VISION

A Computational Investigation into the Human Representation and Processing of Visual Information

David Marr
Late of the Massachusetts Institute of Technology

W. H. Freeman and Company
New York

Project Editor: Judith Wilson
Copy Editor: Paul Monsour
Production Coordinator: Linda Jupiter
Illustration Coordinator: Richard Quiñones
Designer: Ron Newcomer
Artists: Catherine Brandel and Victor Royer
Compositor: Graphic Typesetting Service
Printer and Binder: The Maple-Vail Book Manufacturing Group

Library of Congress Cataloging in Publication Data

Marr, David, 1945-1980.
Vision.

Bibliography: p.
Includes index.
1. Vision—Data processing. 2. Vision—Mathematical models. 3. Human information processing. I. Title.
QP475.M27 1982 152.1'4028'54 81-15076
ISBN 0-7167-1284-9

4 5 6 7 8 9 MP 1 0 8 9 8 7 6 5 4

To my parents and to Lucia

Note: Readers may require stereoscopic viewers in order to obtain the three-dimensional effects of the stereo images illustrated in this book. These viewers may be ordered from the following companies; please write to request current prices.

Hubbard Scientific Company
P.O. Box 104
Northbrook, Illinois 60062

Edmund Scientific Company
1776 Edscorp Building
Barrington, New Jersey 08007

The reader may be able to obtain the stereoscopic effect without an optical device: Hold the stereo image about ten inches away from the eyes and relax the eyes as if staring into the distance. Eventually the left-hand member of the pair seen by the right eye and the right-hand member of the pair seen by the left eye will merge to produce what will appear to be a three-dimensional image.

It will help to hold a fingertip about halfway between the stereo pair and your eyes. Adjust the position of the finger so that when looking with only your left eye, you see the finger in front of the right edge of the right-hand member of the pair. At the same time, when looking with your right eye only, try to see the finger in front of the right edge of the left-hand member of the pair. When your finger is so positioned, look at the finger with both eyes. This procedure will bring the two members of the stereo pair into registration, but they will be out of focus. Now relax your eyes and try to focus the stereo pair without losing the fixation on your finger. This trick seems to get easier as you get older.

Contents

Detailed Contents

PART II

VISION

Chapter 2

Representing the Image 41

Chapter 3

From Images to Surfaces 99

Preface

This book is meant to be enjoyed. It describes the adventures I have had in the years since Marvin Minsky and Seymour Papert invited me to the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology in 1973. Working conditions were ideal, thanks to Patrick Winston's skillful administration, to the generosity of the Advanced Research Projects Agency of the Department of Defense and of the National Science Foundation, and to the freedom arranged for me by Whitman Richards, under the benevolent eye of Richard Held. I was fortunate enough to meet and collaborate with a remarkable collection of people, most especially, Tomaso Poggio. Included among these people were many erstwhile students who became colleagues and from whom I learned much—Keith Nishihara, Shimon Ullman, Ken Forbus, Kent Stevens, Eric Grimson, Ellen Hildreth, Michael Riley, and John Batali. Berthold Horn kept us close to the physics of light, and Whitman Richards, to the abilities and inabilities of people.

In December 1977, certain events occurred that forced me to write this book a few years earlier than I had planned. Although the book has important gaps, which I hope will soon be filled, a new framework for studying vision is already clear and supported by enough solid results to be worth setting down as a coherent whole.

Many people have helped me to live through this somewhat difficult period. Particularly, my parents, my sister, my wife Lucia, and Jennifer, Tomaso, Shimon, Whitman, and Inge gave to me more than I often deserved; although mere thanks are inadequate, I thank them. William Prince steered me to Professor F. G. Hayhoe and Dr. John Rees at Addenbrooke's Hospital in Cambridge, and them I thank for giving me time.

Summer 1979 David Marr

We should like to express our gratitude to those who helped us bring David Marr's *Vision* to fulfillment.

We thank Gunther Stent, whose friendship brought David Marr and W. H. Freeman and Company together and whose sound guidance helped us prepare the book for publication.

We thank David Marr's colleague, Keith Nishihara, for his skill and great effort; the work could not have been finished without him.

