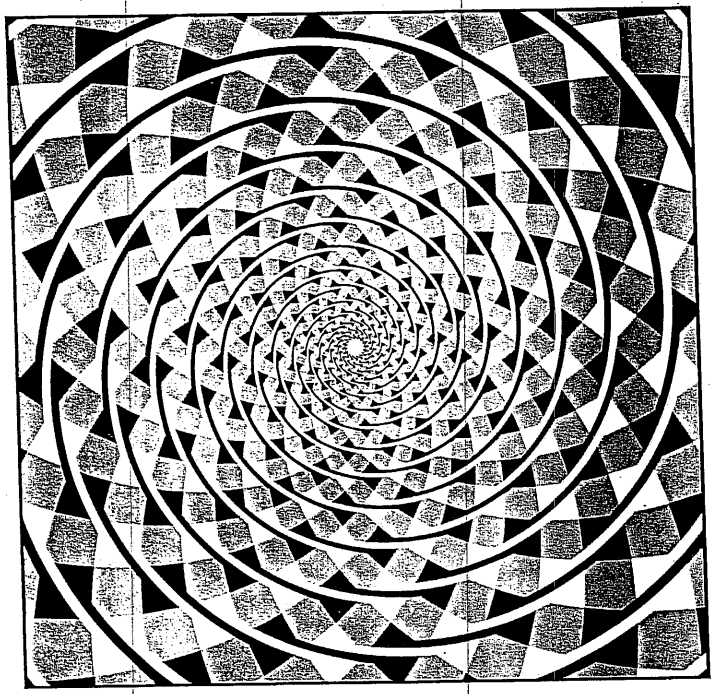


Figure 1.1 Fraser's spiral. Human vision is not quite as infallible as we tend to believe. A case in point is the illusory spiral evoked here by a collection of concentric circles, each circle comprising segments angled toward the center. To convince yourself that there is indeed no spiral, try tracing a spiral with your finger. (After [Fraser 1908].)

Chapter 1

Introduction

If our long-sought quest to create autonomous anthropomorphic automata is to succeed, we must first impart human perceptual capabilities to machines. It has now been well over two decades since several individuals and groups first made concerted efforts to automate visual perception, and yet a sense of frustration seems to prevail. Clearly, this sentiment is prompted to no small extent by the existence of sophisticated human perceptual abilities. We must bear in mind, however, not only that these abilities are the outcome of millions of years of evolution, but also that human perception is fallible; the fallibility of human vision is revealed, for instance, by the well-known Fraser's illusory spiral in Figure 1.1. Nevertheless, perhaps the impatience is justified. Perhaps, computer-vision research has indeed been painstakingly unproductive; or, perhaps, the working framework adopted by the research community is hopelessly flawed. It is too early to tell. This book is an account of the current state of our understanding.

1.1 What Is Computer Vision?

Computer vision, sometimes called image understanding, describes the automatic deduction of the structure and properties of a possibly dynamic three-dimensional world from either a single or multiple two-dimensional images of the world. The images may be monochromatic (i.e., "black and white") or colored, they may be captured by a single or multiple cameras, and each camera may be either stationary or mobile.

The structure and properties of the three-dimensional world that we seek to deduce in computer vision include not only geometric properties, but also material properties and the lighting of the world. Examples of geometric properties are the shapes, sizes, and locations of objects, and examples of material properties are the lightness or darkness of surfaces, their colors, their textures, and their material compositions. If the world is changing while it is being imaged, we might also wish to infer the nature of this change, and, perhaps, predict the future.

Why is computer vision difficult to realize? Because, as we shall explore at some length in this book, image formation is a many-to-one mapping: A variety of surfaces with different material and geometrical properties, possibly under different lighting conditions, could lead to identical images. As a result, the inverse imaging problem, given a single image—that is, the problem of inferring from a single image the scene that led to the image—has no unique solution. Equivalently, a single image of a scene does not constrain the imaged scene sufficiently to allow us to recover the scene unambiguously. There are two ways out of this predicament: (1) gather more data (images), and (2) make assumptions about the world. It is, of course, important that every assumption we invoke be tenable, and that we understand the exact role of the assumption well. Hence, we must make all our assumptions explicit, as we shall endeavor to do throughout this book.

Even when the inverse imaging problem is sufficiently constrained to allow a unique solution in principle, there remain the twin practical problems of computability and robustness. Is the solution computable using reasonable resources? Resources include both computing machinery and time. And, is the solution robust? That is, is the solution insensitive to errors in data (i.e., to signal noise), and to errors in computation (e.g., owing to limited precision arithmetic)? Failure on any of these fronts can render an otherwise promising approach useless in practice.

You might wonder whether computer vision is not a form of image processing or pattern recognition. Although there is some overlap, the differences are distinct. Image processing is a generic term for the

processing of images to produce new images that are more desirable in some fashion; [Pratt 1991] and [Rosenfeld and Kak 1982] are two standard references for image processing. Image processing encompasses the following: **image enhancement**, which modifies images to improve their appearance to human viewers; **image restoration**, which corrects images for degradations (such as motion blur); and **image compression**, which represents images compactly while maintaining acceptable image quality. **Pattern recognition**, or **pattern classification**, on the other hand, classifies patterns into one of a finite (usually small) number of prespecified categories; [Duda and Hart 1973] is the classic reference for pattern recognition. For the most part, the emphasis in pattern recognition is on two-dimensional patterns—for instance, on the letters of the alphabet. Computer vision, in contrast, is concerned with generating descriptions of three-dimensional scenes—scenes that are not constrained to be members of predetermined sets—from two-dimensional images of the scenes.

The purpose of computer vision is to infer the state of the physical world from the inherently noisy and ambiguous images of the world. If the current state of the art is any indication, such a goal is difficult to accomplish in a reliable, robust, and efficient manner. By the same token, such a goal is challenging. One problem in vision is diversity: the inherent diversity in any nontrivial domain with respect to the objects and their relative configurations, and the ensuing diversity in the images, each of which may be acquired from anywhere in any given scene. This diversity necessitates opportunism: Of the many possible sources of information in an image, only a few are typically present, and we must do what we can with what we have.

In addition to the information provided by the images themselves, at times, it might also be possible to bring to bear knowledge about objects, their behavior, and the context. We shall not explore the use of domain-specific knowledge here. Further, we shall restrict ourselves to **passive sensing**—that is, to the sensing of radiation that is already present in the scene, rather than to the sensing of radiation that is actively controlled and introduced by the observer. The latter is termed **active sensing**. Although active sensing—for example, using laser stripes—can greatly simplify the computation of scene structure, it interferes with the state of the world, and is, therefore, not always tolerable. Neither is active sensing always feasible.

The approach to computer vision we shall adopt in this book is a modular one; an alternative approach, for instance, might seek to perform one grand optimization. We, in contrast, shall endeavor to identify and isolate the various sources of information in an image; these sources include,

for instance, brightness discontinuities, shading, and motion. One may use each of these sources of information either individually, or in conjunction with the others, to make deductions about the scene. The benefits of a modular approach are more than pedagogic: A modular approach makes it easier to control and monitor the performance of a system, to debug the system, and to understand and improve the system.

