually realized, and the results that have been achieved.

The third part is somewhat unorthodox and consists of a set of questions and answers that are designed to help the reader to understand the way of thinking behind the approach—to help him acquire the right prejudices, if you like—and to relate these explanations to his personal experience of seeing. I have often found that one or two of the remarks set out in Part III have helped a person to see the point of part of the theory or to circumvent some private difficulty with it, and I hope they may serve a similar purpose here. The reader may find this section more after having read the first two parts of the book, but an early glance at it may provide the motivation to take the trouble.

The detailed exposition comes, then, in Part II. Of course, the subject of human visual perception is not solved here by a long way. But over the last six years, my colleagues and I have been fortunate enough to see the establishment of an overall theoretical framework as well as the solution of several rather central problems in visual perception. We feel that the combination amounts to a reasonably strong case that the representational approach is a useful one, and the point of this book is to make that case. How far this approach can be pursued, of course, remains to be seen.
The Philosophy and the Approach

1.1 BACKGROUND

The problems of visual perception have attracted the curiosity of scientists for many centuries. Important early contributions were made by Newton (1784), who laid the foundations for modern work on color vision, and Helmholtz (1910), whose treatise on physiological optics generated interest even today. Early in this century, Wertheimer (1912, 1923) noticed the apparent motion not of individual dots but of wholes, or "fields," in images presented sequentially as in a movie. In much the same way we perceive the migration across the sky of a flock of geese, the flock somehow constitutes a single entity; and is not seen as individual birds. This observation started the Gestalt school of psychology, which was concerned with describing the qualities of wholes by using terms like solidarity and opposition, and, with trying to formulate the "laws" that governed the creation of these wholes. The attempt failed for various reasons, and the Gestalt school disintegrated into the fog of subjectivism. With the death of the school, many

The two dimensional image seen by a single eye.
The Philosophy and the Approach

right eye is made by copying the left image, shifting a square-shaped region at its center slightly to the left, and then providing a new random pattern to fill the gap that the shift creates. If each of the eyes sees only one matrix, as if the matrices were both in the same physical place, the result is the sensation of a square floating in space. Plainly such percepts are caused solely by the stereo disparity between matching elements in the images presented to each eye; from such experiments, we know that the analysis of stereoscopic information, like the analysis of motion, can proceed independently in the absence of other information. Such findings are of critical importance because they help us to subdivide our study of perception into more specialized parts which can be treated separately. I shall refer to these as independent modules of perception.

The most recent contribution of psychophysics has been of a different kind but of equal importance. It arose from a combination of adaptation and threshold detection studies and originated from the demonstration by Campbell and Robson (1968) of the existence of independent, spatial-frequency-tuned channels—that is, channels sensitive to intensity variations in the image occurring at a particular scale or spatial interval—in the early stages of our perceptual apparatus. This paper led to an explosion of articles on various aspects of these channels, which culminated ten years later with quite satisfactory quantitative accounts of the characteristics of the first stages of visual perception (Wilson and Bergen, 1979). I shall discuss this in detail later on.

Recently a rather different approach has attracted considerable attention. In 1971, Roger N. Shepard and Jacqueline Metzler made line drawings of simple objects that differed from one another either by a three-dimensional rotation or by a rotation plus a reflection (see Figure 1–2). They asked how long it took to decide whether two depicted objects differed by a rotation and a reflection or merely a rotation. They found that the time taken depended on the three-dimensional angle of rotation necessary to bring the two objects into correspondence. Indeed, the time varied linearly with this angle. One is led thereby to the notion that a mental rotation of sorts is actually being performed—that a mental description of the first shape in a pair is being adjusted incrementally in orientation until it matches the second, such adjustment requiring greater time when greater angles are involved.

The significance of this approach lies not so much in its results, whose interpretation is controversial, as in the type of questions it raised. For until then, the notion of a representation was not one that visual psychologists took seriously. This type of experiment means that the notion had to be considered. Although the early thoughts of visual psychologists were naive compared with those of the computer vision community, which had had to face the problem of representation from the beginning, it was not long before the thinking of psychologists became more sophisticated (see Shepard, 1979).

But what of exploration? For a long time, the best hope seemed to lie along another line of investigation, that of electrophysiology. The developers of amplifiers allowed Adrian (1928) and his colleagues to record the minute voltage changes that accompanied the transmission of nerve signals. Their investigations showed that the character of the sensation so produced depended on which fiber carried the message, not how the fiber
was stimulated—as one might have expected from anatomical studies. This led to the view that the peripheral nerve fibers could be thought of as a simple mapping supplying the sensorium with a copy of the physical events at the body surface (Adrian, 1947). The rest of the explanation, it was thought, could safely be left to the psychologists.

The next development was the technical improvement in amplification that made possible the recording of single neurones (Grant and Szentágothai, 1939; Hartline, 1938; Galambos and Davis, 1946). This led to the notion of a cells' "receptive field" (Hartline, 1940) and the Harvard School's famous study of the behavior of neurones at successively deeper levels of the visual pathway (Kuffler, 1953; Hubel and Wiesel, 1962, 1965). But perhaps the most exciting development was the new view that questions of psychological interest could be illuminated and perhaps even explained by neurophysiological experiments. The clearest early example of this was Barlow's (1955) study of ganglion cells in the frog retina, and I cannot put it better than he did:

If one explores the responsiveness of single ganglion cells in the frog's retina using bright field targets, one finds that one particular type of ganglion cell is most effectively driven by something like a black disc subtending a degree or so more rapidly than the animal's receptive field. This causes a vigorous discharge which is maintained without much decrement as long as the movement is continued. Now, if the stimulus which is optimal for this class of cells is presented to intact frogs, the behavioral response is often a jump and snap. The selectivity of the retinal neurones and the selectivity of the animal's reaction in this case is striking. But frogs react more slowly to the same stimulus, and Hubel & Wiesel (1962) found these neurones to be "broadly tuned" to the different intensities of the visual image onto the sensorium, but instead they detect pattern elements, discriminate the depth of objects, and ignore irrelevant causes of variation and are thus arranged in an intriguing hierarchy. Furthermore, there is evidence that they give prominence to what is informationally important, can respond with great reliability, and can have a pattern selectivity permanently modified by early visual experience. This research brought about by neurons and we should not use phrases like "visual activity reflects, reveals, or monitors thought processes" because the activities of neurons provide a basis for understanding how the brain controls behavior" (Barlow, 1972, p. 380).

