An Introduction to Vision Science

1.1 Visual Perception
1.1.1 Defining Visual Perception
1.1.2 The Evolutionary Utility of Vision
1.1.3 Perception as a Constructive Act
   Adaptation and Aftereffects
   Reality and Illusion
1.1.4 Perception as Modeling the Environment
   Visual Completion
   Impossibility
   Predicting the Future
1.1.5 Perception as Apprehension of Meaning
   Attention and Consciousness

1.2 Optical Information
1.2.1 The Behavior of Light
   Illumination
   Interaction with Surfaces
   The Ambient Optic Array
1.2.2 The Formation of Images
   Optical Images
   Projective Geometry
   Perspective and Orthographic Projection
1.2.3 Vision as an "Inverse" Problem

1.3 Visual Systems
1.3.1 The Human Eye
   Eye and Brain
   Anatomy of the Eye
   Physiological Optics
1.3.2 The Retina
   Neurons
   Photo receptors
   Peculiarities of Retinal Design
   Pathways to the Brain
1.3.3 Visual Cortex
   Localization of Function
   Occipital Cortex
   Parietal and Temporal Cortex
   Mapping Visual Cortex
   The Physiological Pathways Hypothesis
Most of us take completely for granted our ability to see the world around us. How we do it seems no great mystery: We just open our eyes and look! When we do, we perceive a complex array of meaningful objects located in three-dimensional space. For example, Figure 1.1.1 shows a typical scene on the Berkeley campus of the University of California: some students walking through Sather Gate, with trees and the distinctive Campanile bell tower in the background. We perceive all this so quickly and effortlessly that it is hard to imagine there being anything very complicated about it. Yet, when viewed critically as an ability that must be explained, visual perception is so incredibly complex that it seems almost a miracle that we can do it at all.

The rich fabric of visual experience that results from viewing natural scenes like the one in Figure 1.1.1 arises when the neural tissues at the back of the eyes are stimulated by a two-dimensional pattern of light that includes only bits and pieces of the objects being perceived. Most of the Campanile, for example, is hidden behind the trees, and parts of the trees are occluded by the towers of the gate. We don’t perceive the Campanile as floating in the air or the trees as having tower-shaped holes cut in them where we cannot currently see them. Even objects that seem to be fully visible, such as the gate towers and the students, can be seen only in part because their far sides are occluded by their near sides. How, then, are we able so quickly and effortlessly to perceive the meaningful, coherent, three-dimensional scene that we obviously do experience from the incompletely two-dimensional pattern of light that enters our eyes?

This is the fundamental question of vision, and the rest of this book is an extended inquiry into its answer. The modern study of vision certainly fits this inductive mold. It is rapidly becoming a tightly interdisciplinary science, neuropsychology, linguistics, and cognitive anthropology—all of which are part of the emerging field of cognitive science. The premise of cognitive science is that the problems of cognition will be solved more quickly and completely by attacking them from as many perspectives as possible.

The modern study of vision certainly fits this interdisciplinary mold. It is rapidly becoming a tightly integrated field at the intersection of many related disciplines, each of which provides different pieces of the jigsaw puzzle. This interdisciplinary field, which I will call vision science, is part of cognitive science.

In this book, I try to convey a sense of the excitement that it is generating among the scientists who study vision and of the promise that it holds for reaching a new understanding about how we see.

In this initial chapter, I will set the stage for the rest of the book by providing an introductory framework for understanding vision in terms of three domains:

1. phenomena of visual perception,
2. the nature of optical information, and
3. the physiology of the visual nervous system.

The view presented in this book is that an understanding of all three domains and the relations among them is required to explain vision. In the first section of this chapter, we will consider the nature of visual perception itself from an evolutionary perspective, asking what it is for. We will define it, talk about some of its most salient properties, and examine its usefulness in coupling organisms to their environments for survival. Next, we will consider the nature of optical information, because all vision ultimately rests on the structure of light reflected into the eyes from surfaces in the environment. Finally, we will describe the physiology of the part of the nervous system that underlies our ability to see. The eyes are important, to be sure, but just as crucial are huge portions of the brain, much of which vision scientists are only beginning to understand. In each domain, the coverage in this introductory chapter will be rudimentary and incomplete. But it is important to realize from the very beginning that only by understanding all three domains and the relations among them can we achieve a full and satisfying scientific explanation of what it means to see. What we learn here forms the scaffold onto which we can fit the more detailed presentations in later chapters.

1.1 Visual Perception

Until now, I have been taking for granted that you know what I mean by “visual perception.” I do so in large part because I assume that you are reading the words on this page using your own eyes and therefore know what
The knowledge achieved by visual perception concerns objects and events in the environment. Perception is not merely about an observer's subjective visual experiences, because we would not say that even highly detailed hallucinations or visual images would count as visual perception. We will, in fact, be very interested in the nature of people's subjective experience—particularly in Chapter 13 when we discuss visual awareness in detail—but it is part of visual perception only when it signifies something about the nature of external reality.

3. Visual knowledge about the environment is obtained by extracting information. This aspect of our definition implies a certain "metatheoretical" approach to understanding visual perception and cognition, one that is based on the concept of information and how it is processed. We will discuss this information processing approach in more detail in Chapter 13, but now suffice it is to say that it is an approach that allows vision scientists to talk about how people see in the same terms as they talk about how computers might be programmed to see. Again, we will have more to say about the prospects for sighted computers in Chapter 13 when we discuss the problem of visual awareness.

4. The information that is processed in visual perception comes from the light that is emitted by or reflected off objects. Optimal information is the foundation of all vision. It results from the way in which physical surfaces interact with light in the environment. Because this re-structuring of light determines what information about objects is available for vision in the first place, it is the appropriate starting point for any systematic analysis of vision (Gibson, 1950). As we will see in Section 1.2, most of the early problems in understanding vision arise from the difficulty of undoing what happens when light projects from a three-dimensional world onto the two-dimensional surfaces at the back of the eyes. The study of what information is contained in these projected images is therefore an important frontier of research in vision science, one that computational theorists are constantly exploring to find new sources of information that vision might employ.

1.1.2 The Evolutionary Utility of Vision

Now that we have considered what visual perception is, we should ask what it is for. Given its biological importance to a wide variety of animals, the answer must be that vision evolved to aid in the survival and successful reproduction of organisms. Desirable objects and situations—such as nourishing food, protective shelter, and desirable mates—must be sought out and approached. Dangerous objects and situations—such as precipitous drops, falling objects, and hungry or angry predators—must be avoided or fled from. Thus, to behave evolutionarily and appropriately, an organism must possess a sensory system capable of forming an image from the light it receives. We will see in Chapter 13 that such an image can be formed by the visual nervous system. Figure 1.1.3 shows an example called a striped grating. First, examine the two stripes on the right to convince yourself that they are vertical and identical to each other. Then look by fixating on the bar between them and moving your eyes back and forth along the bar. The top bar will look tilted to the right, and the bottom one looks tilted to the left. This is because your visual system was not sensitive enough.

Another everyday example of dark adaptation arises in gazing at stars. When you leave a brightly lit room to go outside on a cloudless night, the stars at first may look at for the next few minutes, altering your subsequent visual experiences so that you see something that is not there. Clearly, this is not veridical perception because the afterimage lasts long after the physical flash is gone. Not all aftereffects make you see things that are not there; others cause you to misperceive properties of visible objects. Figure 1.1.3 shows an example called an orientation aftereffect. First, examine the two striped gratings on the right to convince yourself that they are vertical and identical to each other. Then look at the two tilted gratings on the left for about a minute by fixating on the bar between them and moving your eyes back and forth along the bar. The top grating now looks tilted to the left, and the bottom one looks tilted to the right. These errors in perception are further evidence that what you see results from an interaction between the external world and the present state of your visual nervous system.

1It may be confusing that during dark adaptation the visual system becomes less sensitive to light. This apparent difference from other forms of adaptation can be eliminated if you realize that during dark adaptation the visual system is, in a sense, becoming less sensitive to the dark.
have probably noticed that the moon looks much larger when it is close to the horizon than it does when it is high in the night sky. Have you ever thought about why?

Many people think that it is due to refractive distortions introduced by the atmosphere. Others suppose that it is due to the shape of the moon’s orbit. In fact, the optical size of the moon is entirely constant throughout its journey across the sky. You can demonstrate this by taking a series of photographs as the moon rises; the size of its photographic image will not change in the slightest. It is only our perception of the moon’s size that changes. In this respect, it is indeed an illusion—a non-veridical perception—because its image in our eyes does not change size any more than it does in the photographs. In Chapter 7, we will discuss in detail why the moon illusion occurs (Kaufman & Rock, 1962; Rock & Kaufman, 1962). For right now, the important thing is just to realize that our perception of the apparent difference in the moon’s size at different heights in the night sky is illusory.

There are many other illusions demonstrating that visual perception is less than entirely accurate. Some of these are illustrated in Figure 1.1.4. The two arrow shafts in A are actually equal in length; the horizontal lines in B are actually the same size; the long lines in C are actually vertical and parallel; the diagonal lines in D are actually collinear; and the two central circles in E are actually equal in size.