Computer vision has several applications less ambitious than the creation of anthropomorphic robots. In fact, one may argue that mimicking human behavior is the wrong agenda for robotics to begin with [Whitney 1986]. Applications of computer vision include the following: automation (e.g., on the assembly line), inspection (e.g., of integrated-circuit chips to detect defects in them), remote sensing (e.g., of possibly hostile terrain to generate its relief maps), human–computer communication (e.g., through gesturing), and aids for the visually impaired (e.g., mechanical guide dogs). See [Brady 1982] for a detailed list of applications.

1.2 A Word About Your Sponsor

You might wonder why computer-vision researchers do not simply build systems that emulate the human visual system, especially considering the wealth of literature in neurophysiology, psychology, and psychophysics. Gregory [Gregory 1978] provides a fascinating introduction to the latter topics; see [Levine and Shefner 1991] and [Kaufman 1974] for more detailed treatments, and [Levine 1985] for an engineering perspective. One good reason why computer-vision researchers do not emulate human vision is that what is known about the human visual system beyond the human eye is largely disjointed, speculative, and meager. Further, although human vision is certainly adequate for most frequently encountered tasks—or is it that our lifestyles are adapted to tasks that can be accommodated by our perceptual faculties—adequacy must not be taken to imply infallibility. The fallibility of human vision is amply demonstrated by the existence of visual illusions, ambiguities, and inconsistencies, one of which was illustrated in Figure 1.1, and several others of which we shall examine in Section 1.2.1. Before we proceed any further, however, let us pause for a moment to consider whether perception has any meaning in the absence of what is commonly understood to be intelligence; such a consideration is especially pertinent given that computer vision has its origins in a field called *artificial intelligence*.

Is visual perception an integral component of what we commonly term intelligence? Perhaps not. Although many researchers subscribe to the view

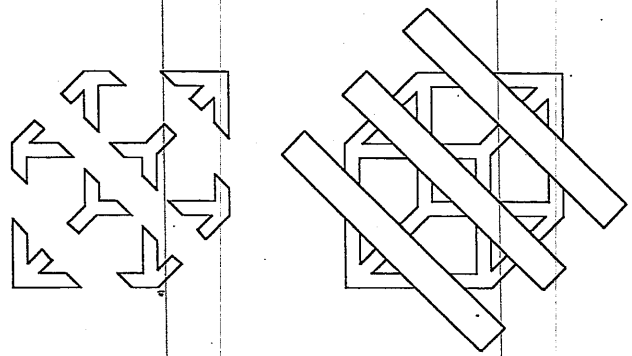


Figure 1.2 Seeing versus thinking. We might argue that seeing and thinking are distinguishable human activities—that seeing is more immediate. Although we can "think" of the fragments in (a) as constituting a cube, it is difficult to "see" a cube. In contrast, we readily "see" a cube in (b), which we can derive from (a) by introducing three "opaque" stripes. Note that the cube in (b) is, in fact, ambiguous: It is the well-known Necker cube, whose vertical face closest to and facing the viewer may reside either in the lower-left corner or in the upper-right corner. (After [Kanizsa 1979], by permission of the Greenwood Publishing Group, Inc., Westport, CT. Copyright © 1979 by Gaetano Kanizsa.)

that perception is inextricably linked to cognition—that what we perceive is often a consequence of logical reasoning—Kanizsa [Kanizsa 1979] argues forcefully that seeing and thinking are clearly distinguishable human activities. He maintains that although thinking and reasoning may lead us to *conceive* of a spatiotemporal phenomenon based on some visual data, *seeing* is more immediate in that it does not require conscious mental effort. Kanizsa buttresses his argument with the two drawings shown in Figure 1.2. Although we can conceive the disjointed fragments in Figure 1.2(a) to constitute a cube, we do not perceive a cube, at least not right away. The introduction of three "opaque" stripes, as in Figure 1.2(b), however, changes our perception: It makes the cube perceptually apparent. Although no one can deny the critical role of perception in the acquisition of information by humans, one could argue that seeing is just a "mechanical act" in that

(barring illusions) it does not originate anything.[1] All that the act of seeing does is infer the state of the world to the extent allowed by the sensed data. And, in doing so, it provides food for thought, to conceptualize and classify—to assign membership in an equivalence class based on form or function, thereby implicitly ascribing properties that are not perceived, but rather are only postulated. Although the question of what constitutes intelligence is of considerable intellectual import, it has no direct bearing on our discussion here; hence, we shall not delve into it further.

Even if we were indeed seeking to duplicate human vision—and say that we had a better understanding of its mechanisms—it is not at all obvious that blind emulation would be the recommended route. For one, evolution took its course under a set of physical and computational constraints that are substantially different from the technological barriers that confront us today. A disclaimer is perhaps in order here. Not for a moment am I suggesting that research in biological vision is of little use to the advancement of computer vision. On the contrary, such research is exciting, enlightening, and stimulating. After all, it is biological systems that provide us with the proof of the possibility of high-performance general-purpose vision. However, given that relatively definitive accounts are currently available only for the workings of the human eye (see Section 1.2.2), we are well advised to be cautious in seeking to ground computer vision in what we think we know about human vision.

1.2.1 Visual Illusions, Ambiguities, and Inconsistencies

As already indicated, the fallibility of the human visual system is amply demonstrated by the existence of visual illusions, ambiguities, and inconsistencies, several of which we shall now examine. See [Gregory 1978] and [Frisby 1980] for additional examples.

We already encountered one visual illusion, the well-known Fraser's spiral, in Figure 1.1. You can confirm easily that the spiral in Figure 1.1 is illusory by trying to trace the spiral. Figure 1.3 illustrates several other classical optical illusions.

1. This line of reasoning has an important historical precedent. It was used by Lady Lovelace in her mid-nineteenth century account to describe the limitation of Babbage's Analytical Engine, the forerunner of the modern-day computer; see, for instance, p. 284, [Babbage et al. 1961]. It also often lies at the heart of arguments denying the possibility of a thinking machine—see, for instance, [Turing 1950]—but that is another matter. See the book by Penrose [Penrose 1989] in this connection.

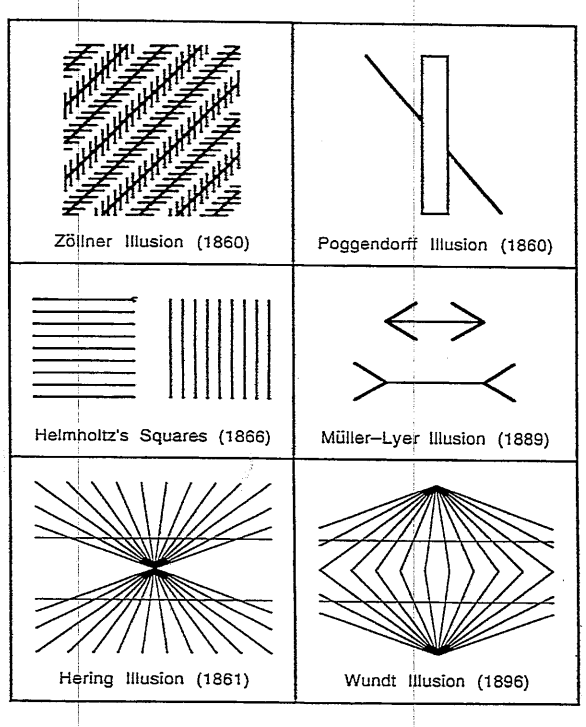


Figure 1.3 Six classical optical illusions. In each of the drawings, geometric truths appear untrue. The illusory effect is so strong that you might wish to have a ruler handy to verify the various assertions. In the Zöllner illusion, the diagonals are parallel, but appear otherwise. In the Poggendorff illusion, the two diagonal straight-line segments seem offset even though they are collinear. Helmholtz's two squares appear rectangular. In the Müller–Lyer illusion, the horizontal line with the reversed arrowheads appears longer than the line with the normal arrowheads, even though both lines have the same length. Finally, in both the Hering and Wundt illusions, the two horizontal and parallel straight lines appear bowed. (See [Boring 1942] for the origins of these optical illusions.)