This aspect of his thinking led Barlow to formulate the first and most important of his five dogmas: A description of the activity of a single nerve cell which is transmitted to and influences other nerve cells and of a nerve cells response to such influences from other cells, is a complete enough description for functional understanding of the nervous system. There is nothing else "looking at" or controlling this activity which must therefore provide a basis for understanding how the brain controls behavior." (Barlow, 1972, p. 380).

I shall return later on to more carefully examine the validity of this point of view, but for now let us just enjoy it. The vigor and excitement of these ideas need no emphasis. At the time the eventual success of a reductionist approach seemed likely, Hubel and Wiesel's (1962, 1968) pioneering studies had shown the way: single-unit studies on stereocats (Barlow, Blake, and Pettigrew, 1967) and on color (DeValois, Abrahams, and Mead, 1967; Gouras, 1968) seemed to confirm the close links between perception and single-cell recordings, and the intriguing results of Groe,
Rocha-Miranda, and Bender (1972), who found "hand-detectors" in the inferotemporal cortex, seemed to show that the application of the reductionist approach would not be limited just to the early parts of the visual pathway.

It was, of course, recognized that physiologies had been lucky. If one probes around in a conventional electronic computer and records the behavior of single elements within it, one is unlikely to be able to discern what a given element is doing. But the brain, thanks to Barlow's first dogs, seemed to be built along more accommodating lines—people were able to determine the functions of single elements of the brain. There seemed no reason why the reductionist approach could not be taken all the way.

I was myself fully caught up in this excitement. Truth, I also believed, was basically neural, and the central aim of all research was a thorough functional analysis of the structure of the central nervous system. My enthusiasm found expression in a theory of the cerebellar cortex (Marr, 1969).

According to this theory, the simple and regular cortical structure is interpreted as a simple but powerful memorizing device for learning motor skills, because of a simple combinatorial trick: each of the 15 million Purkinje cells in the cerebellum is capable of learning over 200 different patterns and discriminating them from unlearned patterns. Evidence is gradually accumulating that the cerebellum is involved in learning motor skills (Ito, 1978), so that something like this theory may in fact be correct.

The way seemed clear. On the one hand we had new experimental techniques of proven power, and on the other, the beginnings of a theoretical approach that could back them up with a fine analysis of cortical structure. Psychophysics could tell us what needed explaining, and the recent advances in anatomy—the Pia-Heimer technique from Naun's laboratory and the recent successful deployment by Szentagothai and others of the electron microscope—could provide the necessary information about the structure of the cerebral cortex.

But somewhere underneath, something was going wrong. The initial discoveries of the 1950s and 1960s were not being followed by equally dramatic discoveries in the 1970s. No neurophysiologist had recorded new and clear high-level correlates of perception. The leaders of the 1960s had turned away from what they had been doing—Hubel and Wiesel concentrated on anatomy; Barlow turned to psychophysics; and the main stream of neurophysiology concentrated on development and plasticity—the concept that neural connections are not fixed or on a more thorough analysis of the cells that had already been discovered (for example, Bishop, Goobies, and Henry, 1971; Schiller, Finlay, and Volman, 1976a, 1976b), or on cells in species like the owl (for example, Pettigrew and Komisar, 1970).

None of the new studies succeeded in elucidating the function of the visual cortex.

It is difficult to say precisely why this happened, because the reasoning was never made explicit and was probably largely unconscious. However, various factors are identifiable. In my own case, the cerebellar study had two effects. On the one hand, it suggested that one could eventually hope to understand cortical structure in functional terms, and this was exciting. But at the same time the study had disappointed me, because even if the theory was correct, it did not much enlighten one about the motor system—it did not; for example, tell one how to go about programming a mechanical arm. It suggested that if one wishes to program a mechanical arm so that it operates in a versatile way then at some point a very large and rather simple type of memory will prove indispensable. But it did not say why, nor what that memory should contain.

The discoveries of the visual neurophysiologists left one in a similar situation. Suppose, for example, that one actually found the apocryphal grandmother cell. Would that really tell us anything much at all? It would tell us that it existed—Grand's hand-detectors tell us almost that—but not why or even how such a thing may be constructed from the outputs of previously discovered cells. Do the single-unit recordings—the simple and complex cells—tell us much about how to detect edges or why one would want to, except in a rather general way through arguments based on economy and redundancy? If we really knew the answers, for example, we should be able to program them on a computer. But finding a hand-detector certainly did not allow us to program one.

As one reflected on these sorts of issues in the early 1970s, it gradually became clearer that something important was missing that was not present in either of the disciplines of neurophysiology or psychophysics. The key observation is that neurophysiologists and psychophysicists have as their business to describe the behavior of cells or of subjects but not to explain such behavior. What are the visual areas of the cerebral cortex actually doing? What are the problems in doing it that need explaining, and at what level of description should such explanations be sought?

The best way of finding out the difficulties of doing something is to try to do it, so at this point I moved to the Artificial Intelligence Laboratory at MIT, where Marvin Minsky had collected a group of people and a powerful computer for the express purpose of addressing these questions.

* A cell that fires only when one's grandmother comes into view.
The second approach was to try for depth of analysis by restricting the scope to a world of single, illuminated, matte white toy blocks set against a black background. The blocks could occur in any shape provided only that all faces were planar and all edges were straight. This restriction allowed more specialized techniques to be used, but it still did not make the problem easy. The limbic–Horn line finder (Horn, 1973) was used to find edges, and both it and its sequel (described in Shafai, 1975) made use of the special circumstances of the environment, such as the fact that all edges there were straight.