**Figure 1.1.4** Visual illusions. Although they do not appear to be so, the two arrow shafts are the same length in A, the horizontal lines are identical in B, the long lines are vertical in C, the diagonal lines are collinear in D, and the middle circles are equal in size in E.

Once the lesson of illusions has been learned, it is easier to see that there is really no good reason why perception should be a clear window onto reality. The objects that we so effortlessly perceive are not the direct cause of our perceptions. Rather, perceptions are caused by the two-dimensional patterns of light that stimulate our eyes. (To demonstrate the truth of this assertion, just close your eyes. The objects are still present, but they no longer give rise to visual experiences.) To provide us with information about the three-dimensional environment, vision must therefore be an **interpretive process** that somehow transforms complex, moving, two-dimensional patterns of light at the back of the eyes into stable perceptions of three-dimensional objects in three-dimensional space. We must therefore conclude that the objects we perceive are actually interpretations based on the structure of images rather than direct registrations of physical reality.

**Ambiguous Figures.** Potent demonstrations of the interpretive nature of vision come from ambiguous figures: single images that can give rise to two or more distinct perceptions. Several compelling examples are shown in Figure 1.1.5. The vase/faces figure in part A can be perceived either as a white vase on a black background (A1) or as two black faces in silhouette against a white background (A2). The Necker cube in Figure 1.1.5B can be perceived as a cube in two different orientations relative to the viewer: with the observer looking down and to the right at the cube (B1) or looking up and to the left (B2). When the percept “reverses,” the interpretation of the depth relations among the lines changes; front edges become back ones, and back edges become front ones. A somewhat different kind of ambiguity is illustrated in Figure 1.1.5C. This drawing can be seen either as a duck facing left (C1) or as a rabbit facing right (C2). The interpretation of lines again shifts from one percept to the other, but this time the change is from one body part to another: The duck’s bill becomes the rabbit’s ears, and a bump on the back of the duck’s head becomes the rabbit’s nose.

There are two important things to notice about your perception of these ambiguous figures as you look at them. First, the interpretations are mutally exclusive. That
is, you perceive just one of them at a time: a duck or a rabbit, not both. This is consistent with the idea that perception involves the construction of an interpretive model because only one such model can be fit to the sensory data at one time. Second, once you have seen both interpretations, they are multistable perceptions, that is, dynamic perceptions in which the two possibilities alternate back and forth as you continue to look at them. This suggests that the two models compete with each other in some sense, with the winner eventually getting "tired out" so that the loser gains the advantage. These phenomena can be modeled in neural network theories that capture some of the biological properties of neural circuits, as we will see in Chapter 6.

1.1.4 Perception as Modeling the Environment

Ambiguous figures demonstrate the constructive nature of perception because they show that perceivers interpret visual stimulation and that more than one interpretation is sometimes possible. If perception were completely determined by the light stimulating the eye, there would be no ambiguous figures because each pattern of stimulation would map onto a unique percept. This position is obviously incorrect. Something more complex and creative is occurring in vision, going beyond the information strictly given in the light that stimulates our eyes (Bruner, 1978).

But how does vision go beyond the optical information, and why? The currently favored answer is that the observer is constructing a model of what environmental situation might have produced the observed pattern of sensory stimulation. The important and challenging idea here is that people's perceptions actually correspond to the models that their visual systems have constructed rather than (or in addition to) the sensory stimulation on which the models are based. That is why perceptions can be illusory and ambiguous despite the nonillusory and unambiguous status of the raw optical images on which they are based. Sometimes we construct the wrong model, and sometimes we construct two or more models that are equally plausible, given the available information.

The view that the purpose of the visual system is to construct models of the environment was initially set forth by the brilliant German scientist Hermann von Helmholtz in the latter half of the 1800s. He viewed perception as the process of inferring the most likely environmental situation given the pattern of visual stimulation (Helmholtz, 1867/1925). This view has been the dominant framework for understanding vision for more than a century, although it has been extended and elaborated by later theorists, such as Richard Gregory (1970), David Marr (1982), and Irvin Rock (1983), in ways that we will discuss throughout this book.

Care must be taken not to misunderstand the notion that visual perception is based on constructing models. Invoking the concept of models does not imply that perception is "pure fiction." If it were, it would not fulfill the evolutionary demand for accurate information about the environment. To satisfy this requirement, perceptual models must (a) be closely coupled to the information in the projected image of the world and (b) provide reasonably accurate interpretations of this information. Illusions show that our models are sometimes inaccurate, and ambiguous figures show that they are sometimes not unique, but both tend to occur only under unusual conditions such as in the books and laboratories of vision scientists. Everyday experience tells us that our perceptual models are usually both accurate and unique. Indeed, if the sensory information is rich and complex enough, it is nearly impossible to fool the visual system into interpreting the environment incorrectly (Gibson, 1966).

Visual Completion. Perhaps the clearest and most convincing evidence that visual perception involves the construction of environmental models comes from the fact that our perceptions include portions of surfaces that we cannot actually see. Look at the shapes depicted in Figure 1.1.6A. No doubt you perceive a collection of three simple geometrical figures: a square, a circle, and a long rectangle. Now consider carefully how this description relates to what is actually present in the image. The circle is partly occluded by the square, so its lower left portion is absent from the image, and only the ends of the rectangle are directly visible, the middle being hidden (or occluded) behind the square and circle. Nevertheless, you perceive the partial circle as complete and the two ends of the rectangle as parts of a single, continuous object. In case you doubt this, compare this perception with that of Figure 1.1.6B, in which exactly the same regions are present but not in a configuration that allows them to be completed.

This perceptual filling in of parts of objects that are hidden from view is called visual completion. It happens automatically and effortlessly whenever you perceive the environment. Take a moment to look at your present surroundings and notice how much of what you "see" is actually based on completion of unseen or partly seen surfaces. Almost nothing is visible in its entirety, yet almost everything is perceived as whole and complete.

You may have noticed in considering the incompleteness of the sensory information about your present environment that visual perception also includes information about self-occluded surfaces: those surfaces of an object that are entirely hidden from view by its own visible surfaces. For example, only half of the cube that you perceive so clearly in Figure 1.1.7A is visible. Your perception somehow manages to include those hidden surfaces that are occluded by the three visible ones. You would be more than a little surprised if you changed your viewpoint by walking to the other side and saw that the cube now appeared as in Figure 1.1.7B. Indeed, there are infinitely many possible physical situations that are consistent with Figure 1.1.7A, yet you automatically perceived just one: a whole cube.

Completion presents an even more compelling case for the model-constructive view of visual perception than do illusions and ambiguous figures. It shows that what you perceive actually goes a good deal beyond what is directly available in the light reaching your eyes. You have very strong expectations about what self-occluded and partly occluded surfaces are like. These must be constructed from something more than the light entering your eyes, because the image itself contains no direct stimulation corresponding to these perceived, but unseen, parts of the world.

Impossible Objects. There is another phenomenon that offers an especially clear demonstration of the modeling aspect of visual perception. Impossible objects are two-dimensional line drawings that initially give the clear perception of coherent three-dimensional objects but are physically impossible. Figure 1.1.8 shows some famous examples. The "butterfly" in Figure 1.1.8A looks sensible enough at first glance, but on closer inspection, it becomes clear that such an object cannot exist because the three round prongs on the left end do not match up with the two square ones on the right end. Similarly, the continuous three-dimensional triangle that we initially perceive in Figure 1.1.8B cannot exist because the surfaces of the locally interpretable sides do not match up properly (Penrose & Penrose, 1958).
One of the most interesting things about impossible objects is how clearly they show that our perceptions are internal constructions of a hypothesized external reality. If visual perception were merely an infallible reflection of the world, a physically impossible object simply could not be perceived. It would be impossible perceptually as it is physically. Yet people readily perceive such objects when viewing properly constructed images. This fact suggests that perception must be performing an interpretation of visual information in terms of the three-dimensional (3-D) objects in the environment that might have given rise to the images registered by our eyes. Moreover, the kinds of errors that are evident in perceiving impossible objects seem to indicate that at least some visual processes work initially at a local level and only later fit the results into a global framework. The objects in Figure 1.1.8 actually make good sense locally; it is only in trying to put these local pieces together more globally that the inconsistencies become evident.

Predicting the Future. Supposing that the visual system does construct hypothetical models of reality rather than just sticking to information available in sensory stimulation, why might such a system have evolved? At some level, the answer must be that the models are more useful from an evolutionary standpoint than the images that gave rise to them, but the reason for this is not entirely clear. The usefulness of visual completion, for example, would seem to be that 3-D models representing hidden surfaces contain much more comprehensive information about the world than purely stimulus-based perceptions. The additional information in the constructed model is valuable because it helps the perceiving organism to predict the future. We have already considered one example in our discussion of Figure 1.1.7. Perceiving a whole three-dimensional cube provides the basis for expecting what we would see if we were to move to see that new surfaces come into view. This is terribly important for creatures (like us) who are constantly on the move. A stable three-dimensional model frees us from having to recreate everything from scratch as we move about in the world.