We thank David Marr's assistant, Carol Papineau, for attending so well to the needs of the manuscript and the publisher.

We thank the vision group at the MIT Artificial Intelligence Laboratory, especially Ellen Hildreth and Eric Grimson, who participated in ways large and small to bring this book to life.

The Publisher

PART I

Introduction and Philosophical Preliminaries

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 primarily 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 $0.00—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 pejorative 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 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 scientists, the physiologists and anatomists 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 neuronal dogmas of Barlow (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. "If," he might say, "vision really 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 hint about the best algorithms for doing so. He doesn't want to know about rhodopsin, 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 computation capable of 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 machine 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 a little about 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.

[illegible] 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 satisfiable.

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 informally, 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 means 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.

CHAPTER 1

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 (1704), who laid the foundations for modern work on color vision, and Helmholtz (1910), whose treatise on physiological optics generates 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 *distinctness*, and with trying to formulate the "laws" that governed the creation of these wholes. The attempt failed for various reasons, and the Gestalt school dissolved into the fog of subjectivism. With the death of the school, many

Figure 1–1. A random-dot stereogram of the type used extensively by Bela Julesz. The left and right images are identical except for a central square region that is displaced slightly in one image. When fused binocularly, the images yield the impression of the central square floating in front of the background.

of its early and genuine insights were unfortunately lost to the mainstream of experimental psychology.

Since then, students of the psychology of perception have made no serious attempts at an overall understanding of what perception is, concentrating instead on the analysis of properties and performance. The trichromatism of color vision was firmly established (see Brindley, 1970), and the preoccupation with motion continued, with the most interesting developments perhaps being the experiments of Miles (1931) and of Wallach and O'Connell (1953), which established that under suitable conditions an unfamiliar three-dimensional shape can be correctly perceived from only its changing monocular projection.*

The development of the digital electronic computer made possible a similar discovery for binocular vision. In 1960 Bela Julesz devised computer-generated random-dot stereograms, which are image pairs constructed of dot patterns that appear random when viewed monocularly but fuse when viewed one through each eye to give a percept of shapes and surfaces with a clear three-dimensional structure. An example is shown in Figure 1–1. Here the image for the left eye is a matrix of black and white squares generated at random by a computer program. The image for the

*The two dimensional image seen by a single eye.

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 meant 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

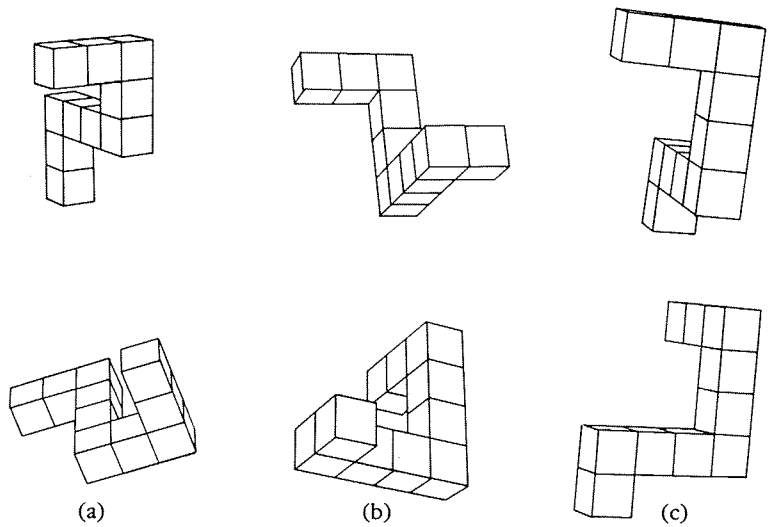

Figure 1–2. Some drawings similar to those used in Shepard and Metzler's experiments on mental rotation. The ones shown in (a) are identical, as a clockwise turning of this page by 80° will readily prove. Those in (b) are also identical, and again the relative angle between the two is 80°. Here, however, a rotation in depth will make the first coincide with the second. Finally, those in (c) are not at all identical, for no rotation will bring them into congruence. The time taken to decide whether a pair is the same was found to vary linearly with the angle through which one figure must be rotated to be brought into correspondence with the other. This suggested to the investigators that a stepwise mental rotation was in fact being performed by the subjects of their experiments.