These techniques did work reasonably well, however, and they allowed a preliminary analysis of later problems to emerge—roughly what does one do once a complete line drawing has been extracted from a scene? Studies of this had begun sometime before with Roberts (1965) and Guzman (1969), and they culminated in the works of Wald (1975) and Mackworth (1975), which essentially solved the interpretation problem for line drawings derived from images of prismatic solids. Wald's work had a particularly dramatic impact, because it was the first to show explicitly that an exhaustive analysis of all possible local physical arrangements of surfaces, edges, and shadows could lead to an effective and efficient algorithm for interpreting an actual image. Figure 1–3 and its legend convey the main ideas behind Wald's theory.

The hope that lay behind this work was, of course, that once the toy world of white blocks had been understood, the solutions found there could be generalized, providing the basis for attacking the more complex problems posed by a richer visual environment. Unfortunately, this turned out not to be so. For the roots of the approach that was eventually successful, we have to look at the third kind of development that was taking place then.

Two pieces of work were important here. Neither is probably of very great significance to human perception for what it actually accomplished—in the end, it is likely that neither will particularly reflect human visual processes—but they are both of importance because of the way in which they were formulated. The first was Land and McCann's (1971) work on the retinex theory of color vision, as developed by them and subsequently by Horn (1974). The starting point is the traditional one of regarding color as a perceptual approximation to reflectance. This allows the formulation of a clear computational question, namely: How can the effects of reflectance changes be separated from the vagaries of the prevailing illumination? Land and McCann suggested using the fact that changes in illumination are usually gradual, whereas changes in reflectance of a surface or of an object boundary are often quite sharp. Hence by filtering out slow changes, these changes due to the reflectance alone could be isolated. Horn devised a

*Operator refers to a local calculation to be applied at each location in the image making use of the intensity there and in the immediate vicinity.
The other piece of work was Horn's (1975) analysis of shape from shading, which was the first in what was to become a distinguished series of articles on the formation of images. By carefully analyzing the way in which the illumination, surface geometry, surface reflectance, and viewpoint conspired to create the measured intensity values in an image, Horn formulated a differential equation that related the image intensity values to the surface geometry. If the surface reflectance and illumination are known, one can solve for the surface geometry (see also Horn, 1977). Thus from shading one can derive shape.

The message was plain. There must exist an additional level of understanding at which the character of the information processing tasks carried out during perception are analyzed and understood in a way that is independent of the particular mechanisms and structures that implement them in our heads. This was what was missing—the analysis of the problem as an information-processing task. Such analysis does not usurp an understanding at the other levels—of neurons or of computer programs—but it is a necessary complement to them, since without it there can be no real understanding of the function of all those neurons.

This realization was arrived at independently and formulated together by Tomaso Poggio in Tübingen and myself (Marr and Poggio, 1977; Marr, 1977b). It was not even quite a new—Leon D. Harmon was saying something similar at about the same time, and others had paid lip service to a similar distinction. But the important point is that if the notion of different types of understanding is taken very seriously, it allows the study of the information-processing basis of perception to be made rigorous. It becomes possible, by separating explanations into different levels, to make explicit statements about what is being computed and why and to construct theories stating that what is being computed is optimal in some sense or is guaranteed to function correctly. The ad hoc element is removed, and heuristic computer programs are replaced by solid foundations on which a real subject can be built. This realization—the formulation of what was missing, together with a clear idea of how to supply it—formed the basic foundation for a new integrated approach, which it is the purpose of this book to describe.

1.2 UNDERSTANDING COMPLEX INFORMATION-PROCESSING SYSTEMS

Almost never can a complex system of any kind be understood as a simple extrapolation from the properties of its elementary components. Consider, for example, some gas in a bottle. A description of thermodynamic effects—
temperature, pressure, density, and the relationships among these fac-
tors—is not formulated by using a large set of equations, one for each of
the particles involved. Such effects are described at their own level, that of
an enormous collection of particles; the effect is to show that in principle
the microscopic and macroscopic descriptions are consistent with one
another. If one hopes to achieve a full understanding of a system as com-
plicated as a nervous system, a developing embryo, a set of metabolic
pathways, a bottle of gas, or even a large computer program, then one must
be prepared to contemplate different kinds of explanation at different lev-
els of description that are linked, at least in principle, into a coherent whole,
even if linking the levels in complete detail is impractical. For the specific
case of a system that solves an information-processing problem, there are
in addition the twin strands of process and representation, and both these
ideas need some discussion.

Representation and Description

A representation is a formal system for making explicit certain entities or
types of information, together with a specification of how the system does
this. And I shall call the result of using a representation to describe a given
entity a description of the entity in that representation (Marc and Nishihara,
1978).

For example, the Arabic, Roman, and binary numeral systems are all
formal systems for representing numbers. The Arabic representation con-
sists of a string of symbols drawn from the set (0, 1, 2, 3, 4, 5, 6, 7, 8, 9),
and the rule for constructing the description of a particular integer \( n \) is that
one decomposes \( n \) into a sum of multiples of powers of 10 and unites these
multiples into a string with the largest powers on the left and the
smallest on the right. Thus, thirty-seven equals \( 3 \times 10^2 + 7 \times 10^0 \), which
becomes 37, the Arabic numeral system's description of the number. What
this description makes explicit is the number's decomposition into powers
of 10. The binary numeral system's description of the number thirty-seven
is 100101, and this description makes explicit the number's decomposition
into powers of 2. In the Roman numeral system, thirty-seven is represented
as XXXVII.

This definition of a representation is quite general. For example, a
representation for shape would be a formal scheme for describing some
aspects of shape; together with rules that specify how the scheme is applied
to any particular shape. A musical score provides a way of representing a
symphony; the alphabet allows the construction of a written representation
of words; and so forth. The phrase "formal scheme" is critical to the defi-
nition, but the reader should not be frightened by it. The reason is simply
that we are dealing with information-processing machines, and the way
such machines work is by using symbols to stand for things—to represent
things, in our terminology. To say that something is a formal scheme means
only that it is a set of symbols with rules for putting them together—no
more and no less.