A perceptual model of the three-dimensional environment does not need to be modified much as we move around because the only thing that changes is our viewpoint relative to a largely stable landscape of objects and surfaces. In fact, the only time the model needs major modification is when model-based expectations are disconfirmed by unexpected surfaces that enter into view. Every day experience tells us that this does not happen nearly as often as confirmation of our expectations. Thus, although constructing a three-dimensional model of the environment may initially seem like a poor evolutionary strategy, its short-term costs appear to be outweighed by its long-term benefits. It takes more time and effort to construct the complete model initially, but once it is done, it requires far less time and effort to maintain it. In the final analysis, the completed model is a remarkably economical solution to the problem of how to achieve stable and accurate knowledge of the environment.

The ability to predict the perceptual future is also evolutionarily crucial because we live in a world that includes moving objects and other mobile creatures. It is useful to know the current position of a moving object, but it is far more useful to know its direction and speed so that you can predict its future trajectory. This is particularly important when something is coming toward you, because you need to decide whether to approach, sidestep, flee, or ignore it. Without a perceptual model that somehow transcends momentary stimulus information, vision would not be able to guide our actions appropriately.

The view that the purpose of the brain is to compute dynamic, predictive models of the environment was set forth by British psychologist Kenneth Craik in 1943. He argued forcefully that organisms that can rapidly extrapolate the present situation into the future have an evolutionary advantage over otherwise identical organisms that cannot. An organism that can predict accurately is able to plan future actions, whereas one that cannot predict can only react once something has happened. There is an important caveat here, however: The process of extrapolation must work faster than the prediction itself. Not surprisingly, then, most perceptual predictions are generated very quickly. Indeed, they are usually generated so quickly that we have no conscious experience of them unless they are violated. Even then, our conscious experience reflects the violation rather than the expectation itself.

1.1.5 Perception as Apprehension of Meaning

Our perceptual constructions of the external world go even further than completing unseen surfaces in a three-dimensional model, however. They include information about the meaning or functional significance of objects and situations. We perceive an object not just as having a particular shape and being in a particular location, but as a person, a dog, a house, or whatever. Being able to classify (or recognize or identify) objects as members of known categories allows us to respond to them in appropriate ways because it gives us access to vast amounts of information that we have stored from previous experiences with similar objects.

Classification. Perhaps the easiest way to understand the importance of classification is to imagine encountering some completely foreign object. You could perceive its physical characteristics, such as its color, texture, size, shape, and location, but you wouldn't know what it was or what you should do with it. Is it alive? Can it be eaten? Is it dangerous? Should you approach it? Should you avoid it? Such questions can seldom be answered directly from an object's physical characteristics, for they also depend on what kind of object it is. We embrace loved ones, flee angry dogs, walk around pillars, eat hamburgers, and sit in chairs. All this is obvious that it scarcely seems worth mentioning, but without perceptually classifying things into known categories, it would be difficult to behave appropriately with the enormous variety of new objects that we encounter daily. We can simply walk around the pillar because past experience informs us that such objects do not generally move. But angry dogs can and do!

Classification is useful because objects within the same category share so many properties and behaviors. All chairs are not exactly alike, nor are all hamburgers, but one chair is a lot more like another than it is like any hamburger, and vice versa. Previous experience with members of a given category therefore allows us to predict with reasonable certainty what new members of that same class will do. As a consequence, we can deal with most new objects at the most abstract level of their category, even though we have never seen that particular object before.

Classifying objects as members of known categories seems simple, but it is actually quite an achievement. Consider the wide variety of dogs shown in Figure 1.1.9, for example. How can we recognize almost immediately that they are all dogs? Do dogs have some unique set of properties that enable us to recognize them? If so, what might they be? These are problems of object identification, one of the most difficult—and as yet unsolved—puzzles of visual perception. In Chapters 8 and 9, we will consider some current ideas about how this might happen.

Attention and Consciousness. It is an undeniable fact that the visible environment contains much more information than anyone can fully perceive. You must therefore be selective in what you attend to, and what you select will depend a great deal on your needs, goals, plans, and desires. Although there is certainly an important sense in which a hamburger is always a hamburger, how you react to one depends a great deal on whether you have just finished eating a two-course meal. After fasting, your attention would undoubtedly be drawn immediately to the hamburger; right after a big Thanksgiving dinner, you would probably ignore it, and if you did not, the sight of it might literally nauseate you. This example demonstrates that perception is not an entirely stimulus-driven process; that is, perceptions...
depending on what we are trying to accomplish, and we may perceive them differently as a result. This point is perhaps so obvious that it goes without saying, but it is important nevertheless. One of the functions of attention is to bring visual information to consciousness. Certain properties of objects do not seem to be experienced consciously unless they are attended, yet unattended objects are often processed fully enough outside of consciousness to attract your attention. Everyday examples abound. You may initially not notice a stationary object in your visual periphery, but if it suddenly starts moving toward you, you look in its direction without knowing why, only then becoming consciously aware of its presence. While driving your car, you sometimes look over at the car next to you, without knowing why, only to find that the driver has been looking at you. In both cases, visual processing has taken place outside of consciousness, directing your attention to the interesting or important aspects of the environment: the moving object or the person looking at you. Once the object is attended, you become conscious of its detailed properties and are able to identify it and discern its meaning in the present situation. Attending to an object visually usually means moving your eyes to fixate on it, but attention and visual fixation are not the same. You are probably familiar with the fact that you can be looking directly at something without attending to it in the slightest. Your thoughts may wander to some completely different topic, and once attention returns to the visual information, you may realize that you had no awareness of what was in your visual field during the diversion. Conversely, you can attend to an object without fixating on it. To demonstrate this, hold your hand out in front of you and fixate directly on your middle finger. Now, without moving your eyes, try attending to each of the other fingers in turn. It is not terribly easy, because you want to move your eyes at the same time as you shift your attention, but it clearly can be done. Many high-level aspects of perception seem to be fully conscious. For example, when you look around the room trying to find your keys, you are certainly aware of the key-finding goal that directs your attention to various likely places in the room. Other aspects of perception are clearly not conscious, even in the same situation, such as knowing what makes an object "keylike" enough to direct your eyes at it during this visual search. In general, lower levels of perception do not seem to be accessible to, or modifiable by, conscious knowledge and expectations, whereas higher levels do.

Not much is yet known about the role of consciousness in perception. Indeed, we know surprisingly little even about the evolutionary advantage of conscious perception. There is a general belief that there must be one, but nobody has yet managed to give a good account of what it is. The basic question is what advantage there might be for a consciously perceiving organism over one that can perform all the same perceptual tasks but without having conscious visual experiences. The unconscious automaton can, by definition, engage in all of the same evolutionarily useful activities—successfully finding food, shelter, and mates while avoiding cliffs, predators, and other dangerous objects—so it is unclear on what basis consciousness could be evolutionarily selected. One possibility is that the problem is ill-posed. Perhaps the automaton actually could not perform all the tasks that the consciously perceiving organism could. Perhaps consciousness plays some crucial and as-yet-unspecified role in our perceptual abilities. We will return to these conjectures in Chapter 13 when we consider what is known about the relation between consciousness and perception.

1.2 Optical Information
Our definition of visual perception implies that vision depends crucially on the interaction among three things: light, surfaces that reflect light, and the visual system of an observer that can detect light. Remove any one of these ingredients, and visual perception of the environment simply does not occur. It seems reasonable, therefore, to begin our study of vision by considering some basic facts about each of them. The present section will describe how light interacts with surfaces to produce the optical events that are the starting point of vision. The next section will describe the overall structure of the human visual system that processes information in these optical events. The remainder of the book discusses in detail how the visual system goes about extracting relevant information from the light to produce the useful perceptions of environmental scenes and events. I argued in the preceding section that the evolutionary role of visual perception is to provide an organism with accurate information about its environment. For this to happen, the light that enters our eyes must somehow carry information about the environment. It need not carry all the information we ultimately get from looking at things, but it must carry enough that the rest can be inferred with reasonable accuracy. In this section, we will consider how light manages to carry information about the world of visible objects around us.

1.2.1 The Behavior of Light
The science concerned with the behavior of light is a branch of physics called optics. According to prevailing physical theory, light consists of minute packets of energy called photons that behave like waves in some respects and like particles in others. Throughout most of this book we will be concerned mainly with the particle behavior of photons, although the discussion of color vision in Chapter 3 will require consideration of its wavelike properties as well. Photons radiate outward from their source—a hot body such as the sun, a fire, or the filament of an incandescent light bulb—like infinitely tiny bullets that travel through air in perfectly straight lines at the enormous speed of 186,000 miles per second. When photons strike the surface of an object, we say that it is illuminated. The amount of visible light—that is, the number of photons—falling on a given surface per unit of time is called its luminance. The luminance of a light covaries to some degree with its perceived brightness, but the relation is far from simple, as we will discover in Chapter 3.