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 explanation? For a long time, the best hope seemed to lie along another line of investigation, that of electrophysiology. The development 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 neurons (Granit and Svaetichin, 1939; Hartline, 1938; Galambos and Davis, 1943). This led to the notion of a cell's "receptive field" (Hartline, 1940) and to the Harvard School's famous series of studies of the behavior of neurons at successively deeper levels of the visual pathway (Kuffler, 1953; Hubel and Wiesel, 1962, 1968). 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 (1953) 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 handheld 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 moved rapidly to and fro within the unit's receptive field. This causes a vigorous discharge which can be 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 behavioural response is often dramatic; they turn towards the target and make repeated feeding responses consisting of a jump and snap. The selectivity of the retinal neurons and the frog's reaction when they are selectively stimulated, suggest that they are "bug detectors" (Barlow 1953) performing a primitive but vitally important form of recognition.
>
> The result makes one suddenly realize that a large part of the sensory machinery involved in a frog's feeding responses may actually reside in the retina rather than in mysterious "centres" that would be too difficult to understand by physiological methods. The essential lock-like property resides in each member of a whole class of neurons and allows the cell to discharge only to the appropriate key pattern of sensory stimulation. Lettvin *et al.* (1959) suggested that there were five different classes of cell in the frog, and Barlow, Hill and Levick (1964) found an even larger number of categories in the rabbit. [Barlow *et al.*] called these key patterns "trigger features," and Maturana *et al.* (1960) emphasized another important aspect of the behaviour of these ganglion cells; a cell continues to respond to the same trigger feature in spite of changes in light intensity over many decades. The properties of the retina are such that a ganglion cell can, figuratively speaking, reach out and determine that something specific is happening in front of the eye. Light is the agent by which it does this, but it is the detailed pattern of the light that carries the information, and the overall level of illumination prevailing at the time is almost totally disregarded. (p. 373)

Barlow (1972) then goes on to summarize these findings in the following way:

> The cumulative effect of all the changes I have tried to outline above has been to make us realise that each *single neuron can perform a much more complex and subtle task than had previously been thought* (emphasis added). Neurons do not loosely and unreliably remap the luminous intensities of the visual image onto our sensorium, but instead they detect pattern elements, discriminate the depth of objects, ignore irrelevant causes of variation and are 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 their pattern selectivity permanently modified by early visual experience. This amounts to a revolution in our outlook. It is now quite inappropriate to regard unit activity as a noisy indication of more basic and reliable processes involved in mental operations: instead, we must regard single neurons as the prime movers of these mechanisms. Thinking is brought about by neurons and we should not use phrases like "unit activity reflects, reveals, or monitors thought processes," because the activities of neurons, quite simply, are thought processes.
>
> This revolution stemmed from physiological work and makes us realize that the activity of each single neuron may play a significant role in perception. (p. 380)

This aspect of his thinking led Barlow to formulate the first and most important of his five dogmas: 'A description of that activity of a single nerve cell which is transmitted to and influences other nerve cells and of a nerve cell's 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 behaviour' (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 stereopsis (Barlow, Blakemore, and Pettigrew, 1967) and on color (DeValois, Abramov, and Mead, 1967; Gouras, 1968) seemed to confirm the close links between perception and single-cell recordings, and the intriguing results of Gross,

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 physiologists 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 dogma, 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 Fink-Heimer technique from Nauta'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 neurophysiologists 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 mainstream 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, Coombs, and Henry, 1971; Schiller, Finlay, and Volman, 1976a, 1976b), or on cells in species like the owl (for example, Pettigrew and Konishi, 1976). 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 has 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—Gross'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 clear that something important was missing that was not present in either of the disciplines of neurophysiology or psychophysics. The key observation is that neurophysiology and psychophysics 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 first great revelation was that the problems are difficult. Of course, these days this fact is a commonplace. But in the 1960s almost no one realized that machine vision was difficult. The field had to go through the same experience as the machine translation field did in its fiascoes of the 1950s before it was at last realized that here were some problems that had to be taken seriously. The reason for this misperception is that we humans are ourselves so good at vision. The notion of a feature detector was well established by Barlow and by Hubel and Wiesel, and the idea that extracting edges and lines from images might be at all difficult simply did not occur to those who had not tried to do it. It turned out to be an elusive problem: Edges that are of critical importance from a three-dimensional point of view often cannot be found at all by looking at the intensity changes in an image. Any kind of textured image gives a multitude of noisy edge segments; variations in reflectance and illumination cause no end of trouble; and even if an edge has a clear existence at one point, it is as likely as not to fade out quite soon, appearing only in patches along its length in the image. The common and almost despairing feeling of the early investigators like B.K.P. Horn and T.O. Binford was that practically anything could happen in an image and furthermore that practically everything did.