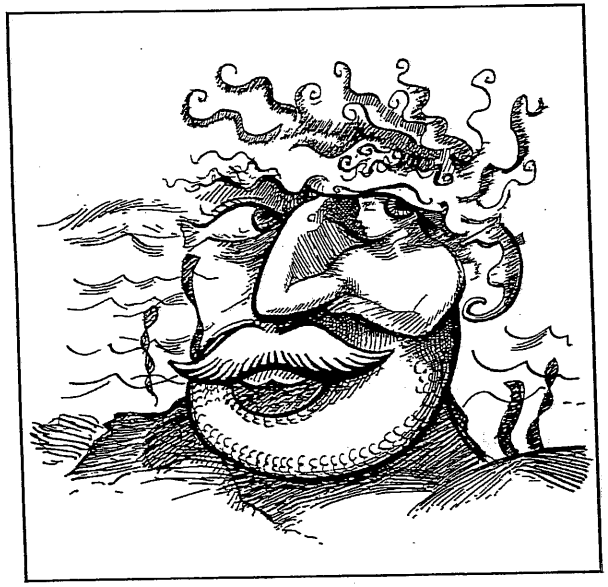


Figure 1.4 *Poseidon and the Mermaid*, by David M. Weimer, 1990. This India-ink drawing is ambiguous: It may be perceived either as Poseidon, the Greek god of the sea, or as a mermaid. The mermaid's tail fin is Poseidon's moustache. (Courtesy, David M. Weimer.)

At times, a figure may evoke more than a single interpretation—that is, the figure may be ambiguous. The multiple interpretations of a figure may either coexist, or one interpretation may dominate the other(s). Weimer's *Poseidon and the Mermaid*, illustrated in Figure 1.4, is an excellent example of an ambiguous figure: It may be perceived either as Poseidon or as a mermaid. Several other visually ambiguous figures are shown in Figure 1.5.

Finally, it is possible that, although what we perceive from a figure is neither ambiguous nor completely illusory, it is globally unrealizable in that we cannot physically construct the perceived three-dimensional object in its entirety in three-dimensional space. This possibility is illustrated beautifully

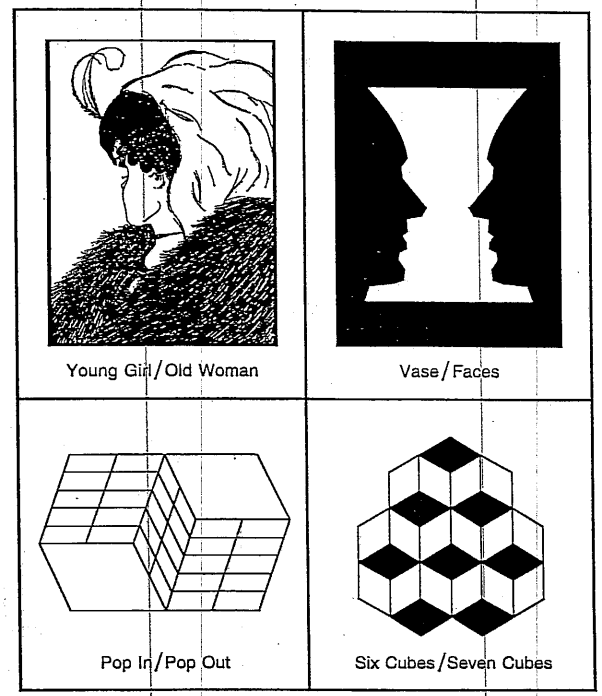


Figure 1.5 Four well-known visual ambiguities. In each of the drawings, two interpretations compete for attention. In the young-girl/old-woman illustration, the young girl's chin is the old woman's nose. The interpretation of the vase/faces illustration depends on whether the black region is seen as the background or as the foreground. Finally, the pop-in/pop-out and six-cubes/seven-cubes ambiguities depend on the spontaneous reversal of the perceived concavities and convexities. (See [Boring 1930] for the origin of the young-girl/old-woman ambiguity, which is based on a cartoon by Hill in 1915; see [Boring 1942] for the origin of the vase/faces ambiguity, which is based on an ambiguous figure by Rubin in 1915; the pop-in/pop-out and six-cubes/seven-cubes ambiguities are based on the Schröder staircase (illustrated in Figure 4.2), which was proposed in 1858 and whose origin is described in [Boring 1942].)

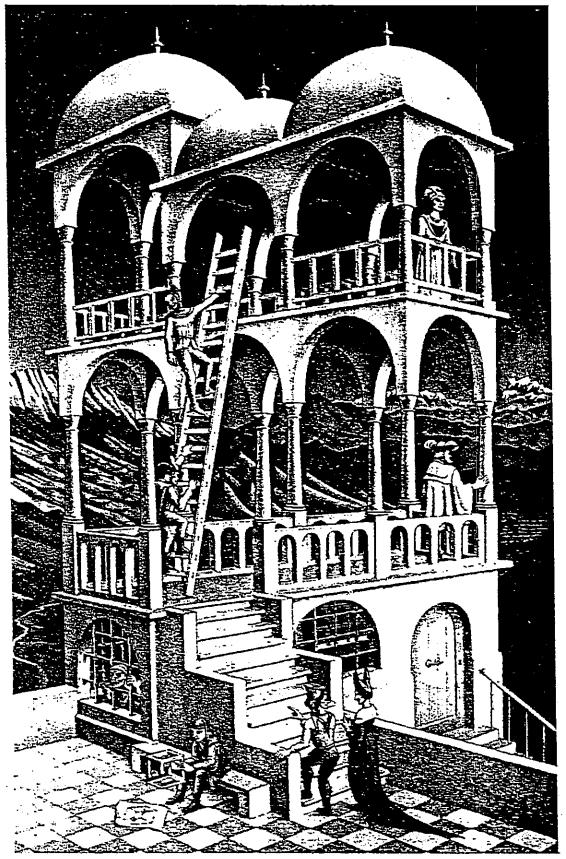


Figure 1.6 *Belvedere*, by Maurits C. Escher, 1958. This well-known lithograph exhibits several geometrical inconsistencies that are not readily apparent. The middle-level pillars cross from the front to the rear, and vice versa; the ladder's base is inside the building, but its top is outside; the topmost level is at right angles to the middle level; the cube being examined by the person on the bench is geometrically impossible. (Copyright © 1990 by M. C. Escher Heirs / Cordon Art – Baarn – Holland.)