Illumination. Illumination refers to the lighting conditions in the environment. The simplest condition from an analytical standpoint is called point-source illumination. It refers to an idealized situation in which all the light illuminating a scene comes from a single, point-sized light source at a specific location. A single incandescent light bulb in an otherwise dark space would be a reasonable approximation, as would the sun on a cloudless day. Point-source illumination produces dark, well-defined shadows behind illuminated surfaces and strong shading effects on the illuminated surfaces themselves, as illustrated in Figure 1.2A.1A. Both effects result from the fact that all of the direct (nonreflected) light is coming from a single location. In fact, one seldom encounters lighting conditions this simple in the real world. Point-source illumination is a generalization of the kind of lighting conditions that can exist in the real world.
source illumination is primarily of interest to vision theorists as a way to reduce the mathematical complexity of certain problems. It is used as a simplifying assumption, for example, in determining the shape of an object from the shading on its illuminated surfaces. If there are multiple point sources, such as a room with two or more incandescent lights, there are correspondingly two or more different shadows and shading patterns for each surface. Each additional light source thus complicates the optical structure of the environment.

In many real situations, light comes from diffuse illumination, in which light radiates from a relatively large region of space. To take an extreme example, the light of the sun on an overcast day is diffused almost uniformly over the entire sky, to nearly equal amounts of radiant light are coming from everywhere in the whole upper half of the visual environment. Under such conditions, both the shadow cast by illuminated surfaces and the shading on the surfaces themselves are much weaker and less well defined than under point-source illumination. If there are multiple point sources, such as a room with two or more incandescent lights, there are correspondingly two or more different shadows and shading patterns for each surface. Each additional light source thus complicates the optical structure of the environment.

Interaction with Surfaces. We said that photons travel in perfectly straight lines, but only until they strike the surface of an object. In almost every case, the surface produces a radical change in the behavior of the photons that strike it. It is these surface-induced changes in the behavior of photons that ultimately provide vision with information about the surfaces in the environment that produced them. The only surfaces that do not change the behavior of photons are completely transparent ones, and such surfaces would be literally invisible—if they existed. All real surfaces interact with light strongly enough under most conditions that they are visible to a vigilant observer.

When a photon strikes the surface of an object, one of three basic events takes place: It is either transmitted through the surface, absorbed by it, or reflected off it (see Figure 1.2.2). Transmitted light can either pass straight through the surface or be bent (refracted), as in Figure 1.2.3, which shows the lower part of a spoon in a glass of water as being displaced laterally. Of these photonsurface interactions, reflection is the most important for vision for two reasons. First, reflected light has been changed by its interaction with the surface, so it contains information about the surface. Second, reflected light is subsequently available to strike the receptive surface of an observer's eye, so it can transmit that information to a higher level of the nervous system. (See Figure 1.2.4.) Perfectly specular and perfectly matte surfaces are just the two idealized endpoints of a continuum, and all real surfaces fall somewhere in between. Figure 1.2.4 shows how different the same spherical shape looks with a highly specular surface compared to a highly matte surface. The reflectance properties of surfaces are actually even more complex than we have considered here because the degree of specularity can vary as a function of the angle of incident light, as is the case with semigloss surfaces. When viewed in a direction nearly parallel to the surface, they are much more specular than when viewed nearly perpendicular to it.

Let us now try to put together what we have said about light radiation from sources of illumination and light reflection by surfaces to come to a more complete understanding of the behavior of light in a real environment. All surfaces reflect some light except completely black ones (which absorb it all) and completely transparent ones (which transmit it all). And most surfaces are more matte than specular. Together, these facts imply that some light bounces in almost every direction off almost every surface in the environment. As a result, light does not come just from the direction of radiant light sources, such as the sun and light bulbs; it also comes by reflection from virtually every surface in the environment. Surfaces thus act as secondary light sources by illuminating other surfaces with their reflected light. Moreover, all of these photons are bouncing from source illumination to partially extended illumination does not (B).

![Figure 1.2.2](image1.png) Interaction between light and surfaces. A photon can be transmitted through a surface, absorbed by it, or reflected off it.

**Figure 1.2.3** Refraction of light. When light is transmitted through a transparent object, it can be bent (refracted), leading to erroneous perceptions, such as the misaligned appearance of the spoon handle in this glass of water.

The change that a surface produces in a reflected photon is to alter its trajectory. The photon bounces off the surface in a direction that depends on both the direction from which it came and the microscopic structure of the surface. If the surface is highly polished, or specular, such as a mirror, the light is reflected in the straight direction that is symmetric to the direction from which it came (see Figure 1.2.4B); the angle of incidence is equal to the angle of reflection. On the other hand, if the surface is dull or matte, such as a typical piece of paper, the light is scattered diffusely in many directions (see Figure 1.2.4A). Perfectly specular and perfectly matte surfaces are just the two idealized endpoints of a continuum, and all real surfaces fall somewhere in between. Figure 1.2.4 shows how different the same spherical shape looks with a highly specular surface compared to a highly matte surface. The reflectance properties of surfaces are actually even more complex than we have considered here because the degree of specularity can vary as a function of the angle of incident light, as is the case with semigloss surfaces. When viewed in a direction nearly perpendicular to the surface, they are much more specular than when viewed nearly parallel to it.
face to surface at enormous speed, often being reflected off many surfaces before finally being absorbed.

Color Plate 1.1 illustrates the sometimes profound effect of different levels of light reflection in the appearance of a bathroom scene. Image A was generated by a sophisticated computer program that shows only the direct illumination entering the eye from the light source, before any light from the fixture has been reflected from surfaces into the eye. Image B shows the same scene with the addition of primary reflections from nonluminous surfaces. Image C then adds secondary reflections, and image D adds further reflections up to the fifth-order. Notice that nonluminous surfaces do not appear in the image until the primary reflections are included, that the reflection in the mirror does not appear until secondary reflections are included, and that specular highlights on the shower door are not apparent until higher-order reflections are modeled. Thus, the rich appearance of natural scenes depends importantly on the complex interactions of ambient light with the structure of the physical environment.

The net effect of all these reflections is that light is reverberating around the environment, filling it with light from virtually every direction. This fact is of paramount importance for vision, because it is this complex optical structure that enables vision to occur. More light comes from some directions than others, and that is why we are able to see surfaces of different colors in different directions. Equal amounts of light from all directions, called a Ganzfeld, just looks like an all-encompassing gray fog.

The Ambient Optic Array. The pioneering perceptual psychologist James J. Gibson conceptualized the optical information available in light in terms of what he called the ambient optic array (or AOA). The AOA refers to the light converging at a given point of observation from all directions. It is called “ambient” because the observation point is literally surrounded by light converging on it from all directions. This means that if your eye were at this observation point, either light reflected from environmental surfaces or light emitted directly from the radiant source would be available from every direction. Vision is possible at that observation point because environmental surfaces structure the light in the AOA in complex, but lawful, ways. This lawfulness in the optical structure of the AOA provides the information that enables vision to occur.

When conceived in this way, vision can be likened to solving a puzzle. Surfaces in the world alter light by reflecting it in a way that forms the AOA at the current observation point with a particular complex structure. The visual system registers this structure and then tries, in effect, to reverse the process by determining the arrangement of surfaces that must exist in the environment to have structured the AOA in just that way. This all happens so quickly and effortlessly that we have no conscious knowledge about how it is done.

To appreciate more fully the nature of the AOA, let us consider a few examples. Figure 1.2.5A shows schematically the structure of the AOA from an observer’s perspective in a room containing a stool and a window that looks onto a tree. Notice that light comes from all directions toward the point of observation and that the AOA defined at this point exists independently of the observer occupying it. Figure 1.2.5B illustrates the relationship between an observer’s view of the world and the AOA: The eye samples a directional subset of the AOA. The shaded part of the AOA is not currently visible because the eye admires light only from the front. Figure 1.2.5C shows the resulting pattern of light that would be entering the observer’s eye at this observation point. It depicts the momentary optical image that falls on the light-sensitive cells at the back of the eye. This is the starting point of vision, at least for that instant in time.

It is important to realize that there is a different AOA at every point in the environment. Each one is unique, providing slightly different information about the environment. To illustrate this fact, consider what happens when this observer stands up. As he or she rises, the eye moves along a trajectory of different observation points, and visual stimulation is determined at each one by its unique AOA. Figure 1.2.5D depicts the final AOA of this event with solid lines, showing how certain parts of surfaces that were previously visible have become occluded and how other parts of surfaces that were previously occluded have now become visible. Figure 1.2.5E shows what optical image is entering the observer’s eye at this moment so that you can see how it has changed.

The changes in optical information caused by the change in observation point highlight a distinction of great importance between the momentary static AOA and the temporally extended, dynamic AOA. Whereas the static AOA can be characterized by a pattern of light converging at the observation point, the dynamic AOA can be fully characterized only by the optic flow of light over time. Thus, the dynamic AOA provides the observer with information from an additional dimension that unfolds over time. This turns out to be enormously important for many perceptual phenomena, such as our ability to perceive the third dimension of the environment (depth or distance from the observer), to determine the shapes of moving objects, and to perceive our own trajectory through the environment as we move. We will consider each of these topics more fully in Chapters 5 and 10 when we discuss perception of depth and motion in detail.

