1. Introduction to Vision Science

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. 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.

How, then, are we able so quickly and effortlessly to perceive the meaningful, coherent, three-dimensional scene that we obviously do experience from the incomplete, two-dimensional pattern of light that enters our eyes?

- Why do objects appear colored?

- How can we determine whether an object is large and distant or small and close?

- How do we perceive which regions in a visual image are parts of the same object?

- How do we know what the objects that we see are for?

- How can we tell whether we are moving relative to objects in the environment or they are moving relative to us?

- Do newborn babies see the world in the same way we do?

- Can people "see" without being aware of what they see?

Different parts of the answers come from a variety of different disciplines—biology, psychology, computer 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 we will call vision science, is part of cognitive science. In this course, 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.

1.1 Visual Perception

Now we will consider the various aspects of visual perception.

1.1.1 The Evolutionary Utility of Vision

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 in an evolutionarily adaptive manner, we must somehow get information about what objects are present in the world around us, where they are located, and what opportunities they afford us. All of the senses—seeing, hearing, touching, tasting, and smelling—participate in this endeavor.

Evolutionarily speaking, visual perception is useful only if it is reasonably accurate. Indeed, by and large, what you see is what you get. When this is true, we have what is called **veridical perception** (from the Latin veridicus meaning to say truthfully): perception that is consistent with the actual state of affairs in the environment. This is almost always the case with vision, and it is probably why we take vision so completely for granted. It seems like a perfectly clear window onto reality. But is it really?

In the remainder of this section, I will argue that perception is not a clear window onto reality, but an actively constructed, meaningful model of the environment.

1.1.2 Perception as a Constructive Act

Adaptation and Aftereffects

One kind of evidence that visual experience is not a clear window onto reality is provided by the fact that visual perception changes over time as it adapts to particular conditions. When you first enter a darkened movie theater on a bright afternoon, for instance, you cannot see much except the images on the screen. After just a few minutes, however, you can see the people seated near you, and after 20 minutes or so, you can see the whole theater surprisingly well. This increase in sensitivity to light is called **dark adaptation**. The theater walls and distant people were there all along; you just could not see them at first because your visual system was not sensitive enough.

When someone takes a picture of you with a flash, you first experience a blinding blaze of light. This is a veridical perception, but it is followed by a prolonged experience of a dark spot where you saw the initial flash. This **afterimage** is superimposed on whatever else you 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 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 gaze back and forth along it. Then look at the square between the two gratings on the right. 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.

Reality and Illusion

There are many other cases of systematically nonveridical perceptions, usually called illusions. One particularly striking example with which you may already be familiar is the **moon illusion**. You 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?

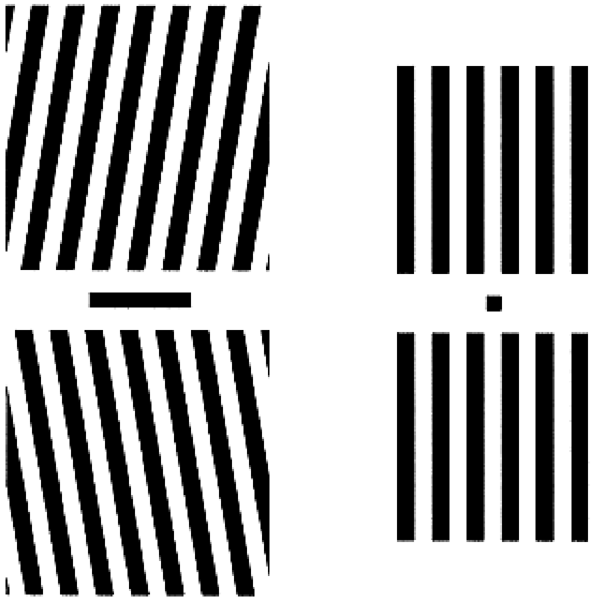

Figure 1.1 An orientation aftereffect. Run your eyes along the central bar between the gratings on the left for 30-60 seconds. Then look at the square between the two identical gratings on the right. The upper grating should now appear tilted to the left of vertical and the lower grating tilted to the right.