A representation, therefore, is not a foreign idea at all—we all use
representations all the time. However, the notion that one can capture
some aspect of reality by making a description of it using a symbol and
that to do so can be useful seems to me a fascinating and powerful idea.
But even the simple examples we have discussed introduce some rather
general and important issues that arise whenever one chooses to use one
particular representation. For example, if one chooses the Arabic numeral
representation, it is easy to discover whether a number is a power of 10
but difficult to discover whether it is a power of 2. If one chooses the binary
representation, the situation is reversed. Thus, there is a trade-off; any
particular representation makes certain information explicit at the expense
of information that is pushed into the background and may be quite hard
to recover.

This issue is important, because how information is represented can
greatly affect how easy it is to do different things with it. This is evident
even from our numbers example. It is easy to add, to subtract, and even to
multiply if the Arabic or binary representations are used, but it is not at all
easy to do these things—especially multiplication—with Roman numerals.
This is a key reason why the Roman culture failed to develop mathematics
in the way the earlier Arabic cultures did.

An analogous problem faces computer engineers today. Electronic
technology is much more suited to a binary number system than to the
conventional base 10 system, yet humans supply their data and require the
results in base 10. The design decision facing the engineer, therefore, is,
Should one pay the cost of conversion into base 2, carry out the arithmetic
in a binary representation, and then convert back into decimal numbers
to output, or should one sacrifice efficiency of circuitry to carry out oper-
ations directly in a decimal representation? On the whole, business com-
puters and pocket calculators take the second approach, and general pur-
pose computers take the first. But even though one is not restricted to
using just one representation system for a given type of information, the
choice of which to use is important and cannot be taken lightly. It deter-
nines what information is made explicit and hence what is pushed further
into the background, and it has a far-reaching effect on the ease and
difficulty with which operations may subsequently be carried out on that information.

**PROCESS**

The term process is very broad. For example, addition is a process, and so is taking a Fourier transform. But so is making a cup of tea, or going shopping. For the purposes of the book, I want to restrict our attention to the meanings associated with machines that are carrying out information-processing tasks. So let us examine in depth the nature behind one simple such device, a cash register at the checkout counter of a supermarket.

There are several levels at which one needs to understand such a device, and it is perhaps most useful to think in terms of three of them.

The most abstract is the level of what the device does and why. What it does is arithmetic, so our first task is to master the theory of addition.

Addition is a mapping, usually denoted by +, from pairs of numbers into single numbers; for example, + maps the pair (3, 4) to 7, and I shall write this in the form (3 + 4) → 7. Addition has a number of abstract properties, however. It is commutative: both (3 + 4) and (4 + 3) are equal to 7, and associative: the sum of 3 + (4 + 5) is the same as the sum of (3 + 4) + 5. Then there is the unique distinguished element, zero, the adding of which has no effect: (4 + 0) → 4. Also, for every number there is a unique "inverse," written (−4) in the case of 4, which when added to the number gives zero: [4 + (−4)] → 0.

Notice that these properties are part of the fundamental theory of addition. They are true no matter how the numbers are written—whether in binary, Arabic, or Roman representation—and no matter how the addition is executed. This part of the first level is something that might be characterized as what is being computed.

The other half of this level of explanation has to do with the question of why the cash register performs addition and not, for instance, multiplication when combining the prices of the purchased items to arrive at a final bill. The reason is that the rules we intuitively feel to be appropriate for combining the individual prices in fact define the mathematical operation of addition. These can be formulated as constraints in the following way:

1. If you buy nothing, it should cost you nothing; and buying nothing and something should cost the same as buying just the something. (The rules for zero.)

2. The order in which goods are presented to the cashier should not affect the total. (Commutativity.)

3. Arranging the goods into two piles and paying for each pile separately should not affect the total amount you pay. (Associativity; the basic operation for combining prices.)

4. If you buy an item and then return it for a refund, your total expenditure should be zero. (Inverses.)

It is a mathematical theorem that these conditions define the operation of addition, which is therefore the appropriate computation to use.

This whole argument is what I call the computational theory of the cash register. Its important features are (1) that it contains separate arguments about what is computed and why and (2) that the resulting operation is defined uniquely by the constraints it has to satisfy. In the theory of visual processes, the underlying task is to reliably derive properties of the world from images of it; the business of isolating constraints that are both powerful enough to allow a process to be defined and generally true of the world is a central theme of our inquiry.

In order that a process shall actually run, however, one has to realize in some way and therefore choose a representation for the entities that the process manipulates. The second level of the analysis of a process, therefore, involves choosing two things: (1) a representation for the input and for the output of the process and (2) an algorithm by which the transformation may actually be accomplished. For addition, of course, the input and output representations can both be the same, because they both consist of numbers. However, this is not true in general. In the case of a Fourier transform, for example, the input representation may be the time domain, and the output, the frequency domain. If the first of our levels specifies what and why, this second level specifies how: For addition, we might choose Arabic numerals for the representations, and for the algorithm we could follow the usual rules about adding the least significant digits first and "carrying" if the sum exceeds 9. Cash registers, whether mechanical or electronic, usually use this type of representation and algorithm.

There are three important points here. First, there is usually a wide choice of representation. Second, the choice of algorithm often depends rather critically on the particular representation that is employed. And third, even for a given fixed representation, there are often several possible algorithms for carrying out the same process. Which one is chosen will usually depend on any particularly desirable or undesirable characteristics that the algorithms may have; for example, one algorithm may be much
more efficient than another, or another may be slightly less efficient but more robust (that is, less sensitive to slight inaccuracies in the data on which it must run). Or again, one algorithm may be parallel, and another, serial. The choice, then, may depend on the type of hardware or machinery in which the algorithm is to be embodied physically.