1.2.2 The Formation of Images

If vision is to provide accurate information about the external world, then there must be a consistent rela-
ship between the geometry of environmental surfaces and the light that enters the eye at a particular observation point. Indeed, there is. Figure 1.2.5 depicts the nature of this relationship by illustration. In this section, we will take a closer look at how optical images are formed on the back of the eye and how the laws of perspective projection describe this image formation process.

Optical Images. The situation involved in image formation is diagrammed in Figure 1.2.6. The external world has three spatial dimensions. Illumination bathes the objects in this three-dimensional space with light, and that light is reflected by surfaces into the observer's eye along straight lines (Figure 1.2.6A). Photons pass into the eye to form a two-dimensional, upside-down image on its back surface (Figure 1.2.6B and C). The object in the external world is often referred to as the distal stimulus (meaning distant from the observer), and its optical image at the back of the eye as the proximal stimulus (meaning close to the observer). As Figure 1.2.6A illustrates, the size of an object's image in the eye is usually specified by its visual angle: the number of degrees subtended by the image from its extremes to the focal point of the eye. It is important to understand that this angle measures the spatial dimensions of the object in environmental space, not the distal stimulus. The same external object will subtend a smaller angle when it is farther away and a larger angle when it is closer to the observer's eye. This relationship between object size, object distance, and image size is important in understanding the size-distance relation, the size at different distances, project images of different sizes, as illustrated in parts B and C.

Projective Geometry. The image formation process that maps the 3-D world to the 2-D image is highly lawful, and like most lawfulness, can be analyzed mathematically. The most appropriate mathematics for the task is projective geometry: the study of how a higher-dimensional space is mapped onto a lower-dimensional one. In the case of static vision, the projective mapping of interest is from the 3-D space of the environment onto the 2-D space of the image plane. Projective geometry can therefore specify for a given 3-D scene of objects exactly where each point in the scene will project onto a given 2-D image plane and what properties of these images will be invariant over different projections. In dynamic vision, the projection of interest is from the 4-D structure of space-time onto the 3-D space of optic flow that unfolds over time on the 2-D surface at the back of the eye.

Projective geometry thus seems to be the ideal mathematical tool for understanding image formation. The problem is that projective geometry alone cannot model the full complexity of optical phenomena because it does not contain the appropriate structure for modeling reflection, absorption, or refraction of light. In a world filled with opaque surfaces, for example, only light reflected from the closest surface in a particular direction will enter the eye. Photons from all farther points will be either absorbed or reflected by the next closer surface, thus preventing them from reaching the eye. The complications that are introduced by the interaction of light with surfaces make projective geometry only partly useful for modeling the process of forming 2-D optical images from the 3-D world. Rather, one must understand in more detail how light from the 3-D world forms an image when it is projected onto a 2-D plane.

Perspective and Orthographic Projection. One way to form a 2-D image of the 3-D world is to place a pinhole at a given distance in front of an image plane or other 2-D surface inside an otherwise lightproof box. Such a device is called a pinhole camera. Because photons travel in straight lines, the light falling on each point of the image plane of the pinhole camera got there by being reflected (or emitted) from a particular point in environmental space. That point lies along the ray starting at its image-plane point and passing through the pinhole (see Figure 1.2.7A). This situation gives the basic geometry of perspective projection (or polar projection): the process of image formation in which the light converges toward a single focal point (or pole). Good pinhole images are not as easy to create as the above description makes it seem. To get a crisp, clear image, the pinhole must be very small; about 0.4 mm in diameter is ideal. Because of this small aperture size, however, very little light falls on the image plane, so it must be observed under very dark conditions. If the hole is made larger to let more light in, the image becomes blurred because all the light no longer goes through the single point of the pinhole but through many different points (see Figure 1.2.7B). This problem can be overcome by supplying the camera with a transparent convex lens at its opening to bend the incoming light inward to a point (called its focal point) some distance behind the lens (see Figure 1.2.7C). Thus, the lens provides a "virtual pinhole" at its focal point that makes

* A good way to do this is to make a light-tight box with a single pinhole on one side and a translucent surface behind it. When the back surface is viewed under darkened conditions—as such as with a dark cloth covering one's head and the back of the box—you can observe the upside-down and backward image on the translucent surface.

Chapter 1

Introduction to Vision Science

Environmental Object (Distal Stimulus)

Camera

Projected Image (Proximal Stimulus)

A. Pinhole Aperture without Lens → Sharp Image

B. Large Aperture without Lens → Fuzzy Image

C. Large Aperture with Lens → Sharp Image

Figure 1.2.7 The optics of pinhole cameras and lenses. (A) A pinhole camera with a small aperture produces sharp images without a lens. (B) A camera with a larger aperture but no lens produces fuzzy, out-of-focus images. (C) A camera with a large aperture and a lens can produce clear, well-focused images if the focal length of the lens is appropriate for the distance to the imaging surface.

the projected image at the back sharp and clear again, like the pinhole camera's image, only brighter because more light comes through the larger opening. As we will see in Section 1.3, the human eye contains such a lens whose job is to focus the image on the back surface of the eye.

Assuming that the complicating effects of light-surface interaction can be incorporated into the model, the mathematics of perspective (or polar) projection are useful for modeling image formation by the human eye (see Figure 1.2.8A and B). Unfortunately, they are also

Chapter 1

Introduction to Vision Science
rather complex, much more so than has been alluded to here. To simplify matters, visual theorists often employ orthographic projection (or parallel projection) instead of perspective projection to model the geometry of image formation (see Figure 1.2.8C). In this case, the image is conceptualized as being formed by light rays that travel parallel to each other and perpendicular to the image plane, rather than rays that converge at the pinhole.

The mathematical simplification that results from orthographic projection is that the depth dimension of the world—distances from image to objects—is simply ignored, whereas all spatial information in the plane perpendicular to the viewing direction is preserved without change. This means that when the distance from the image to the object is large relative to the depth of the object (Figure 1.2.8B), orthographic projection is a good approximation of perspective projection. Close up, however, the differences between orthographic and perspective projection become quite significant, as illustrated in Figure 1.2.8A.

One way of understanding the relationship between perspective and orthogonal projection is to consider what happens when an object is moved farther and farther away from the pinhole of perspective projection (compare Figure 1.2.8A and Figure 1.2.8B). As this happens, the light rays projecting through the pinhole become more and more parallel so that, at an infinite distance, the light rays would be parallel, as they are in orthographic projection (Figure 1.2.8C). Thus, orthographic projection can be conceived as a limiting case of perspective projection, at which the distance between object and focal point is infinite. The important difference is that the perspective image of an object at an infinite distance is a single point, unlike its true orthographic projection which results in a spatially extended image.

1.2.3 Vision as an “Inverse” Problem

We have now described how light reflected from the 3-D world produces 2-D images at the back of the eye where vision begins. This process of image formation is completely determined by the laws of optics, so for any given scene with well-specified lighting conditions and a point of observation, we can determine with great accuracy what image would be produced. In fact, the field of computer graphics is concerned with exactly this problem: how to render images on a computer display screen that realistically depict scenes of objects by modeling the process of image formation. Many of the problems in this domain are now very well understood, as one can appreciate by examining some examples of state-of-the-art computer images that have been generated without recourse to any real optical processes whatsoever. The images in Color Plate 1.1, for example, were rendered by a ray-tracing algorithm that simulates image formation from an internal model of the surfaces in the room and the behavior of the light that illuminates them. In effect, the program simulates the optical events of photon emission, reflection, transmission, and absorption to construct an image of a “virtual” environment that does not exist in the physical world. Such programs allow the effects of different orders of light reflection to be illustrated (e.g., in Color Plate 1.1A-D) because the program can be stopped after each cycle of simulated reflection to see what the image looks like. This is not possible with real optical image formation.

The early stages of visual perception can be viewed as trying to solve what is often called the inverse problem: how to get from optical images of scenes back to knowledge of the objects that gave rise to them. From this perspective, the most obvious solution is for vision to try to invert the process of image formation by undoing the optical transformations that happen during image formation.

Unfortunately, there is no easy way to do this. The difficulty is that the mathematical relation between the environment and its projective image is not symmetrical. The projection from environment to image goes from three dimensions to two and so is a well-defined function: Each point in the environment maps into a unique point in the image. The inverse mapping from image to environment goes from two dimensions to three, and

this is not a well-defined function: Each point in the image could map into an infinite number of points in the environment. Therefore, logic dictates that for every 2-D image on the back of our eyes, there are infinitely many distinct 3-D environments that could have given rise to it.

Figure 1.2.9 illustrates the indeterminacy of inverse projection by showing that a single line segment in an optical image could have resulted from the projection of an infinite number of lines in the environment. The reason is that the inverse problem is underspecified (or under-constrained or underdetermined) by the sensory data in the image. There is no easy way around this problem, and that is why visual perception is so complex. In fact, were it not for the fact that our brains manage to come up with the correct solution most of the time, it would be tempting to conclude that 3-D visual perception is simply impossible!