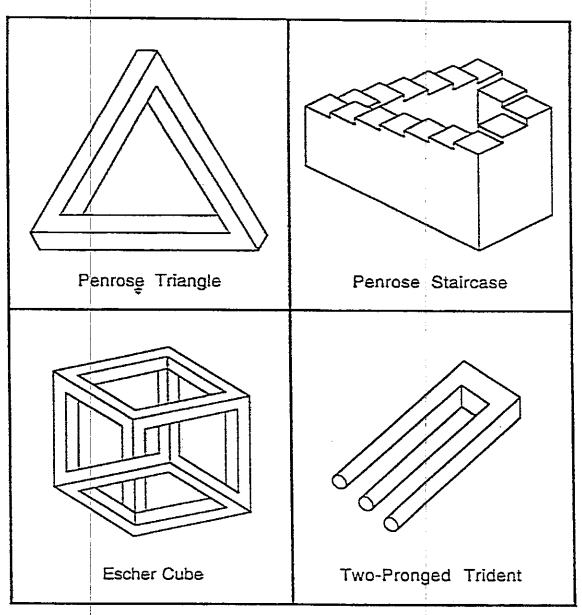


Figure 1.7 Four well-known visual inconsistencies. In each of the drawings, geometrical inconsistency makes the perceived object physically unrealizable. The Penrose triangle (after [Penrose and Penrose 1958]) appears to be a polyhedron, but if it were one, it could not form a closed loop. The Penrose staircase (after [Penrose and Penrose 1958]) descends (ascends) all the way around to the starting step. The Escher cube (after Escher's lithograph, *Belvedere*, 1958) has a top that is inconsistent with its base. Finally, the two-pronged trident (after [Gregory 1965]) appears to terminate in three prongs that have circular cross-sections, whereas the base of the trident appears to have only two prongs with rectangular cross-sections.

by *Belvedere*, the well-known lithograph by Escher reproduced in Figure 1.6. Although the lithograph at first seems to portray a perfectly ordinary scene, a closer examination reveals several geometrical inconsistencies that render the perceived scene physically unrealizable. Figure 1.7 illustrates four other "inconsistent drawings."

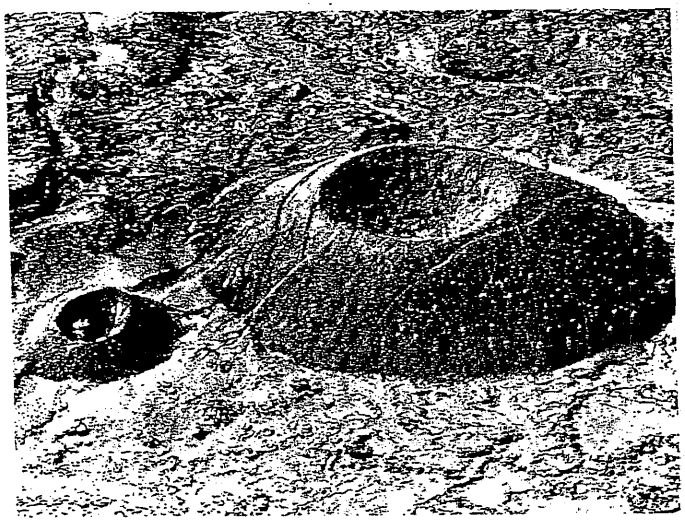


Figure 1.8 An image whose interpretation by humans changes when the image is turned upside down. If you turn this photograph of two lava cones with craters upside down, you will perceive it to be the image of two craters with mounds. This reversal of the perceived concavities and convexities is apparently due to an implicit assumption by you, the viewer, that the scene is lit from above. (After [Rittenhouse 1786]. Photograph provided by Associated Press / Wide World Photos, 1972.)

All the examples of visual illusions, ambiguities, and inconsistencies rendered thus far might strike you as contrived. It may seem to you that their success at "manipulating" our visual system depends at least in part on their lack of "realism." After all, if we cannot believe what we see, what are we to believe? This absolute faith in our perceptual faculties is not entirely justified. Consider, for instance, Figure 1.8—a routine photograph of a natural scene showing lava cones with craters at their tips. There seems no cause for confusion here, now that we have a real scene. However, the photograph remains unambiguous only until we turn it upside down: Then, the cones become craters, and the craters become cones.

The various visual illusions, ambiguities, and inconsistencies furnished here are more than just curiosities. They provide us with valuable insights

into the nature of human vision, and, in addition, they raise the following all-important question: Is human vision just controlled hallucination? That is, do we infer from our retinal images more than what is supported by the geometry and physics of image formation? Helmholtz, in his justly celebrated *Handbook of Physiological Optics*, first published in the middle of the nineteenth century, expresses the view that, "Every image is the image of a thing merely for him who knows how to read it, and who is enabled by the aid of the image to form an idea of the thing" (p. 24, [Helmholtz 1910], English translation).[2] The implication of this assertion is that it is not as though the human visual system is making precise and exact inferences based on the physics of image formation in the eye, but rather that the human visual system is invoking some rules of thumb that are derived from and biased by the prior experience of the individual, and, perhaps, of the species. As a result, humans may "see" what is not (i.e., hallucinate), and they may not "see" what is (i.e., overlook). Whereas we are quite forgiving when it comes to the performance of humans, we are not quite as charitable when it comes to the performance of machines. Hence, we need ask ourselves this: Do we really wish to make machines see as we do?

1.2.2 The Eye and Beyond

Irrespective of whether or not we seek to emulate human vision—in either one or both of form and function—it is pertinent to ask, what is it exactly that we know about the human visual system? If nothing else, the answer to this question would educate and enlighten us, and would perhaps even suggest strategies for machine vision.

2. Hermann von Helmholtz (1821–1894), the author of the *Handbook of Physiological Optics* ([Helmholtz 1909], [Helmholtz 1910], [Helmholtz 1911]), was one of the preeminent scientists of the nineteenth century. He made fundamental contributions to a wide variety of disciplines, including physiology, optics, electrodynamics, mathematics, and meteorology. He is best known, however, for his statement of the law of conservation of energy. Helmholtz was an empiricist—he denied the doctrine of innate ideas and held experience to be the basis of all knowledge—and his empiricism is reflected in his work. Helmholtz's greatest work, the *Handbook of Physiological Optics* (1856–1866), is an epitome of the scientific method. It is the single most important treatise on the physics and physiology of human vision to this day. Helmholtz's inventions in connection with human vision included the ophthalmoscope, which is an instrument for viewing the interior of the eye, and the ophthalmometer, which is an instrument for making measurements of the eye. Helmholtz's other great work on sensory perception, *On the Sensations of Tone as a Physiological Basis for the Theory of Music* (1862), laid the foundations for the science of acoustics.

That the eye is an organ of sight is obvious: All you need to do to verify this assertion is to close your eyes. However, it is not equally obvious what the exact role of the eye might be. Pythagoras and his followers, circa 500 B.C., supported the "emanation hypothesis of vision." According to this hypothesis, visual rays emanate in straight lines from the eye, spreading with immeasurable speed and consummating the act of vision on touching the perceived object. You may think of this mechanism of vision as analogous to how a blind person might discover her surroundings by groping with a long cane. The emanation hypothesis, however strange it may sound now, remained widely accepted in various forms for centuries, until Kepler, in 1604, correctly proposed that the eye is an optical instrument that forms a real inverted image on its back surface; see Figure 2.1. You are referred to Polyak's monumental *The Vertebrate Visual System* [Polyak 1957] both for an extended historical account of investigations into the eye, and for intricate physiological details of the eye.