This brings us to the third level, that of the device in which the process is to be realized physically. The important point here is that, once again, the same algorithm may be implemented in quite different technologies. The child who mechanically adds two numbers from right to left, carrying a digit when necessary, may be using the same algorithm that is implemented by the wires and transistors of the cash register in the neighborhood supermarket, but the physical realization of the algorithm is quite different in these two cases. Another example: Many people have written computer programs to play tic-tac-toe, and there is a more or less standard algorithm that cannot lose. This algorithm has in fact been implemented by W. D. Wirth and R. Silverman in a quite different technology, in a computer made out of Tinker toys, a children's wooden building set. The whole monotonously ungainly engine, which actually works, currently resides in a museum at the University of Missouri in St. Louis.

Some styles of algorithms will suit some physical substrates better than others. For example, in conventional digital computers, the number of connections is comparable to the number of gates, while in a brain, the number of connections is much larger (\( \times 10^7 \)) than the number of nerve cells. The underlying reason is that wires are rather cheap in biological architecture, because they can grow individually and in three dimensions. In conventional technology, wire laying is more or less restricted to two dimensions, which quite severely restricts the scope for using parallel techniques and algorithms; the same operations are often better carried out serially.

The Three Levels

We can summarize our discussion in something like the manner shown in Figure 1-4, which illustrates the different levels at which an information-processing device must be understood before one can be said to have understood it completely. At one extreme, the top level, is the abstract computational theory of the device, in which the performance of the device is characterized as a mapping from one kind of information to another; the abstract properties of this mapping are defined precisely, and its appropriateness and adequacy for the task at hand are demonstrated. In the middle is the choice of representation for the input and output and the algorithm to be used to transform one into the other. And at the other extreme are the details of how the algorithm and representation are realized physically—the detailed computer architecture, so to speak. These three levels are coupled, but only loosely. The choice of an algorithm is influenced for example, by what it has to do and by the hardware in which it must run. But there is a wide choice available at each level, and the explication of each level involves issues that are rather independent of the others.

Each of the three levels of description will have its place in the eventual understanding of perceptual information processing, and of course they are logically and causally related. But an important point to note is that since the three levels are only rather loosely related, some phenomena may be explained at only one or two of them. This means, for example, that a correct explanation of some psychophysical observation must be formulated at the appropriate level. In attempts to relate psychophysical problems to physiology, too often there is confusion about the level at which problems should be addressed. For instance, some are related mainly to the physical mechanisms of vision—such as afterimages (for example, the one you see after staring at a light bulb) or such as the fact that any color can be matched by a suitable mixture of the three primaries (a consequence principally of the fact that we humans have three types of cones). On the other hand, the ambiguity of the Necker cube (Figure 1-5) seems to demand a different kind of explanation. To be sure, part of the explanation of its perceptual reversal must have to do with a bistable neural network (that is, one with two distinct stable states) somewhere inside the
1.2 Understanding Complex Information-Processing Systems

Psychophysics can also help to determine the nature of a representation. The work of Roger Shepard (1975), Eleanor Rosch (1978), and Elizabeth Warren (1975) provides some interesting hints in this direction. More specifically, Stevens (1977) argued from psychophysical experiments that surface orientation is represented by the coordinates of slant and tilt, rather than (for example) the more traditional $(p, q)$ of gradient space (see Chapter 3). He also deduced from the uniformity of the size of errors made by subjects judging surface orientation over a wide range of orientations that the representational quantities used for slant and tilt are pure angles and not, for example, their cosines, sines, or tangents.

More generally, if the idea that different phenomena need to be explained at different levels is kept clearly in mind, it often helps in the assessment of the validity of the different kinds of objections that are raised from time to time. For example, one favorite is that the brain is quite different from a computer because one is parallel and the other serial. The answer to this, of course, is that the distinction between serial and parallel is a distinction at the level of algorithm; it is not fundamental at all—anything programmed in parallel can be rewritten serially (though not necessarily vice versa). The distinction, therefore, provides no grounds for arguing that the brain operates so differently from a computer that a computer could not be programmed to perform the same tasks.

Importance of Computational Theory

Although algorithms and mechanisms are empirically more accessible, it is at the top level, the level of computational theory, which is critically important from an information-processing point of view. The reason for this is that the nature of the computations that underlie perception depends more upon the computational problems that have to be solved than upon the particular hardware in which their solutions are implemented. To phrase the matter another way: an algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism (and the hardware) in which it is embodied.

In a similar vein, trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: it just cannot be done. In order to understand how, we have to understand aerodynamics, only then do the structure of feathers and the different shapes of birds make sense. More to the point, as we shall see, we cannot understand why retinal ganglion cells and lateral geniculate neurons have the receptive fields they do just by studying their anatomy and physiology. We can understand how these cells and neurons behave...
as they do by studying their writing and interactions, but in order to understand why the receptive fields are as they are—why they are circularly symmetrical and why their excitatory and inhibitory regions have characteristic shapes and distributions—we have to know a little of the theory of differential operators, band-pass channels, and the mathematics of the uncertainty principle (see Chapter 2).

Perhaps it is not surprising that the very specialized empirical disciplines of the neurosciences failed to appreciate fully the absence of computational theory, but it is surprising that this level of approach did not play a more forceful role in the early development of artificial intelligence. For far too long, a heuristic program for carrying out some task was held to be a theory of that task, and the distinction between what a program did and how it did it was not taken seriously. As a result, (1) a style of exploration evolved that invoked the use of special mechanisms to solve particular problems, (2) particular data structures, such as the lists of attribute value pairs called property lists in the Lisp programming language, were held to amount to theories of the representation of knowledge, and (3) there was frequently no way to determine whether a program would deal with a particular case other than by running the program.

Failure to recognize this theoretical distinction between what and how also greatly hampered communication between the fields of artificial intelligence and linguistics. Chomsky's (1965) theory of transformational grammar is a true computational theory in the sense defined earlier; it is concerned solely with specifying what the syntactic decomposition of an English sentence should be, and not at all with how that decomposition should be achieved. Chomsky himself was very clear about this—it is roughly his distinction between competence and performance, though his idea of performance did include other factors, like stopping in mid-sentence—but the fact that his theory was defined by transformations, which look like computations, seems to have confused many people. Winograd (1972), for example, felt able to criticize Chomsky's theory on the grounds that it cannot be inverted and so cannot be made to run on a computer; it had heard reflections of the same argument made by Chomsky's colleagues in linguistics as they turn their attention to how grammatical structure might actually be computed from a real English sentence.