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 nonveridical perception—because its image in our eyes does not change size any more than it does in the photographs.

There are many other illusions demonstrating that visual perception is less than entirely accurate. Some of these are illustrated in Figure l.2. 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. In each case, our visual system is somehow fooled into making perceptual errors about seemingly obvious properties of simple line drawings. These illusions support the conclusion that perception is indeed fallible and therefore cannot be considered a clear window onto external reality. The reality that vision provides must therefore be, at least in part, a construction by the visual system that results from the way it processes information in light. As we shall see, the nature of this construction implies certain **hidden assumptions** , of which we have no conscious knowledge, and when these assumptions are untrue, illusions result.

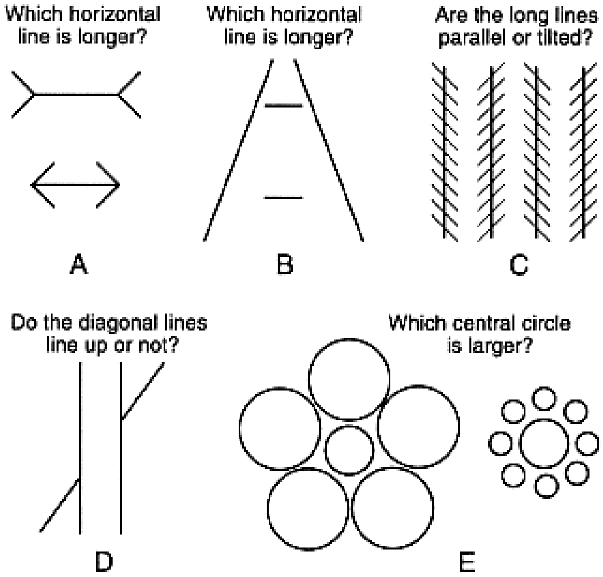
Which horizontal line is longer?
Which horizontal line is longer?
Are the long lines parallel or tilted?
A
B
C
Do the diagonal lines line up or not?
Which central circle is larger?
D
E

Figure 1.2 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.

It is easy to get so carried away by illusions that one starts to think of visual perception as grossly inaccurate and unreliable. This is a mistake. As we said earlier, vision is evolutionarily useful to the extent that it is accurate—or, rather, as accurate as it needs to be. Illusions are not terribly obvious in everyday life; they occur most frequently in books about perception. The important point for the present discussion is that the existence of illusions proves convincingly that perception is not just a simple registration of objective reality. There is a great deal more to it than that. 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 1.3. 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 1.3B 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 change; front edges become back ones, and back edges become front ones. A somewhat different kind of ambiguity is illustrated in Figure 1.3C. 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.

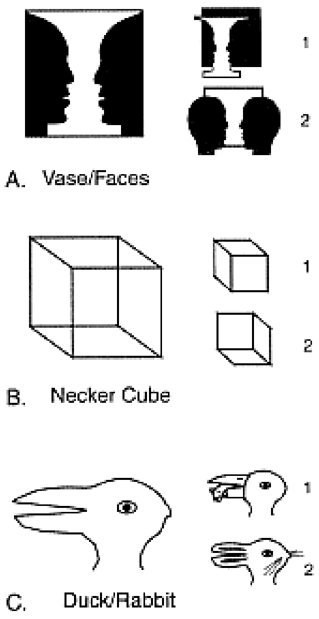

Figure 1.3 Ambiguous figures. Figure A can be seen either as a white vase against a black background or as a pair of black faces against a white background. Figure B can be seen as a cube viewed from above or below. Figure C can be seen as a duck (facing left) or a rabbit (facing right).

There are two important things to notice about your perception of these ambiguous figures as you look at them. First, the interpretations are mutually 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 modelbecause 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.

1.1.3 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. 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.