The human eye is roughly a globe about 2 cm in diameter, free to rotate in its orbit under the control of six extrinsic muscles. Figure 1.9 illustrates the schematic of the horizontal cross-section of the right human eye, viewed from above. Light enters the eye through the tough transparent **cornea**, passes through the watery **aqueous humor** that fills the **anterior chamber**, proceeds to the crystalline **lens**, and then through the gelatinous **vitreous humor**, finally to form an inverted image on the photosensitive **retina**. Directly in front of the lens is an annular opaque muscular membrane, the **iris**, which gives the eye its color. Light can enter the eye only through the circular aperture of the iris, the **pupil**, whose size is controlled by the expansion and contraction of the iris. The lens is held in place by the **suspensory ligament**, through which the **ciliary muscle** adjusts the curvature of the lens. Barring the region of the eye where the cornea is located, the eye is covered by a dense fibrous opaque sheath called the **sclera**, part of which is seen as the white of the eye. Between the sclera and the retina lies the heavily pigmented **choroid**, which absorbs the light that passes through the retina undetected.

The adjustment of the curvature of the lens of an eye by its ciliary muscle is called **accommodation**. Accommodation adapts the eye for clearest vision at a particular distance—in a healthy eye, objects at this distance are imaged on the retina rather than in front of or behind the retina. Objects that are imaged in front of or behind the retina appear blurred to the viewer. **Nearsightedness**, or **myopia**, describes the inability of an eye to bring into focus on its retina objects that are distant from the eye—these objects are

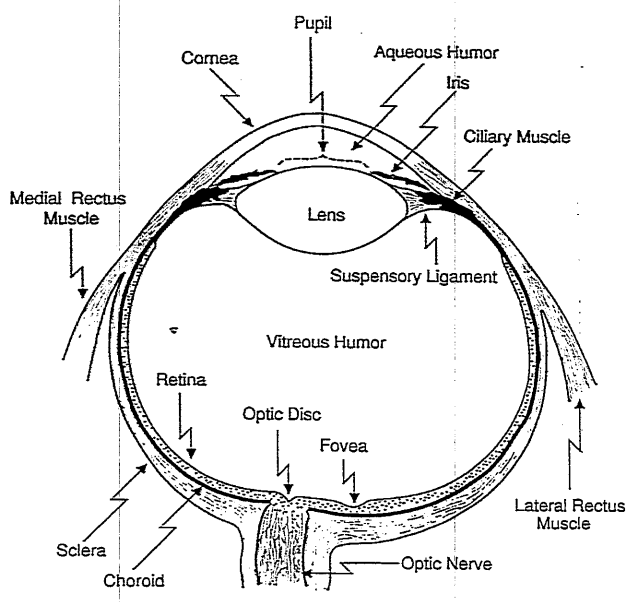


Figure 1.9 Schematic of a horizontal cross-section of a right human eye, viewed from above. Light passes in sequence through the cornea, the aqueous humor, the lens, and the vitreous humor, finally to form an image on the photosensitive *retina*. The retina encodes the image into nerve impulses, and transmits these impulses to the brain via the optic nerve. Clearest vision is realized in a circular depression in the retina called the *fovea*. The optic disc, where the optic nerve leaves the retina, is "blind." The ciliary muscle, which attaches to the lens through the suspensory ligament, controls the curvature of the lens. The iris, which is an opaque annular muscular membrane that gives the eye its color, controls the size of the pupil, which is the aperture in the iris that is the sole entrance for light into the eye. The sclera is a dense, fibrous, opaque sheath covering much of the eye. The choroid is a heavily pigmented layer that lies between the sclera and the retina.

imaged in front of the retina. **Farsightedness**, or **hypermetropia**, describes the inability of an eye to bring into focus on its retina objects that are close to the eye—these objects are imaged behind the retina. Both nearsightedness and farsightedness can usually be corrected by external lenses.

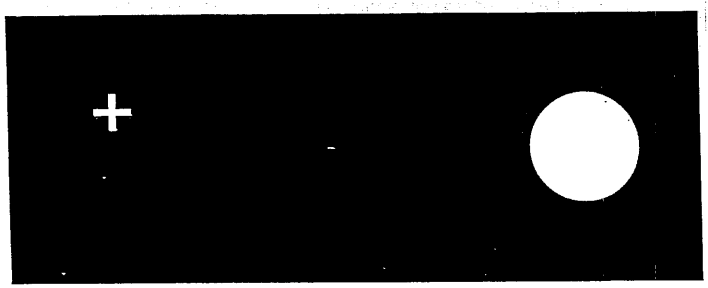


Figure 1.10 Discovery of your blind spot. You can discover your blind spot as follows. First, close your left eye and hold this book in the normal reading position about a foot or so away from your face. Then, look steadily at the white cross with your right eye, and doing so move the book slowly toward and away from your open eye, until, at a particular position of the book, the white disc disappears completely from your view. The white disc will seem to disappear completely from its black background when the retinal image of the disc is formed within your optic disc (see Figure 1.9), which is also known as the *blind spot*. (After [Mariotte 1668] and [Helmholtz 1911].)

The Retina

The **retina** is a complex nervous membrane with a mosaic of photoreceptors that on being stimulated by light produce electrical nervous signals. These signals are transmitted to the brain via the **optic nerve**, evoking the experience of vision. The location on the retina where all the individual nerve fibers constituting the optic nerve come together is called the **optic disc**. This region is free of photoreceptors, and hence, is often called the **blind spot**. (Everyone has a blind spot—you can experience yours by performing the simple experiment outlined in Figure 1.10.) Not far from the blind spot, near the axis of the lens, is a shallow pit with a high concentration of photoreceptors. This pit is the **fovea**. Although only a fraction of a millimeter in diameter, the fovea is of paramount importance as it provides the greatest visual acuity in brightly lit scenes. It is at the fovea that objects are imaged when we direct our gaze toward them.

The retina, which is a fraction of a millimeter thick—that is, about as thick as a few sheets of typing paper—has been the subject of several revealing studies since the advent of the compound microscope. According

to its greatest investigator, Ramón y Cajal, "The preference that the best known anatomists and histologists have given to this area is easily understood, since a knowledge of the arrangements of retinal cells is essential for a full understanding of vision and the many problems associated with it" (p. 781, [Ramón y Cajal 1892–1893], English translation). Figure 1.11 shows a photograph of a vertical section of a human retina taken from about 1.25 mm from the center of the fovea; a schematic of the photograph is shown alongside the photograph. The retina in the photograph was stained so that its structure would be highlighted—in a living eye, the retina is largely transparent. The first important point to note is that the optic nerve fibers run close to the vitreous humor, whereas the **rods** and **cones**—the two types of photoreceptors, each named after its shape—are located near the choroid. This arrangement is counterintuitive as it requires that the light pass through almost the entire depth of retinal tissue—complete with blood vessels (not shown in the figure), nerve cells, and nerve fibers—before it can be sensed. At the fovea, however, unlike the rest of the retina, much of the retinal tissue between the photoreceptors and the vitreous humor is displaced to one side, creating a shallow pit that has direct access to light.