Three types of approach were taken to try to come to grips with these phenomena. The first was unashamedly empirical, associated most with Azriel Rosenfeld. His style was to take some new trick for edge detection, texture discrimination, or something similar, run it on images, and observe the result. Although several interesting ideas emerged in this way, including the simultaneous use of operators* of different sizes as an approach to increasing sensitivity and reducing noise (Rosenfeld and Thurston, 1971), these studies were not as useful as they could have been because they were never accompanied by any serious assessment of how well the different algorithms performed. Few attempts were made to compare the merits of different operators (although Fram and Deutsch, 1975, did try), and an approach like trying to prove mathematically which operator was optimal was not even attempted. Indeed, it could not be, because no one had yet formulated precisely what these operators should be trying to do. Nevertheless, considerable ingenuity was shown. The most clever was probably Hueckel's (1973) operator, which solved in an ingenious way the problem of finding the edge orientation that best fit a given intensity change in a small neighborhood of an image.

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 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 shapes 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 Binford–Horn line finder (Horn, 1973) was used to find edges, and both it and its sequel (described in Shirai, 1973) 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 (1968), and they culminated in the works of Waltz (1975) and Mackworth (1973), which essentially solved the interpretation problem for line drawings derived from images of prismatic solids. Waltz'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 Waltz'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, those changes due to the reflectance alone could be isolated. Horn devised a

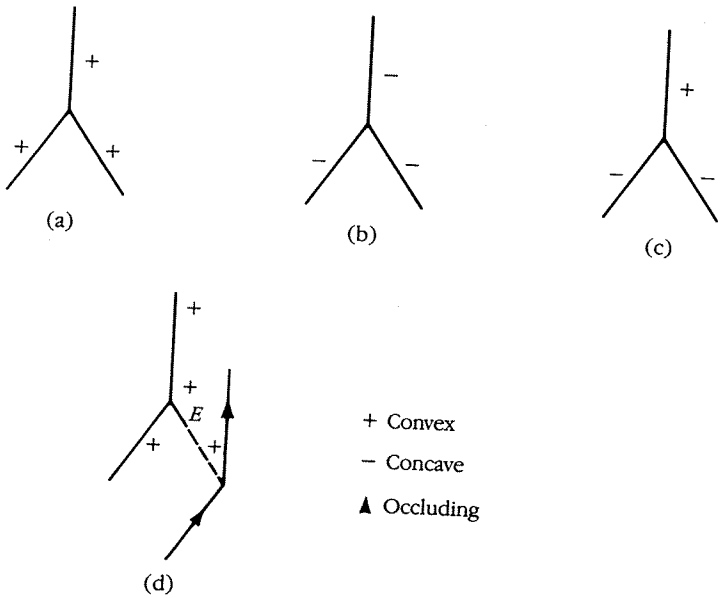

Figure 1–3. Some configurations of edges are physically realizable, and some are not. The trihedral junctions of three convex edges (a) or of three concave edges (b) are realizable, whereas the configuration (c) is impossible. Waltz cataloged all the possible junctions, including shadow edges, for up to four coincident edges. He then found that by using this catalog to implement consistency relations [requiring, for example, that an edge be of the same type all along its length like edge *E* in (d)], the solution to the labeling of a line drawing that included shadows was often uniquely determined.

clever parallel algorithm for this, and I suggested how it might be implemented by neurons in the retina (Marr, 1974a).