We know that 3-D perception is possible precisely because the human visual system manages to do it with such remarkable accuracy under most circumstances. How does it solve this seemingly insoluble problem? Different theorists have taken different approaches, as we shall see in Chapter 5, but the dominant one is to assume that 3-D perception results from the visual system making a lot of highly plausible assumptions about the nature of the environment and the conditions under which it is viewed. These assumptions constrain the inverse problem enough to make it solvable most of the time. If the assumptions are true, the resulting solution will be veridical. Vision is thus a heuristic process in which inferences are made about the most likely environmental condition that could have produced a given image. The process is heuristic because it makes

Figure 1.2.9 An illustration of inverse projection. A single line segment on the retina can be the projection of an infinite variety of lines in the environment.
use of inferential rules of thumb—based on the additional assumptions—that are not always valid and so will sometimes lead to erroneous conclusions, as in the case of perceptual illusions. Under most everyday circumstances, however, the assumptions are true, and so normal visual perception is highly veridical. We will encounter these additional assumptions throughout the book, particularly in Chapter 5 when we discuss depth perception in detail. Then we will be able to see in what sense they allow the visual system to go beyond the information given in the optical image so that the seemingly impossible inverse problem can be solved.

1.3 Visual Systems

We know that the inverse problem can be solved because the human visual system solves it—maybe not all the time and maybe not with perfect accuracy, but enough of the time and with sufficient accuracy to provide us with excellent information about the environment. Much of the rest of this book is concerned with what is currently known about how the visual system accomplishes this feat. To begin, we will now take a quick look at the overall structure of the part of the nervous system that is known to be involved in processing visual information. The description that we give here will be brief and superficial in many respects. This is intentional. Its purpose is merely to provide a scaffolding of background knowledge about the biological structure of the visual system so that it will be available for later discussions that complete the picture. Once we have mastered some of the basic facts about the "hardware" of this system—in its anatomy or physical structure—we can begin to ask better informed questions about what it does—in physiology, or biological function.

1.3.1 The Human Eye

Although it has been known since antiquity that eyes are the sensory organs of vision, the accurate understanding of how they work is a relatively recent achievement. The Greek philosopher Plato (427–347 B.C.) believed that an "inner fire" gave rise to rays that emanated from the eye toward perceived objects. Epicurus (341–270 B.C.) rejected this emanation theory, believing instead that tiny replicas of objects were somehow transmitted rapidly into the mind through the eyes. Galen (A.D. 130–200) later elaborated these ideas with physiological details, proposing that after the rays emanated from the eye, they interacted with the object and then returned to the eye. In the lens of the eye, he believed, these rays interacted with a "visual spirit" that flowed from the brain to the eye and back, bringing with it the replicas of perceived objects.

The modern era of physiological optics did not really begin until the brilliant Arabic philosopher Alhazen (A.D. 965–1040) hit upon the idea that the eye is like a pinhole camera, as we discussed in Section 1.2.2. The important insight that he achieved was that vision occurs when light from external sources is reflected from surfaces of objects and enters the eye. Even so, an accurate understanding of the optics of the eye required the invention and understanding of lenses. The noted astronomer Johannes Kepler (1571–1630) finally put these elements together into a reasonable approximation of the modern theory of physiological optics, discussed below.

Eye and Brain. Although Galen had many important facts about vision wrong, he was right in believing that both eyes and brain are essential (see Figure 1.3.1). We now know that optical information from the eyes is transmitted to the primary visual cortex in the occipital lobe at the back of the head, as shown in Figure 1.3.1. This information is then sent to many other visual centers located in the posterior temporal and parietal cortex, as illustrated in Figure 1.3.2. Some estimates put the percentage of the cortex involved with visual function at more than 50% in the macaque monkey (DeYoe & Van Essen, 1988; Van Essen et al., 1990), although it is probably slightly lower in humans. The complete visual system thus includes much of the brain as well as the eyes, and the whole eye-brain system must function properly for the organism to extract reliable information about the environment from the ambient optic array.

The eyes must collect and register information contained in light, and the brain must then process that information in ways that make it useful for the organism. The fact that both eye and brain are required for vision means that a person who has normal eyes but damage to visual parts of the brain might be as "blind" as a person who has a normal brain but is blind by work. Indeed, both sorts of blindness exist. Damage to the eyes from accidents or disease sometimes prevents them from doing their job of registering optical information and/or sending it on to the visual areas of the brain. Such conditions cause the sort of blindness most people know about: lack of sight because no information from light gets into the system. Although far less common, there are also people who cannot see—for at least do not have visual experiences—yet have eyes that work quite normally. This kind of blindness—called blindsight—results from damage to certain critical parts of the visual cortex due to disease, surgery, or stroke (see Section 13.2.2). Damage to other parts of visual cortex does not result in blindness but can cause debilitating selective deficits in perception. Some patients can see well enough to describe faces accurately but cannot identify even members of their own family by sight (see Section 9.2.5). Others can describe and visual information travels to other parts of the brain. A secondary pathway goes from the optic nerve to the superior colliculus and then to other brain centers. (From Rosenzweig & Leiman, 1982. An Introduction to Vision Science, 10.1.4. We will discuss these fascinating problems in more detail at appropriate places throughout the book, particularly at the end when we consider conscious experiences of seeing.

Anatomy of the Eye. There are some obvious anatomical facts about the eyes that almost everyone knows. Humans have two eyes, which are approximately spherical in shape except for a bulge at the front. Located at about the horizontal midline of the head, they sit in nearly hemispherical holes in the skull, called the eye sockets, that hold them securely in position yet allow
invalid, that is, conditions that are seldom or never present in normal everyday living conditions.\(^4\) He maintained that his theory of direct perception was correct but that it held only for unrestrained observers who were able to actively explore a natural environment. He argued, quite persuasively, that when psychologists require subjects to view a scene from a single static viewpoint, force them to wear unusual glasses to view a projection composed of light and dark regions, and so on, that both eyes are looking at the same environmental scene, that the environment is a three-dimensional world populated with objects rather than a two-dimensional picture that simulates the optical structure of a three-dimensional scene, they violate the conditions under which humankind evolved the ability to perceive in the first place. Perception under such artificial conditions, he therefore claimed, could not be used as evidence against his theory, which was a theory of aspatial perception.

Taken on its own terms, Gibson's defense of his position is both reasonable and persuasive. The conditions under which vision scientists often study perception are undeniably odd and unnatural in many ways. Yet the phenomena do occur and seem to require an explanation of some sort. What is most interesting about Gibson's defense for the present discussion, however, is the close relation that exists between the boundary conditions for ecological perception and the heuristic assumptions of computational theories.

Generally speaking, what Gibson calls ecological conditions are those in which the heuristic assumptions of inference-based theories are true: for example, that observers are not looking from some special vantage point, that both eyes are looking at the same environmental scene, that the environment is a three-dimensional world populated with objects rather than a two-dimensional projection composed of light and dark regions, and so on. Notice that if the boundary conditions of ecological perception are exactly the situations in which the heuristic assumptions of inference theories are true, ecological perception will always be veridical. Thus, Gibson was able to defend his theory of direct perception against evidence of nonveridicity by claiming that such situations were not ecological. This is why Gibson was so effective at deflecting criticisms based on the existence of illusions; he did not deny that illusions existed, but only that they occurred under natural ecological conditions.\(^7\)

The view that will be taken in this book is that perceptual theory must account for all phenomena of visual perception, whether ecologically valid or not. From this perspective, an approach based on heuristic computational inference is preferable to one based on direct perception because it can potentially give an explanation for both ecological and nonecological phenomena. We will therefore adopt the stance that perception involves some form of inductive inference and that it is carried out by performing computations in neural networks of some sort. In the course of our explorations of visual perception, we will attempt to identify the heuristic assumptions that underlie it, the processes that carry it out, and the neural mechanisms through which it occurs.

**Top-Down versus Bottom-Up Processes.** Another important distinction in the processing of perceptual information is its metaphorical "direction": whether it is *bottom-up* or *top-down*. The spatial metaphor underlying the naming of this distinction can be understood in terms of a flowchart of visual processing in which the retinal image is depicted at the bottom and temporally subsequent interpretations farther along the visual pathway are located at higher and higher levels (see Figure 2.3.11). "*Bottom-up*" processing—more descriptively called *data-driven processing*—refers to processes that take a "lower-level" representation as input and create or modify a "higher-level" representation as output. Top-down processing—also called *hypothesis-driven or expectation-driven processing*—refers to processes that operate in the opposite direction, taking a "higher-level" representation as input and producing or modifying a "lower-level" representation as output.

Many people's naive intuition is that vision is essentially a bottom-up process. It begins with the sensory information in the retinal images and goes "upward" to perceptual and then conceptual interpretations. Most theorists concur that the early stages of visual processing are indeed strictly bottom-up. But there are good reasons to think that this cannot be true for the entire process of visual perception.\(^5\) I argued in Section 1.1, for example, that perception of the present state of affairs produces expectations about the future. These expectations imply a top-down component to visual processing, because they suggest that prior higher-level interpretations influence current processing at lower levels. For example, it turns out that how people identify letters depends strongly on whether those letters are part of known words or meaningless letter strings. For this to occur, there must be top-down feedback from some higher-level representation of known words to the lower-level representation of letters. When we examine theories of perceptual categorization in Chapter 9, we will encounter further examples of the idea that top-down processing is involved in fitting stored models of familiar objects and scenes to incoming sensory data. The point at which top-down processes begin to augment bottom-up processes is currently a controversial issue, however. Some theorists believe that it happens early in visual processing; others believe that it happens late.