The second important point to note with respect to the structure of the retina, illustrated in Figure 1.11, is that the photosensitive rods and cones do not have a continuous physical link to the optic nerve fibers by means of which they transmit signals to the brain. This absence of a physical link was first established by Ramón y Cajal, who showed that the retina comprises three distinct layers of nerve cells, and that these nerve cells communicate with one another through junctions, called **synapses** (see [Polyak 1957]). We shall not delve into the details of the structure and function of the retina—see [Rodieck 1973] for these details. Suffice it to say that, as illustrated in Figure 1.11, the rods and cones transmit signals to the optic nerve fibers, which extend from the **ganglion cells**, by means of the **bipolar cells**. At the input end of the bipolar cells are the **horizontal cells**, and at the output end of the bipolar cells are the **amacrine cells**, both horizontal and amacrine cells providing lateral (synaptic) interconnections between the nerve cells that run vertically in Figure 1.11. These lateral interconnections determine the response of a ganglion cell to a photostimulus within an extended retinal area called the ganglion cell's **receptive field** (see the excellent article by Werblin [Werblin 1973] in this connection). It seems that at least one function of the retinal nerve cells that mediate between the photoreceptors and the optic nerve fibers is to condense the information contained in the light impinging on the retina. This hypothesis is supported by the sheer numbers of the photoreceptors and the optic nerve fibers: Whereas the total number of cones in a human eye is about 6 million, and the total number of

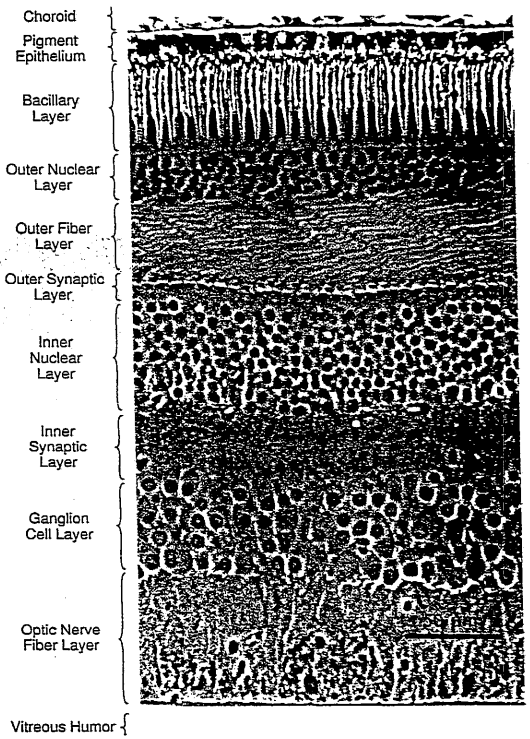
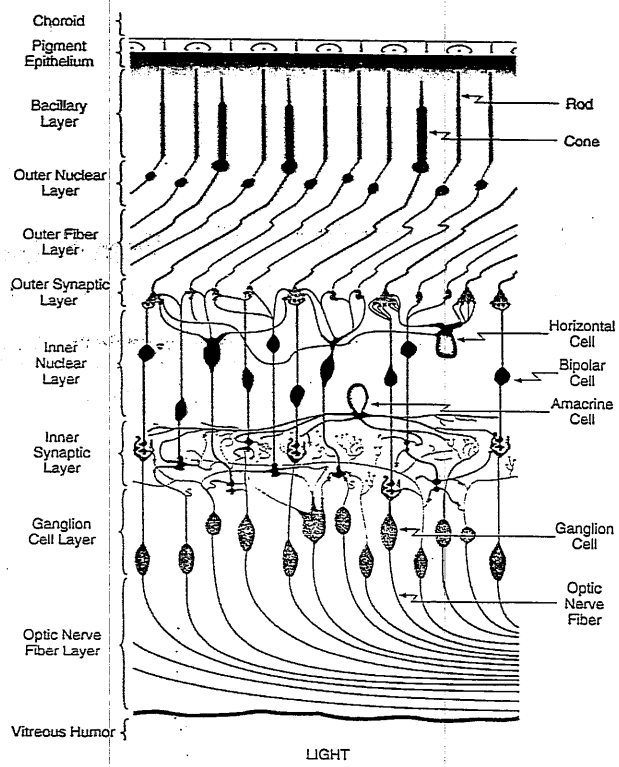


Figure 1.11 The human retina. A photograph of a vertical cross-section of a human retina is shown on this page, and the schematic corresponding to this photograph is shown on the facing page. The photograph, which was obtained through phase-contrast microscopy, is of a retinal section about 0.4 mm thick, this section taken from about 1.25 mm from the center of the fovea. Light strikes the retina through the vitreous humor, which is at the bottom of the photograph, and then proceeds through much of the retinal tissue to be detected by the photoreceptors, which are of two types: *rods* and *cones*. The rods, which are capable of detecting light at much fainter levels than are the cones, facilitate vision in the dark. The cones, on the other hand—which unlike the rods come in three distinct varieties, each variety with a preferred sensitivity to red, green, or blue light—provide us with color vision. At the fovea, which is the region of the retina that provides the clearest vision, the inner layers of the retina that lie between the vitreous humor and the photoreceptors are pushed to one side so that light can impinge more directly on the photoreceptors. It is to accommodate this dislocation of retinal tissue that the



nerve fibers that lead from the nuclei of the rods and cones to the outer synaptic layer have highly slanted trajectories in the neighborhood of the fovea, as in the photograph shown. At the outer synaptic layer, the rod and cone terminals form synapses with the bipolar cells. A *synapse* is a site at which one cell of the nervous system transmits signals to another cell. The bipolar cells lead to the ganglion cells, with which they form synapses at the inner synaptic layer. It is the nerve fibers of the ganglion cells that come together to form the optic nerve, which leads to the brain. As the optic nerve fibers in all number only of the order of 1 million, in contrast to about 125 million rods and cones, the signals that are generated by the rods and cones in response to light impinging on them must of necessity be compacted and encoded before they can be transmitted to the brain. This compaction is accomplished with the aid of horizontal and amacrine cells, which provide lateral synaptic interconnections at the outer and inner synaptic layers, respectively. (Photograph from [Boycott and Dowling 1969] with permission, and schematic after [Polyak 1957].)

rods is about 120 million, the total number of nerve fibers leaving a human eye is only of the order of 1 million (see [Pirenne 1967]).

Rods and cones, and hence human eyes, are sensitive to only a small fraction of the electromagnetic spectrum; the electromagnetic spectrum includes not only the light visible to us but also radio waves, infrared rays, ultraviolet rays, and X rays. Whereas the rods all exhibit a similar variation in sensitivity to light over the spectrum of visible colors, the cones come in three distinct varieties, each variety being most sensitive to either the red, the green, or the blue portion of the visible spectrum. As a result, it is the cones that provide us with color vision. The rods, on the other hand, come in handy for vision in dimly lit scenes owing to their ability to detect light at much fainter levels than can be detected by the cones. In this connection, see, for instance, [Cornsweet 1970]. The fovea, which provides the greatest visual acuity, has only cones, whereas much of the rest of the retina has a much higher concentration of rods than of cones. Consequently, our vision in brightly lit scenes is sharp and colored, whereas our vision in dimly lit scenes is blurred and colorless. (As an aside, it might interest you to learn that it is the absence of rods from the fovea that prompts astronomers to "look off" the fovea when they wish to detect faint stars.)

All in all, the human eye is a truly remarkable device. To this day, no feat of human engineering has come even remotely close in performance. The sensitivity of the human eye approaches the absolute limit set by the quantum nature of light, and the maximum visual acuity of the human eye is high enough for the wave nature of light to have a bearing; see, for instance, [Pirenne 1967] and [Barlow 1981].