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 1.4A. 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 1.4B, in which exactly the same regions are present but not in a configuration that allows them to be completed.

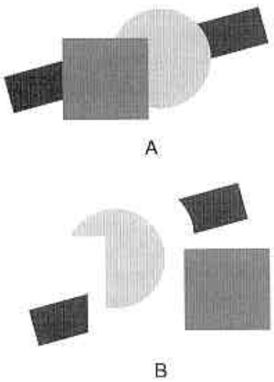

Figure 1.4 Visual completion behind partly occluding objects. Figure A is perceived as consisting of a square, a circle, and a rectangle even though the only visible regions are those shown separated in Figure B.

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.

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.

1.2 Vision as an "Inverse" Problem

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.

In contrast, 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 1.5 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.

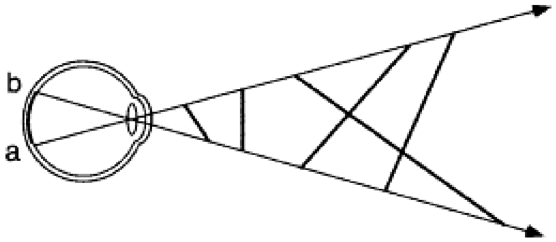

Figure 1.5 An illustration of inverse projection. A single line segment on the retina can be the projection of an infinite variety of lines in the environment.

We know that 3-D perception is possible because the human visual system manages to do it with such remarkable accuracy under most circumstances. How does it solve this seemingly insoluble problem? The dominant approach 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. Under most everyday circumstances, the assumptions are true, and so normal visual perception is highly veridical.

1.3 Theoretical Approaches to Vision

The modern era in vision science began in the 1950s and 1960s when three important developments fundamentally changed the way scientists understood vision: the use of computer simulations to model cognitive processes of various kinds, the application of information processing ideas to psychology, and the emergence of the idea that the brain is a biological processor of information. All three developments have exerted major influences on the evolution of vision science as an interdisciplinary field.

1.3.1 Computer Vision

A breakthrough of immense importance in the development of vision science was the idea that modern digital computers could be used to simulate complex perceptual processing. Vision and other forms of perception and cognition had previously been considered the exclusive province of living organisms. Scientists were therefore in the position of investigating a working system by constructing theories and testing them experimentally on living beings, which is an expensive and difficult enterprise. The emergence of computer simulation techniques changed the situation dramatically, however. It allowed vision scientists to build synthetic systems.

As early as the 1940s, Turing himself understood the incredible possibilities of his computing machine for simulating intelligent thought (Turing, 1950). This idea gave rise to the field of artificial intelligence (or AI), the branch of computer science in which computer programs are written to simulate intelligent behavior. Originally, AI theorists focused on trying to simulate difficult intellectual tasks such as playing chess and proving mathematical theorems (e.g., Newell, Shaw, & Simon, 1958; Newell & Simon, 1963). Only later was it realized that programming computers to perceive the environment visually would be a challenging and useful goal. This endeavor gave rise to the field now known as computer vision: the study of how computers can be programmed to extract useful information about the environment from optical images.

Computer vision promoted two important developments that changed the theoretical branch of vision science dramatically and forever:

1. Real images. Theories of vision simulated on computers can be applied to gray-scale images that have been obtained from video cameras recording real-world scenes. Classical theories of visual perception were generally designed to account for stimulus conditions that never exist in real situations: perfect, noiseless, line drawings of ideal objects. Computer vision changed this by allowing theories to be tested on real images of real objects, warts and all.

2. Explicit theories. Before computer simulations, theories of visual perception were vague, informal, and incomplete, stressing large conceptual issues at the expense of important details. Computer simulations changed this because one hallmark of computer programming is that it forces the theorist to make everything explicit.