### 2.4 Four Stages of Visual Perception

With this general background in the information processing approach, we will now apply some of these concepts to vision. We will begin by decomposing visual perception at the algorithmic level into four major stages beyond the retinal image itself: *image-based, surface-based, object-based, and category-based* processing. (See text for details.)

\(^4\) Arguments about ecological validity were previously employed by Egon Brunswik (1955) to undermine the importance of results based on improbable stimulus situations. Gibson and his followers broadened the meaning of this concept in important ways, however, using it to rule out certain "everyday conditions" in which the stimulus is artificial, such as 2-D photographs and movies that portray 3-D scenes and events. Even though we may encounter such conditions frequently in our modern high-technology society, they do not correspond to the conditions under which human vision evolved.

\(^5\) Many simple geometric illusions are present even under natural, unrestrained viewing conditions; however, Deese and Hardwick (1991) report that the Müller-Lyer Illusion (see Figure 1.1.4A) persists under "gallery" conditions when observers can walk around models of the inward and outward arrows and view it from any angle.
whose center is in the middle of the fovea and whose $x$ and $y$ axes are aligned with retinaally defined horizontal and vertical, respectively. These square image elements—called pixels, a shortened form of "picture elements"—are taken as the primitive, indivisible, explicitly represented visual unit of information in the input image. The value of a given pixel in a gray-scale image is usually denoted $I(x,y)$ for the image "intensity" (or luminance) at the given location. These aspects of the image representation are illustrated in Figure 2.4.5 for a small portion of the scene in Figure 2.4.2. Figure 2.4.3A displays an enlargement of the small region inside the white box in Figure 2.4.2 to show the gray-scale levels of individual pixels. Figure 2.4.3B displays the numerical intensity values (from a potential range of 0 to 255) of the individual pixels shown in Figure 2.4.3A.

Regardless of whether such simplifying assumptions are made, the coordinate system of the retinal image is presumed to be explicitly tied to the intrinsic structure of the retina. The center of the retinal coordinate system is identified with the center of the fovea, and its axes are identified with retinaally defined horizontal and vertical. Receptor positions are specified relative to this retinal frame of reference.

To appreciate just how difficult the problem of interpreting the raw output of such an array of sensors or receptors is, consider again the display in Figure 2.4.3B. It indicates the light intensity falling on each receptor as a two-digit number. All of the spatial information of the gray-scale image in Figure 2.4.3A is present but in a numerical form that your visual system cannot interpret in terms of edges, regions, surfaces, and objects, and so forth. It appears completely meaningless and uninterpretable. This is not so when you look at Figure 2.4.2, however, which shows the whole image in gray-scale shadings from which these numbers were derived. Immediately, you perceive edges, regions, surfaces, and objects, the important aspects of the visual scene that were missing when you examined the numerical array.

The reason the numerical version is so difficult to comprehend is that your visual system is finely tuned to process information contained in intensity images and not at all well equipped to process numerical ones. Therefore, all of your edge detection, region-finding, surface location, and object interpretation processes were useless—except in reading the numbers themselves—and you were forced to search laboriously for every scrap of information you could find to get some notion of what might be depicted. When you consider the problem of trying to specify what information processes might be required for visual perception to occur, you should remember how difficult it was to interpret this numerical image, for this is the challenge that all perceptual theories ultimately face: perceiving objects arrayed in the three-dimensional environment just on the basis of a two-dimensional array of numbers. It is a daunting task indeed.

---

2.4.2 The Image-Based Stage

Most theorists currently agree that the initial registration of images in the two eyes is not the only representation based on a two-dimensional retinal organization. We call these additional representations and processes the image-based stage. It includes image-processing operations such as detecting local edges and lines, linking local edges and lines together more globally, matching up corresponding images in the left and right eyes, defining two-dimensional regions in the image, and detecting other image-based features, such as line terminations and "blobs." These two-dimensional features of images characterize their structure and organization before being interpreted as properties of three-dimensional scenes. For example, Figure 2.4.4A shows the cup. (B) The numerical array shows the intensity of the light falling on each square pixel element in the video image shown in part A.
indicates the locations of local edges that would constitute part of the image-based representation for the cup image shown in Figure 2.4.2.

Notice that the luminance edges that have been detected in Figure 2.4.4A are not exactly the same as the edges most people readily identify in the same image, as shown in Figure 2.4.4B. Many of the edges represented in part A are ones that people typically do not notice, either because they are faint or because they are due to differences in illumination (shadows and shading) rather than surface edges. Equally interesting is the fact that some of the most obvious edges that everyone perceives in the image are actually missing in the edge map of Figure 2.4.4A. The light side of the cup shades smoothly into the light table top on the left, and its dark side shades into the dark background and shadow on the right, with no discernible edge in the image to indicate their presence. Pieces of the top edges of the cup are similarly missing in the luminance edges. These simple demonstrations are dramatic proof that the set of luminance edges detected in an image (Figure 2.4.4A) are not the same as a clean line drawing of the objects people typically perceive (Figure 2.4.4B). Would that it were so simple!

Marr (1982) called the representations that resulted from such image-based processes primal sketches and suggested that there are two of them. The first he called the raw primal sketch, which includes just the results of elementary detection processes that locate edges, bars, holes, and line terminations. The second he termed the full primal sketch, which also includes global grouping and organization among the local image features present in the raw primal sketch. Marr's particular proposals about image-based processing may be incorrect—indeed, we will suggest a rather different conception at the end of Chapter 4—but the general idea of constructing an image-based representation of some sort is a useful one. Whatever it might be in detail, the common underlying structure of such an image-based representation is defined by the following properties:

1. **Image-based primitives.** The primitive elements represent information about the 2-D structure of the luminance image (such as edges and lines defined by differences in light intensity) rather than information about the physical objects in the external world that produced the image (such as surface edges or shadow edges). The two kinds of information are correlated, of course, but this correlation can only be used after the image features have been made explicit in the image-based representation.

2. **Two-dimensional geometry.** The geometry of spatial information in image-based representations is inherently two-dimensional and can be represented in the analog format of two-dimensional arrays.

3. **Retinal reference frame.** The coordinate system within which the 2-D features are located is specified relative to the retina in the sense that the principle axes are aligned with the eye (rather than the body, gravity, or the environment).

We will have a great deal more to say about how such image-based representations might be constructed in Chapter 4 when we discuss this topic in detail.

### 2.4.3 The Surface-Based Stage

The second stage of visual processing, which we will call the **surface-based stage**, is concerned with recovering the intrinsic properties of visible surfaces in the external world that might have produced the features that were discovered in the image-based stage. The fundamental difference is that the surface-based stage represents information about the external world in terms of the spatial layout of visible surfaces in three dimensions, whereas the image-based stage refers to image features in the two-dimensional pattern of light falling on the retina.

The notion that the visual system is fundamentally concerned with perceiving the **surface layout**—the spatial distribution of visible surfaces within the three-dimensional environment—was first proposed and strongly advocated by Gibson (1950). While almost every other theorist was talking about perceiving 3-D objects, Gibson realized that perceiving visible surfaces was a more basic task. His idea did not gain wide acceptance, however, until computer vision theorists began to advocate it fairly recently (e.g., Marr, 1978; Marr & Nishihara, 1978). Part of the difficulty was that Gibson never suggested a specific representation for this surface layout or a set of processes that could construct it from retinal images. The idea was not terribly surprising because before the advent of computer vision, there was not an information processing theorist and therefore did not believe in either representations or processing.

Still, he recognized more clearly than anyone else that surface perception was basic and crucial.

The concept of an explicit surface-based representation as an intermediate stage in vision became popular when it was formulated quantitatively by computer vision theorists and implemented in working computer simulations. Marr (1978) and Barrow and Tenenbaum (1978) proposed representations at about the same time and sketched algorithms that might be able to construct them from actual gray-scale images. Marr (1978) named his surface-based representation the **2.5-D surface**—the 2-D assumption that it lies somewhere between the true 2-D structure of image-based representations and the true 3-D structure of object-based representations (see below). Barrow and Tenenbaum (1978) called their surface-based representations **intrinsics images** to emphasize the fact that they represent intrinsic properties of surfaces in the external world rather than properties of the input image.

Constructing a surface-based representation is the first step in recovering the third spatial dimension from two-dimensional images. It does not contain information about all of the surfaces that are present in the environment, only about those that are visible from the current viewpoint. As we will see in Chapter 5, visible surfaces provide a great deal of sensory information about the distance of the objects from the viewer and their slant. They cannot be computed from the retinal images without additional assumptions because doing so is an underconstrained inverse problem. But the additional assumptions that are required to infer the properties of visible surfaces are relatively few and almost always true, especially compared with those needed to infer the same properties of surfaces that are hidden from view. Because the surface-based representation includes only the visible portions of surfaces, it can be conceived as a single, extremely flexible rubber sheet that has been "shrink wrapped" to just those surfaces in the environment that reflect light into the perceiver's eyes. Most current visual theories treat surfaces in this representation as being composed of many small, locally flat pieces. This is possible because even a strongly curved surface is nearly flat over a sufficiently small region, just as the spherical earth seems flat on the scale at which people move around on it. The simplification allows the surface-based representation to be specified completely by just information about the color, slant, and distance from the viewer of each locally flat patch of surface in each direction radially outward from the viewer's position. Figure 2.4.5 illustrates what such a surface-based representation would look like for the ceramic cup in Figure 2.4.2 by showing circles lying on the local surface patches and vectors sticking perpendicularly out of them at a sampling of locations, as though needles were sticking perpendicularly out of the patches of surface.