The explanation is simply that finding algorithms by which Chomsky's theory may be implemented is a completely different endeavor from formulating the theory itself. In our terms, it is a study at a different level, and both tasks have to be done. This point was appreciated by Marcus (1980), who was concerned precisely with how Chomsky's theory can be realized and with the kinds of constraints on the power of the human grammatical processor that might give rise to the structural constraints in syntax that Chomsky found. It even appears that the emerging "trace" theory of grammar (Chomsky and Lasnik, 1977) may provide a way of synthesizing the two approaches—showing that, for example, some of the rather ad hoc restrictions that form part of the computational theory may be consequences of weaknesses in the computational power that is available for implementing syntactical decoding.

The Approach of J. J. Gibson

In perception, perhaps the nearest anyone came to the level of computational theory was Gibson (1966). However, although some aspects of his thinking were on the right lines, he did not understand properly what information processing was, which led him to seriously underestimate the complexity of the information-processing problems involved in vision and the consequent subtext that is necessary in approaching them.

Gibson's important contribution was to take the debate away from the philosophical considerations of sense-data and the affective qualities of sensation and to note instead that the important thing about the senses is that they are channels for perception of the real world outside or, in the case of vision, of the visible surfaces. He therefore asked the critically important question, How does one obtain constant perceptions in everyday life on the basis of continually changing sensations? This is exactly the right question, showing that Gibson correctly regarded the problem of perception as that of recovering from sensory information "valid" properties of the external world. His problem was that he had a much oversimplified view of how this should be done. His approach led him to consider higher-order variables—stimulus energy, ratios, proportions, and so on—as "invariants" of the movement of an observer and of changes in stimulation intensity.

"These invariants," he wrote, "correspond to permanent properties of the environment. They constitute, therefore, information about the permanent environment." This led him to a view in which the function of the brain was to "detect invariants" despite changes in "sensations" of light, pressure, or loudness of sound. Thus, he says that the "function of the brain, when loosed with its perceptual organs, is not to decode signals, nor to interpret messages, nor to accept images, nor to organize the sensory input or to process the data, in modern terminology. It is to seek and extract information about the environment from the flowing array of ambient energy," and he thought of the nervous system as in some way "reusing" these invariants. He then embarked on a broad study of animals in their environments, looking for invariants to which they might...
resonate. This was the basic idea behind the notion of ecological optics (Gibson, 1966, 1979).

Although one can criticize certain shortcomings in the quality of Gibson's analysis, its major and, in my view, fatal shortcoming lies at a deeper level and results from a failure to realize two things. First, the detection of physical invariants, like image surfaces, is exactly and precisely an information-processing problem, in modern terminology. And second, he vastly underrated the sheer difficulty of such detection. In discussing the recovery of three-dimensional information from the movements of an observer, he says that "in motion, perspective information alone can be used" (Gibson, 1966, p. 202). And perhaps the key to Gibson is the following:

The detection of non-change when an object moves in the world is not as difficult as it might appear. It is only made to seem difficult when we assume that the perception of constant dimensions of the object must depend on the correcting of illusions of constant form and size. The information for the constant dimension of an object is normally carried by invariant relations in an optic array. Rigidity is specified. (emphasis added)

Yes, to be sure, but how? Detecting physical invariants is just as difficult as Gibson feared, but nevertheless we can do it. And the only way to understand how is to treat it as an information-processing problem.

The underlying point is that visual information processing is actually very complicated, and Gibson was not the only thinker who was misled by the apparent simplicity of the act of seeing. The whole tradition of philosophical inquiry into the nature of perception seems not to have taken seriously enough the complexity of the information processing involved. For example, Austin's (1962) Sense and Sensibility entertainingly demolishes the argument, apparently favored by earlier philosophers, that since we are sometimes deluded by illusions (for example, a straight stick appears bent if it is partly submerged in water), we see sense-data rather than material things. The answer is simply that usually our perceptual processing does run correctly (it delivers a true description of what is there), but although evolution has seen to it that our processing allows for many changes (like constant illumination), the perturbation due to the refraction of light by water is not one of them. And incidentally although the example of the bent stick has been discussed since Aristotle, I have seen no philosophical inquiry into the nature of the perceptions of, for instance, a heron, which is a bird that feeds by pecking up fish first seen from above the water surface. For such birds the visual correction might be present.

Anyway, my main point here is another one. Austin (1962) spends much time on the idea that perception tells one about real properties of the external world, and one thing he considers is "real shape," (p. 66), a notion which had cropped up earlier in his discussion of a coin that "looked elliptical" from some points of view. Then so,

it had a real shape which remained unchanged. But coins in fact are rather special cases. For one thing their outlines are well defined and vary highly stable, and for another they have a known and a nameable shape. But there are plenty of things of which this is not true. What is the real shape of a cloud... or of a cat? Does its real shape change whenever it moves? If not, in what posture is its real shape on display? Furthermore, is its real shape such as to be fairly smooth outlines, or must it be finely enough erred to take account of each hair? It is pretty obvious that there is no answer to these questions—no reason according to which, no procedure by which, answers are to be determined. (emphasis added), (p. 67)

But there are answers to these questions. There are ways of describing the shape of a cat to an arbitrary level of precision (see Chapter 5), and there are rules and procedures for arriving at such descriptions. That is exactly what vision is about, and precisely what makes it complicated.

1.3 A REpresentational FRAMEWORK FOR VISION

Vision is a process that produces from images of the external world a description that is useful to the viewer and not cluttered with irrelevant information ( Marr, 1976; Marr and Nishihara, 1978). We have already seen that a process may be thought of as a mapping from one representation to another, and in the case of human vision, the initial representation is in no doubt—it consists of arrays of image intensity values as detected by the photoreceptors in the retina.