The first insight derived from these two developments was the realization that vision is extremely difficult. It turns out to be unbelievably hard to get computers to "see" even the simplest things. Processes that had previously been taken for granted by psychological theorists (for example, detecting edges, finding regions, and determining which regions are part of the same three-dimensional object) have required heroic computational efforts. These are all tasks that the human visual system accomplishes with incredible speed and accuracy yet without any apparent effort. In comparison, even state-of-the-art computer programs running on the fastest, most powerful computers that have yet been devised fail to achieve the speed, accuracy, and flexibility of human perceivers on even such simple and basic visual tasks as these.

The second development, that of adopting a mathematical approach to creating working computer vision programs, was most clearly and effectively articulated at the Massachusetts Institute of Technology (M.I.T.), particularly by David Marr and his many talented colleagues. This research is characterized by mathematical analyses of how the luminance structure in two-dimensional images provides information about the structure of surfaces and objects in three-dimensional space (Marr, 1982).

1.3.2 Three Levels of Information Processing

In his influential book Vision, David Marr (1982) distinguished three different levels of description involved in understanding complex information processing systems: the computational, algorithmic, and implementational levels. In so doing, he provided a metatheoretical analysis of the information processing paradigm (see box). A metatheory is a theory about theories, a theory that attempts not to analyze vision itself, but to analyze the nature of theories about vision. Marr argued that there are important conceptual distinctions among these three levels and that all of them are essential for understanding vision—or anything else—as information processing.

Information Processing Paradigm

The information processing paradigm is a way of theorizing about the nature of the human mind as a computational process. It has been applied with considerable success not only to visual perception, but also to a wide range of cognitive phenomena in auditory perception, memory, language, judgment, thinking, and problem solving. In fact, the information processing approach so dominates these topics that several writers have suggested that it constitutes a "Kuhnian paradigm" for cognition.

The noted philosopher of science Thomas Kuhn (1962) defined a scientific paradigm as a set of working assumptions that a community of scientists shares (often implicitly) in conducting research on a given topic. The assumptions of a paradigm usually involve pretheoretical or metatheoretical ways of conceptualizing the major issues and sensible ways of approaching them theoretically. Kuhn describes the Newtonian view of physics as a paradigm that survived largely intact from the seventeenth century until the early part of the twentieth century. Although there had been many important theoretical developments since Newton's day—such as Maxwell's equations describing electromagnetic fields—none of them required rejecting the fundamental assumptions that underlay Newton's ideas about the nature of the physical world. For example, Newton and his successors implicitly or explicitly assumed that there is a qualitative distinction between mass and energy, that time is absolute, and that causality is deterministic. Quantum mechanics and Einstein's theory of relativity eventually brought about the demise of this Newtonian paradigm and ushered in a new paradigm that incorporated a new set of assumptions, including the ideas that mass and energy are equivalent, that time is relative, and that causation is inherently probabilistic.

The claim that information processing constitutes a paradigm for the cognitive sciences—including vision science—is based on the widely held belief that the nature of mental processes can be captured by theories that specify them in terms of information processing events. Although there remains a small but vocal subset of vision scientists who do not view visual perception as information processing, it is certainly the framework within which most current theories of visual perception are cast.

For those who are interested in other more recent paradigms to the Vision/AI problems, refer to Aloimonos (1993) and Ballard & Brown (1992) for the approach known as Active and Purposive Vision, and to Edelman (1987 and 1989) for a more biological approach.

E.g. for the purposive paradigm:

- The key question is to identify the goal of the task, the motivation being to ease the task by making explicit just that piece of information that is needed.
- Collision avoidance for autonomous vehicle navigation is an example where precise shape description is not needed.
- The approach may be heterogeneous and a qualitative answer may be sufficient in some cases.
- The paradigm does not yet have a solid theoretical basis, but the study of biological vision is a rich source of inspiration.

The Computational Level

The most abstract description that Marr proposed was the computational level. He defined it as the informational constraints available for mapping input information to output information. This level of theorizing specifies what computation needs to be performed and on what information it should be based, without specifying how it is accomplished.