Again, Marr's particular conception of the 2.5-D sketch may be flawed, but some kind of surface-based representation seems necessary. Among the crucial properties of such a representation are the following:

1. **Surface primitives.** The primitive elements of the surface-based representation are local patches of 2-D surface at some particular slant located at some distance from the viewer within 3-D space. Each such patch of surface can be further specified by its color and texture.

2. **Three-dimensional geometry.** Although the surfaces themselves are locally only 2-D, their spatial distribution is represented within a 3-D space.

3. **Viewer-centered reference frame.** The coordinate system within which the 3-D layout of surfaces is represented is specified in terms of the direction and distance from the observer's stationpoint to the surface rather than in terms of the retina.

The flow diagram in Figure 2.4.6 indicates that the representation of surfaces is constructed from several similar processes, each contributing to the final representation.
The surfaces visible in Figure 2.4.2 are represented as a set of local estimates of surface orientation (slant and tilt) and depth of imaginary circles on the surface and velocity of points at various distances due to motion between the lateral position of objects in the images of the left and right eyes), motion parallax (differences in primarily with how the surface-based representation shape, and occlusion. We will have a great deal more to say about these factors in Chapter 2.

Figure 2.4.5 The surface-based representation of the cup scene. The surfaces visible in Figure 2.4.2 are represented as a set of local estimates of surface orientation (planar and slant) and depth with respect to the viewer. Surface orientation is depicted by a set of imaginary circles on the surface and "needles" pointing perpendicularly out of them at a sampling of image locations.

Figure 2.4.6 A flowchart showing how the surface-based representation might be derived from the image-based representation. These sources of information about depth and surface orientation are discussed in detail in Chapter 5.

2.4.4 The Object-Based Stage

Visual perception clearly does not end with a representation of just the surfaces that are visible. If it did, we should not be surprised were a change in viewpoint to reveal that the lower back side of the cup in Figure 2.4.2 simply did not exist or that it had some quite different shape from the smooth cylindrical one everyone perceives so effortlessly. But, as I argued in Chapter 1, either of these revelations would surprise us greatly. The fact that we have such expectations about partly and completely hidden surfaces suggests that there is some form of true three-dimensional representation that includes at least some occluded surfaces in the visual world. It is in this object-based stage that the visual representation includes truly three-dimensional information. For the visual system to manage this, further hidden assumptions about the nature of the visual world are required, because now the inferences include information about unseen surfaces or parts of surfaces. We will call this stage of processing object-based because the inclusion of these unseen surfaces implies that they involve explicit representations of whole objects in the environment. Figure 2.4.7 shows as dashed lines the hidden edges that everyone perceives in Figure 2.4.2. The table edge is occluded by the cup, and the back, inner sides, and bottom of the cup are occluded by the parts of the cup that we can actually see. Recovering the 3-D structure of these environmental objects is the goal of object-based processing.

There are at least two ways in which such an object-based representation might be constructed. One is simply to extend the surface-based representation to include unseen surfaces within a fully three-dimensional space. This might be called a boundary approach to object-based representation. The other is to conceive of objects as intrinsically three-dimensional entities, represented as arrangements of some set of primitive 3-D shapes. This might be called the volumetric approach, since it represents objects explicitly as volumes of a particular shape in three-dimensional space.

Figure 2.4.8 illustrates how a human body might be approximated by a hierarchy of parts, each of which is represented in terms of shape primitives based on cylindrical volumes. Influential work on 3-D shape primitives in computer vision by Agin and Binford (1976) and Marr and Nishihara (1978) caused the volumetric approach to dominate theories of object-based processing for many years. It is possible, of course, that some filling-in of occluded surfaces can take place in an intermediate stage before construction of a full volumetric representation. We will discuss in Chapters 6, 7, and 8 how object-based representations might be derived.

Once again, it is important to separate the details of Marr's particular version of an object-based representation in terms of generalized cylinders from the more abstract theoretical concept of a volumetric description. In this case, however, even the general nature of the representation is far from clear. In addition to the issue of whether the primitive elements are surfaces or volumes, there is much debate over the precise nature of the relevant reference frame and geometry. The current best guess is as follows:

1. **Volumetric primitives.** The primitive elements of the object-based representation may be described of truly three-dimensional volumes, thereby including information about unseen surfaces of the object.

2. **Three-dimensional geometry.** The space within which the volumetric primitives are located is also fully three-dimensional.

3. **Object-based reference frames.** The coordinate system within which the spatial relations among the volumetric primitives are represented may be defined in terms of the intrinsic structure of the volumes themselves. (The concept of intrinsic, object-based reference frames is complex and will be discussed in detail in Chapter 8.)

2.4.5 The Category-Based Stage

I argued in Chapter 1 that the ultimate goal of perception is to provide the perceiving organism with accurate information about the environment to aid in its survival and reproduction. This strongly implies that the final stage of perception must be concerned with recovering the functional properties of objects: what they afford the organism, given its current beliefs, desires, goals, and motives. We call this processing the category-based stage because it is widely believed that functional properties are accessed through a process of categorization.

The perception of Figure 2.4.2 as showing a cup is a result of the category-based processing of some type. But what type?

The categorization (or pattern recognition) approach to perceiving evolutionarily relevant function proposes that two operations are involved. First, the visual system classifies an object as being a member of one of a large number of known categories according to its visible properties, such as its shape, size, color, and location. Second, this identification allows access to a large body of stored information about this type of object, including its function and various forms of expectations about its future behavior. The object in Figure 2.4.2 is then known to be useful for containing liquids and for drinking out of them. This two-step scheme has the advantage that any functional property can be associated with any object, because the relation between the form of an object and the information stored about its function, history, and use can be purely arbitrary, owing to its mediation by the process of categorization.

There is also a very different way in which the visual system might be able to perceive an object's function, and that is by registering functional properties of objects more or less directly from their visible characteristics without first categorizing them. Gestalt theorists first suggested this approach, calling them **physiologistic characters.** Koella put it this way: "To primitive man, each thing says what it is and what he ought to do with it: a fruit says, 'Eat me'; water says, 'Drink me'; thunder says, 'Fear me'; and woman says, 'Love me.'" (Koella, 1935, p. 7). Gibson (1979) later advocated a similar approach, expanding his claim of direct perception to include function. He referred to the visible functions of an object as its **affordances** for the perceiver. Examples of affordances would be whether an object affords...
grasping by the observer’s hand or whether it affords sitting upon by the observer’s body. According to this view, one does not first have to classify something as a member of the category “chair” to know that one can sit on it because this affordance is directly perceivable without categorization.

It is possible—indeed, even likely—that people employ both types of processes (direct and indirect) in perceiving function. Some objects such as chairs and cups have functional properties that are so intimately tied to their visible structure that one might not need to categorize them to know what they can be used for. Other objects, such as computers and telephones, have functions that are so removed from their obvious visual characteristics that they almost certainly need to be categorized first. The extent to which people use each of these strategies to perceive functionally relevant information about objects is currently unknown. We will explore these and other issues involved in perceiving function more fully in Chapter 9.

These four proposed stages of visual processing—image-based, surface-based, object-based, and category-based—represent the current best guess about the overall structure of visual perception. We have listed them in the particular order in which they must logically be initiated, but that does not necessarily mean that each is completed before the next begins. The arrows going backward in Figure 2.4.1 indicate that later processes may feed back to influence earlier ones. In the coming chapters, we will use this four-stage framework to structure our discussion of each of the main topics that we consider: perception of color (Chapter 3), spatial properties (Chapters 4–9) and motion (Chapter 10). In the final three chapters, we will examine three related topics that surround this core of visual processing: visual selection by eye movements and attention in Chapter 11, visual memory and imagery in Chapter 12, and visual awareness in Chapter 13.

Suggestions for Further Reading

Classical Perceptual Theories

Koffka, K. (1935). Principles of Gestalt psychology. New York: Harcourt Brace. This is the most comprehensive statement of the Gestalt view of psychology. The vast majority of it concerns vision, but other topics are covered as well.


Rock, I. (1983). The logic of perception. Cambridge: MIT Press. This is the definitive modern statement of a neo-Helmholtzian constructivist position. In Rock’s hands, perception is likened to a problem-solving process in which the visual system must figure out what environmental situation gave rise to the optical image registered on the retina.

History of Information Processing


Information Processing Theory


Naive, V. S. (1993). A guided tour of computer vision. Reading, MA: Addison-Wesley. This is an excellent, general-purpose introduction to the field of computer vision.

Four Stages of Vision