It is quite proper to think of an image as a representation; the items that are made explicit are the image intensity values at each point in the array, which we can conveniently denote by I(x,y) at coordinate (x,y). In order to simplify our discussion, we shall neglect for the moment the fact that there are several different types of receptor, and imagine instead that there is just one, so that the image is black-and-white. Each value of I(x,y) thus specifies a particular level of gray; we shall refer to each detector as a picture element or pixel and to the whole array f as an image.

But what of the output of the process of vision? We have already agreed that it must consist of a useful description of the world, but that requirement is rather nebulous. Can we not do better? Well, it is perfectly true that, unlike the input, the result of vision is much harder to discern, let
alone specify precisely, and an important aspect of this new approach is that it makes quite concrete proposals about what that end is. But before we begin that discussion, let us step back a little and spend a little time formulating the more general issues that are raised by these questions.

The Purpose of Vision

The usefulness of a representation depends upon how well suited it is to the purpose for which it is used. A pigeon uses vision to help it navigate, fly, and seek out food. Many types of jumping spiders use vision to tell the difference between a potential meal and a potential mate. One type, for example, has a curious retina formed of two diagonal strips arranged in a V. If it detects a real V on the back of an object lying in front of it, the spider has found a mate. Otherwise, maybe a meal. The frog, as we have seen, detects bugs with its retina; and the rabbit retina is full of special gadgets, including what is apparently a hawk detector, since it responds well to the patterns made by a preying hawk hovering overhead. Human vision, on the other hand, seems to be very much more general, although it clearly contains a variety of special-purpose mechanisms that can, for example, direct the eye toward an unexpected movement in the visual field or cause one to blink or otherwise avoid something that approaches one's head too quickly.

Vision, in short, is used in such a bewildering variety of ways that the visual systems of different animals must differ significantly from one another. Can the type of formulation that I have been advocating, in terms of representations and processes, possibly prove adequate for them all? I think so. The general point here is that because vision is used by different animals for such a wide variety of purposes, it is inconsiderate to believe that all seeing animals use the same representations; each can confidently be expected to use one or more representations that are nicely tailored to the owner's purposes.

As an example, let us consider briefly a primitive but highly efficient visual system that has the added virtue of being well understood. Werner Reichardt's group in Tübingen has spent the last 14 years patiently unravelling the visual flight-control system of the housefly, and in a famous collaboration, Reichardt and Tomaso Poggio have gone far toward solving the problem (Reichardt and Poggio, 1976, 1979; Poggio and Reichardt, 1976). Roughly speaking, the fly's visual apparatus controls its flight through a collection of about five independent, rigidly inflexible, very fast responding systems (the time from visual stimulus to change of torque is only 21 ms). For example, one of these systems is the landing system; if the visual field "explodes" fast enough (because a surface looms nearby), the fly automatically "lands" toward its center. If this center is above the fly, the fly automatically inverts to land upside down. When the feet touch, power to the wings is cut off. Conversely, to take off, the fly jumps; when the feet no longer touch the ground, power is restored to the wings, and the insect flies again.

In flight control is achieved by independent systems controlling the fly's vertical velocity (through control of the lift generated by the wings) and horizontal direction (determined by the torque produced by the asymmetry of the horizontal thrust from the left and right wings). The visual input to the horizontal control system, for example, is completely described by the two terms

\[ r(\phi) + D(\phi) \]

where \( r \) and \( D \) have the form illustrated in Figure 1-6. This input describes how the fly tracks an object that is present at angle \( \phi \) in the visual field and has angular velocity \( \dot{\phi} \). The system is triggered to track objects of a certain angular dimension in the visual field, and the motor strategy is such that if the visible object was another fly a few inches away, then it would be

![Figure 1-6](image-url)
intercepted successfully if the target was an elephant 100 yd away, interception would fail because the fly's built-in parameters are for another fly nearby, not an elephant far away.

Thus, fly vision delivers a representation in which at least these three things are specified: (1) whether the visual field is looming sufficiently fast that the fly should contemplate landing; (2) whether there is a small patch—it could be a black speck or it turns out, a textured figure in front of a textured ground—having some kind of motion relative to its background; and if there is such a patch, (3) φ and γ for this patch are delivered to the motor system. And that is probably about 60% of fly vision. In particular, it is extremely unlikely that the fly has any explicit representation of the visual world around him—no true conception of a surface, for example, but just a few triggers and some specifically fly-centered parameters like φ and γ.

It is clear that human vision is much more complex than this, although it may well incorporate subsystems not unlike the fly's to help with specific and rather low-level tasks like the control of pursuit eye movements. Nevertheless, as Poggio and Reichardt have shown, even these simple systems can be understood in the same sort of way as information-processing tasks. And one of the fascinating aspects of their work is how they have managed not only to formulate the differential equations that accurately describe the visual control system of the fly but also to express these equations, using the Volterra series expansion, in a way that gives direct information about the minimum possible complexity of connections of the underlying neuronal networks.

Advanced Vision

Visual systems like the fly's serve adequately and with speed and precision the needs of their owners, but they are not very complicated; very little objective information about the world is obtained. The information is all very much subjective—the angular size of the stimulus as the fly sees it rather than the objective size of the object out there, the angle that the object has in the fly's visual field rather than its position relative to the fly or to some external reference, and the object's angular velocity again in the fly's visual field, rather than any assessment of its true velocity relative to the fly or to some stationary reference point.

One reason for this simplicity must be that these facts provide the fly with sufficient information for it to survive. Of course, the information is not optimal and from time to time the fly will fritter away its energy chasing a falling leaf a medium distance away or an elephant a long way away as a direct consequence of the inadequacies of its perceptual system. But this apparently does not matter very much—the fly has sufficient excess energy for it to be able to absorb these extra costs. Another reason is certainly that translating these rather subjective measurements into more objective qualities involves much more computation. How, then, should one think about more advanced visual systems—human vision, for example. What are the issues? What kind of information is vision really delivering, and what are the representational issues involved?