The Visual Pathways to the Brain

Although the past two centuries have witnessed substantial gains in our understanding of the structure and function of the human eye, "the more central parts of the visual system have little that is not at the moment mysterious" (p. 7, [Barlow 1981]). It is the conversion of representations of retinal images into knowledge of the world that constitutes the barrier to our understanding of human vision. Although we do have an idea of the major visual pathways from the eyes to the brain, we know little of what happens in the brain.

Figure 1.12 illustrates the major visual pathways from the eyes to the brain. From each eye emerges an **optic nerve**, which carries electrical nervous signals from the eye to the brain. The fibers constituting each optic nerve can be divided into two groups: those that originate on the inner nasal side of the eye, and those that originate on its outer temporal side. Fibers

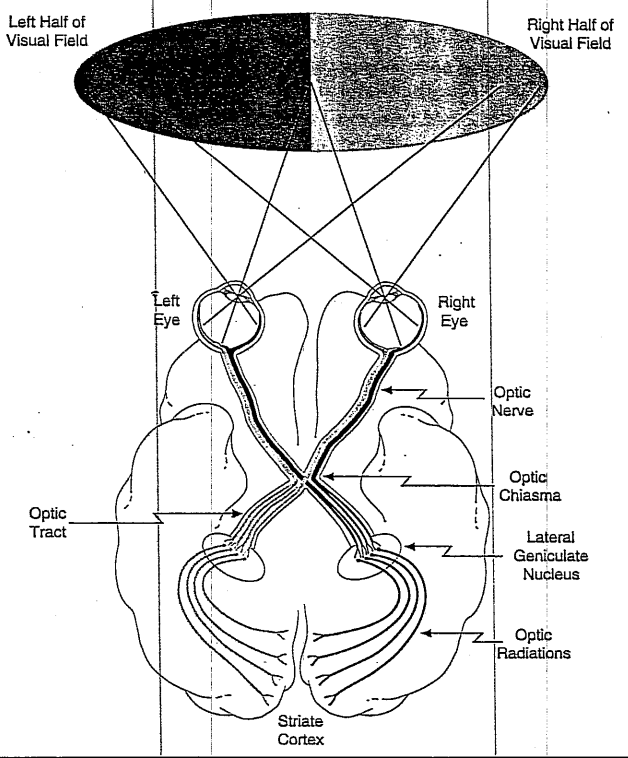


Figure 1.12 The major visual pathways from the eyes to the brain. The pattern of light striking each retina is encoded into nerve impulses, and these impulses are transmitted to the brain via the optic nerve that emerges from the retina. The left half of the visual field, which is imaged on the right half of each retina, is transmitted to the right half of the brain; the right half of the visual field, which is imaged on the left half of each retina, is transmitted to the left half of the brain. The cross-over of optic nerve fibers that is necessary to realize such a mapping takes place at the optic chiasma. From the optic chiasma, the nerve fibers proceed to the lateral geniculate nuclei via the optic tracts; from the lateral geniculate nuclei, nerve impulses are transmitted to the striate cortex of the brain via the optic radiations. It is important for binocular stereoscopic depth perception that each of the possibly two retinal images of a point in the visual field be mapped onto the same region of the brain.

originating on the temporal side of each eye go to the same side of the brain as the eye where they originate. In contrast, fibers originating on each nasal side cross over at the **optic chiasma** and proceed to the opposite side of the brain as the eye where they originate. Thus, as indicated in Figure 1.12, the left half of the visual field is mapped onto the right half of the brain, and the right half of the visual field is mapped onto the left half of the brain. What is important is that the two retinal images of any point in the scene that is visible to both eyes are mapped onto the same region of the brain. It is the disparity between the two retinal images that makes stereoscopic depth perception possible within the field of view shared by the two eyes, also called the **stereoscopic field.**[3] (The term *stereoscopic* literally implies the seeing of objects in three dimensions. However, its use here is more specific: *Stereoscopic* is taken to mean the seeing of objects in three dimensions on the basis of the retinal disparity between the images formed in the left and the right eye. As we shall see at length in this book, it is possible to perceive an object in three dimensions without the use of both eyes.)

To appreciate the significance of stereoscopic vision, you may perform the following simple experiment. Set up an uncapped pen on its end on a flat surface, perhaps with the aid of a few books, and then withdraw your hands some few feet away from the pen. Then, keeping your head stationary, try to cap the pen, first with one eye closed, and then with both eyes open. You should find it much easier to cap the pen with both eyes open.

Returning to the primary visual pathways from the eyes to the brain illustrated in Figure 1.12, from the optic chiasma, the nerve fibers proceed in two groups, the **optic tracts,** each tract comprising fibers that originate on the same side of each of the two retinas. Most of the nerve fibers constituting each of the two optic tracts proceed to the corresponding **lateral geniculate nucleus,** which is mainly a relay station where the nerve fibers make their first postretinal synapse. From the two lateral geniculate nuclei, signals are transmitted via the **optic radiations** to the **striate cortex,** which gets its name from its striped appearance in a fresh human brain. The importance of the

3. It seems that binocular animals that are predators have both eyes in front of their heads, like humans, and as a result, they have a large stereoscopic field of view. Binocular stereoscopic vision is important to predators as it allows them to judge accurately without moving the distance to their prey. In contrast to binocular predators, binocular prey have one eye on either side of their heads, an arrangement that provides prey with a large total field of view at the expense of a large stereoscopic field. A large total field of view allows prey to determine whether a predator is present over a large range of directions, and such a determination is clearly of higher priority to prey than is the accurate estimation of the distance of a predator.

fovea is manifested in the striate cortex by the disproportionately large area dedicated to the fovea there. In comparison to what we know of the human eye, little is known of what happens in the region of the brain dedicated to vision; see [Kuffler, Nicholls, and Martin 1984], [Peters and Jones 1985], and [Hubel 1988] for perspectives on the state of our understanding.

1.3 What Is to Come

This book aims to lead you gently through a guided tour of computer vision, stopping along the way to emphasize concepts and their significance. For the more seasoned among you, there shall be details and pointers, especially to recent developments. Although these details and pointers, which can safely be ignored without loss of continuity, may at first seem obscure to those among you who are unfamiliar with the terrain, they could later serve as guide maps for unaccompanied explorations. There are no prerequisites per se to join this tour. Curiosity and interest are taken for granted.

Figure 1.13 provides a sketch of a plausible schema for a general-purpose computer-vision system. By **general purpose** here is meant "without restriction to a particular task or domain." The boxes in the figure represent data, and the arrows indicate processes and the direction of data flow. The computer-vision paradigm in Figure 1.13 is by far the most common; hence, for purposes of discussion, let us adopt it here too.