I do not now believe that this is at all a correct analysis of color vision or of the retina, but it showed the possible style of a correct analysis. Gone are the ad hoc programs of computer vision; gone is the restriction to a special visual miniworld; gone is any explanation *in terms of* neurons—except as a way of implementing a method. And present is a clear understanding of what is to be computed, how it is to be done, the physical assumptions on which the method is based, and some kind of analysis of algorithms that are capable of carrying it out.

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 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 factors—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 effort 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 complicated 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 levels of description that are linked, at least in principle, into a cohesive 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 (Marr and Nishihara, 1978).

For example, the Arabic, Roman, and binary numeral systems are all formal systems for representing numbers. The Arabic representation consists 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 $3 \times 10^1 + 7 \times 10^0$, 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 definition, 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 had.

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 on output; or should one sacrifice efficiency of circuitry to carry out operations directly in a decimal representation? On the whole, business computers and pocket calculators take the second approach, and general purpose 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 determines 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 this 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 notions 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 $(3 + 4) \rightarrow 7$. 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: $(4 + 0) \rightarrow 4$. Also, for every number there is a unique "inverse," written (-4) in the case of 4, which when added to the number gives zero: $[4 + (-4)] \rightarrow 0$.

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. Thus part of this 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 it 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 methodically 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. Hillis and B. Silverman in a quite different technology, in a computer made out of Tinkertoys, a children's wooden building set. The whole monstrously ungainly engine, which actually works, currently resides in a museum at the University of Missouri in St. Louis.

Some styles of algorithm 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 ($\times 10^4$) 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 center is the choice of representation for the input and output and the

Computational theory	Representation and algorithm	Hardware implementation
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?

Figure 1–4. The three levels at which any machine carrying out an information-processing task must be understood.

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 other two.

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

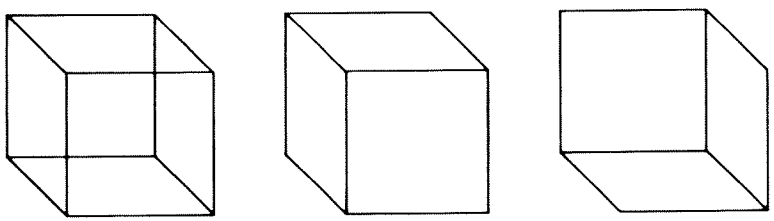

Figure 1–5. The so-called Necker illusion, named after L. A. Necker, the Swiss naturalist who developed it in 1832. The essence of the matter is that the two-dimensional representation (a) has collapsed the depth out of a cube and that a certain aspect of human vision is to recover this missing third dimension. The depth of the cube can indeed be perceived, but two interpretations are possible, (b) and (c). A person's perception characteristically flips from one to the other.

brain, but few would feel satisfied by an account that failed to mention the existence of two different but perfectly plausible three-dimensional interpretations of this two-dimensional image.

For some phenomena, the type of explanation required is fairly obvious. Neuroanatomy, for example, is clearly tied principally to the third level, the physical realization of the computation. The same holds for synaptic mechanisms, action potentials, inhibitory interactions, and so forth. Neurophysiology, too, is related mostly to this level, but it can also help us to understand the type of representations being used, particularly if one accepts something along the lines of Barlow's views that I quoted earlier. But one has to exercise extreme caution in making inferences from neurophysiological findings about the algorithms and representations being used, particularly until one has a clear idea about what information needs to be represented and what processes need to be implemented.

Psychophysics, on the other hand, is related more directly to the level of algorithm and representation. Different algorithms tend to fail in radically different ways as they are pushed to the limits of their performance or are deprived of critical information. As we shall see, primarily psychophysical evidence proved to Poggio and myself that our first stereo-matching algorithm (Marr and Poggio, 1976) was not the one that is used by the brain, and the best evidence that our second algorithm (Marr and Poggio, 1979) *is* roughly the one that is used also comes from psychophysics. Of course, the underlying computational theory remained the same in both cases, only the algorithms were different.

Psychophysics can also help to determine the nature of a representation. The work of Roger Shepard (1975), Eleanor Rosch (1978), or Elizabeth Warrington (1975) provides some interesting hints in this direction. More specifically, Stevens (1979) 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 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 bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings 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