My approach to these problems was very much influenced by the fascinating accounts of clinical neurology such as Critchley (1955) and Warrington and Taylor (1979). Particularly important was a lecture that Elizabeth Warrington gave at MIT in October 1973, in which she described the capacities and limitations of patients who had suffered left or right parietal lesions. For me, the most important thing that she did was to draw a distinction between the two classes of patient (see Warrington and Taylor, 1979). For those with lesions on the right side, recognition of a common object was possible provided that the patient's view of it was in some sense straightforward. She used the words conventional and unconventional—a water-pail or a clairvoy seen from the side gave "conventional" views but seen end-on gave "unconventional" views. If these patients recognized the object at all, they knew its name and its semantics—that is, its use and purpose, how big it was, how much it weighed, what it was made of, and so forth. If their view was unconventional—a pail seen from above, for example—not only would the patients fail to recognize it, but they would vehemently deny that it could be a view of a pail. Patients with left parietal lesions behaved completely differently. Often these patients had no language, so they were unable to name the viewed object or state its purpose and semantics. But they could convey that they correctly perceived its geometry—that is, its shape—even from the unconventional view.

Warrington's talk suggested two things. First, the representation of the shape of an object is stored in a different place and is therefore a quite different kind of thing from the representation of its use and purpose. And second, vision alone can deliver an internal description of the shape of a viewed object, even when the object was not recognized in the conventional sense of understanding its use and purpose.

This was an important moment for me for two reasons. The general trend in the computer vision community was to believe that recognition was so difficult that it required every possible kind of information. The results of this point of view duly appeared a few years later in programs like Freedman's (1974) and Tenenbaum and Barrow's (1976). In the latter program, knowledge about objects—for example, that desks have telephones on them and that telephones are black—was used to help "segment" out a blackblob halfway up an image and "recognize" it as a telephone. Freedman's program used a similar approach to "segment" and
"recognize" a hammer in a scene. Clearly, we do use such knowledge in real life; I once saw a brown blob quivering amongst the lettuce in my garden and correctly identified it as a rabbit, even though the visual information alone was inadequate. And yet here was this young woman calmly telling us not only that her patients could convey to her that they had grasped the shapes of things that she had shown them, even though they could not name the objects or say how they were used, but also that they could happily continue to do so even if she made the task extremely difficult visually by showing them peculiar views or by illuminating the objects in peculiar ways. It seemed clear that the intuitions of the computer vision people were completely wrong and that even in difficult circumstances shapes could be determined by vision alone.

The second important thing, I thought, was that Elizabeth Warrington had put her finger on what was somehow the quintessential fact of human vision—that it tells about shape and space and spatial arrangement. Here lay a way to formulate its purpose—building a description of the shapes and positions of things from images. Of course, that is by no means all that vision can do; it also tells about the illumination and about the reflectances of the surfaces that make the shapes—their brightnesses and colors and visual textures—and about their motion. But these things seemed secondary; they could be hung off a theory in which the main job of vision was to derive a representation of shape.

To the Desirable via the Possible

Finally, one has to come to terms with cold reality. Desirable as it may be to have vision deliver a completely invariant shape description from an image (whatever that may mean in detail), it is almost certainly impossible in only one step. We can only do what is possible and proceed from there toward what is desirable. Thus we arrived at the idea of a sequence of representations, starting with descriptions that could be obtained straightforwardly from an image but that are carefully designed to facilitate the subsequent recovery of gradually more objective, physical properties about an object's shape. The main stepping stone toward this goal is describing the geometry of the visible surfaces, since the information encoded in images, for example by stereopsis, shading, texture, contours, or visual motion, is due to a shape's local surface properties. The objective of many early visual computations is to extract this information.

However, this description of the visible surfaces turns out to be unsuitable for recognition tasks. There are several reasons why, perhaps the most prominent being that like all early visual processes, it depends critically on the vantage point. The final step therefore consists of transforming the viewer-centered surface description into a representation of the three-dimensional shape and spatial arrangement of an object that does not depend upon the direction from which the object is being viewed. This final description is object centered rather than viewer centered.

The overall framework described here therefore divides the derivation of shape information from images into three representational stages (Table 1-3): (1) the representation of properties of the two-dimensional image,

<table>
<thead>
<tr>
<th>Name</th>
<th>Purpose</th>
<th>Primitives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image(s)</td>
<td>Represents intensity</td>
<td>Intensity value at each point in the image</td>
</tr>
<tr>
<td>Primal sketch</td>
<td>Makes explicit important information about the multidimensional image, primarily the intensity changes there and their geometrical distribution and organization.</td>
<td>Zero-crossings</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lines</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Terminations and discontinuities</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Edge segments</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Virtual lines</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Groups</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cartesianal organization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Boundaries</td>
</tr>
<tr>
<td>2-D sketch</td>
<td>Makes explicit the orientation and rough depth of the visible surfaces, and contours of discontinuities in these quantities in a viewer-centered coordinate frame.</td>
<td>Local surface orientation (the &quot;needles&quot; primitives)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Distance from viewer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Discontinuities in depth</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Discontinuities in surface orientation</td>
</tr>
<tr>
<td>3-D model representation</td>
<td>Describes shapes and their spatial organization in an object-centered coordinate frame, using a modular hierarchical representation that includes volumetric primitives (i.e., primitives that represent the volume of space that a shape occupies) as well as surface primitives.</td>
<td>3-D models arranged hierarchically, each one based on a spatial configuration of a few sides or arcs, to which volumetric or surface shape primitives are attached</td>
</tr>
</tbody>
</table>
such as intensity changes and local two-dimensional geometry, (2) the representation of properties of the visible surfaces in a viewer-centered coordinate system, such as surface orientation, distance from the viewer, and discontinuities in these quantities; surface reflectance; and some coarse description of the prevailing illumination; and (3) an object-centered representation of the three-dimensional structure and of the organization of the viewed shape, together with some description of its surface properties. This framework is summarized in Table 1–3. Chapters 2 through 5 give a more detailed account.

PART II

Vision