In the computer-vision paradigm of Figure 1.13, the three-dimensional scene is first viewed by one or multiple cameras to produce either monochromatic or colored images. We shall restrict ourselves to monochromatic images in this book. The images thus acquired are processed so that the brightness discontinuities within the images are detected—these discontinuities are termed *edges*—and perhaps also so that the images are segmented into relatively homogeneous image regions. Then, the images and their edge and homogeneous-region maps are used to constrain the possible interpretations of the imaged world. Currently, the most widely investigated processes used to constrain the three-dimensional world are based on the following image characteristics: line drawings (i.e., edge maps of images), shading (i.e., variation in image brightness), variation in image texture, the disparity between images in stereo image pairs, and image evolution under motion of the camera relative to the scene. The significance of edges, shading, texture, stereo, and motion in the interpretation of images is highlighted by Figures 3.1, 5.1, 6.1, 7.1, and 8.1, respectively. The three-dimensional constraints derived from images may be either local (e.g., the depths, orientations, and reflectances of surface points), or global (e.g., the

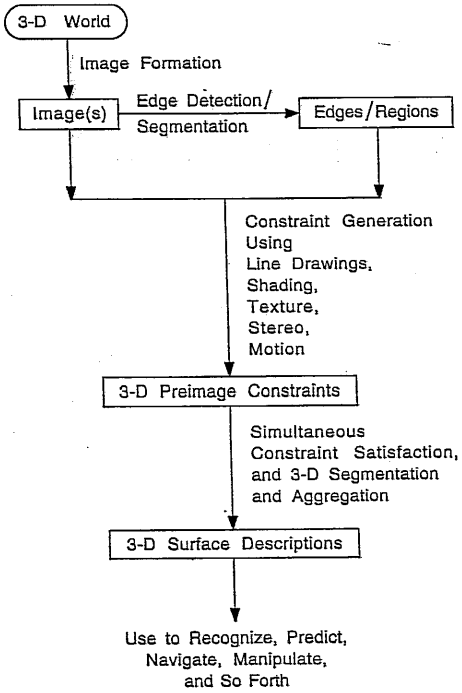


Figure 1.13 A plausible schema for general-purpose computer vision. *Computer vision* describes the automatic deduction of the structure and properties of a (possibly dynamic) three-dimensional world from its two-dimensional image(s). The boxes in the figure denote data, the arrows indicate processes and the direction of data flow, and 3-D is an abbreviation for three-dimensional. The term *preimage* in the figure refers to the imaged scene. In general, the preimage of any point in the range of a mapping is the point or collection of points in the domain of the mapping that map onto the particular point in the range—in the context of imaging, the mapping is image formation, whose domain is the three-dimensional world and whose range is the two-dimensional image.

restriction of a surface to be a surface of revolution). At any rate, once a collection of such constraints has been derived, surface descriptions may be generated by simultaneously enforcing all the preimage constraints (i.e., constraints on the imaged scene) and partitioning the data into sets arising from independent surfaces. These descriptions, preferably simultaneously metric and symbolic, may serve a variety of purposes: recognition, prediction, navigation, manipulation, and the performance of other tasks that require cognition and planning. (The **preimage** of any point in the range of a mapping is the point or collection of points in the domain of the mapping that map onto the particular point in the range—in the context of imaging, the mapping is image formation, whose domain is the three-dimensional world and whose range is the two-dimensional image.)

Although modularity is a convenient design procedure, and the absence of feedback avoids tricky control problems, it is clear that in Figure 1.13, the strict demarcation between the various processes (and data) and the restriction to forward data flow are both potentially limiting to robustness. For instance, edge detection is likely to proceed better in conjunction with the interpretation of edges. All the processes in Figure 1.13 could conceivably be modulated by succeeding data; as in all feedback loops, stability would be the primary ensuing concern. Nevertheless, despite its obvious limitations, as Figure 1.13 does represent the currently most popular paradigm, its components are what we shall discuss.

In Chapter 2, we shall discuss three aspects of image formation: geometry, radiometry, and sensing. This discussion will lay the foundation for the treatment of other topics. In Chapter 3, we shall examine edge detection and image segmentation. First, we shall consider popular schemes to detect edge fragments in images, and then, we shall turn our attention to the organization of such fragments into extended edges, and the description of these edges; subsequently, we shall review image-segmentation techniques. In Chapters 4, 5, 6, 7, and 8, we shall review the fundamentals of, and examine the progress made toward, constraining the imaged world using line drawings, shading, texture, stereo, and motion, respectively. These topics are at the heart of computer-vision research today. Owing to the limited success to this point of such efforts in generating robust three-dimensional preimage constraints, simultaneous constraint satisfaction and three-dimensional segmentation and aggregation have received relatively scant attention in the literature. Hence, we shall not devote a separate chapter to these topics. In Chapter 9, we shall first examine the attributes that make a representation desirable, and then, we shall discuss several shape-representation strategies in this light. Finally, in Chapter 10, we shall consider pointers to some of the

topics that have not previously received our attention. These topics include the following: so-called high-level tasks, such as object recognition, that may require three-dimensional surface descriptions; industrial applications; active range finding; and color vision.

1.4 A Bibliographical Note

Computer vision, as we know it today, had its beginnings in the seminal work of Roberts [Roberts 1965], who developed computer programs to deduce the three-dimensional structure and arrangement of a few simple trihedral-vertex polyhedra from their digital images. Roberts's modest success with his *blocks world* prompted high hopes. Back in the 1960s, the emerging and ambitiously named field, *artificial intelligence*, was arousing spectacular short-term expectations. Soon, however, researchers realized that visual perception is a nontrivial intellectual enterprise, and that techniques developed to analyze polyhedral scenes almost never lend themselves to more general settings.

Since the early work of Roberts, substantial time and effort have been devoted to computer vision, and the ensuing results have been documented in several publications. The principal among these publications are the journals *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*IEEE PAMI*), *International Journal of Computer Vision* (*IJCV*), and *Computer Vision, Graphics, and Image Processing* (*CVGIP*), now with the subtitle *Image Understanding* (*CVGIP: IU*). Other journals of interest include *Artificial Intelligence, Biological Cybernetics, Journal of the Optical Society of America A, Pattern Recognition Letters, Pattern Recognition, IEEE Transactions on Robotics and Automation, International Journal of Robotics Research,* and *IEEE Transactions on Systems, Man, and Cybernetics*. Occasionally, survey articles surface in the *Computing Surveys* and the *Proceedings of the IEEE*. The most prominent conferences in the field are the *International Conference on Computer Vision* (*ICCV*) and the *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR*). Other related conferences include the *International Conference on Pattern Recognition*, the *International Symposium on Robotics Research*, the *IEEE International Conference on Robotics and Automation*, and the *International Joint Conference on Artificial Intelligence*. New results obtained at several U.S. universities are first reported in the proceedings of the *Image Understanding Workshop*, which is organized at regular intervals by the Defense Advanced Research Projects Agency (DARPA) of the United States. It is not uncommon to also find references to reports and memoranda—however, as only a minuscule fraction of the audience has ready access to any such document, this author finds the practice ill-advised.

Two popular books on the subject are *Computer Vision* by Ballard and Brown [Ballard and Brown 1982], and *Robot Vision* by Horn [Horn 1986]; the latter provides an in-depth coverage of relatively few topics. Several survey articles and collections of papers have also been published. Among the surveys, the two most prominent are [Barrow and Tenenbaum 1981a] and [Brady 1982]; both are slightly dated; however, the former provides a valuable historical perspective. Among the collections of papers, the most significant one that comprises papers not published elsewhere is [Hanson and Riseman 1978]; this collection includes two fairly influential position papers, one by Marr [Marr 1978], and the other by Barrow and Tenenbaum [Barrow and Tenenbaum 1978]. In addition to the various books and surveys, every year, Rosenfeld publishes a handy exhaustive bibliography in *CVGIP*. Given that most theoretical results are of largely untested utility, and experimental claims of unsubstantiated robustness, computer-vision books and surveys of necessity reflect personal perspectives. This book will be no different.