To illustrate this concept, we will consider a very simple information processing system: a household thermostat. The "computation" that a thermostat must perform is to map both the current temperature of the air and the user's setting of preferred temperature (the input information) into an on/off signal for the furnace (the output information), depending on whether the air temperature is lower or higher than the set-point. We can summarize this computational level description in mathematical form as a binary function in two variables, where 0 is the output function, T is the temperature, and S is the set-point. Notice that we have not yet said anything about how this mathematical function is to be achieved; we have merely defined what the inputs are and how they are formally related to the outputs. This is a computational-level description of thermostat.

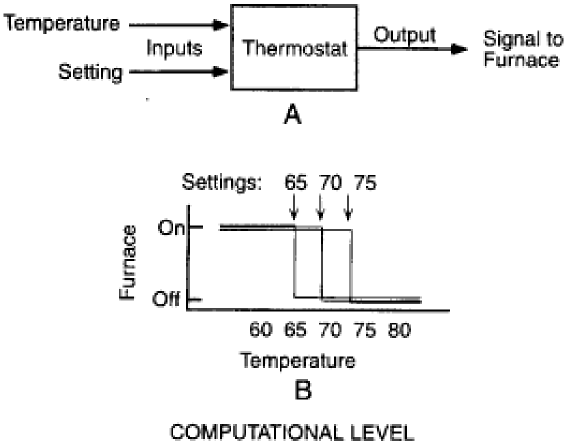

Figure 2.3.1 Computational-level description of a thermostat (A) The box diagram shows the inputs and output of a thermostat. (B) The graph plots the input/output behavior of the thermostat for three particular settings as a function of temperature.

The Algorithmic Level

At the middle level in Marr's hierarchy is the algorithmic level of description for an information processing system. Algorithmic descriptions are more specific than computational descriptions in that they specify how a computation is executed in terms of information processing operations. There are, in principle, many different ways in which a given computational-level mapping of input to output might be accomplished, in the same sense that there are many different computer programs that would accomplish the same computational task. Thus, the algorithmic level corresponds most closely to the concept of a program as it is understood in computer science.

To construct an algorithm for a given computationally defined task, one must decide upon a representation for the input and output information and construct a set of processes that will transform the input representation into the output representation in a well-defined manner. You can think of a representation as a way of encoding information about something and a process as a way of changing one representation into another. In the case of our thermostat example, the most obvious algorithm is to use one continuous variable to encode the temperature and another to encode the set-point and then to perform a comparison operation between these two magnitudes to determine whether the temperature is higher or lower than the set-point.

This is the standard algorithm for most thermostats, but others are also possible. One alternative would be to represent the temperature as a series of binary (two-valued) variables—say, above 60° versus below 60°, above 61° versus below 61°, and so on—and then to have the set-point select which of these temperature variables controls the output. (Note that the representation of temperature and set-point would be discrete in this case, in contrast to their continuous representations in the previous algorithm.) Still other algorithms are possible even for this simple information processing system, but the important point is that more than one algorithm can satisfy a given computational description.

There are many controversies about the nature of visual representations: whether the representation of a given fact is localized in a particular representing element or distributed over many such elements, whether a certain fact is represented explicitly or implicitly, whether all visual representations can be reduced to a finite set of primitive atoms or constitute an open-ended system.

The Implementational Level

The lowest level description is at the implementational level. It specifies how an algorithm actually is embodied as a physical process within a physical system. Just as the same program can be run on many computers that differ in their physical construction, so the same algorithm can be implemented using many physically different devices. We should stress here that by "different devices," we mean to include the possibility that the same algorithm might be implemented on brains as well as various different kinds of computers.

To illustrate the implementational level concretely, Figure 1.6 shows one way to construct a physical thermostat using the first algorithm that we described. The double curved line depicts a bimetallic strip, made by putting together two strips of metal that have different thermal expansion rates. The differential expansion of the two metals at different temperatures causes the strip to bend more or less as the temperature changes. The free end of this strip is part of a contact switch that completes an electrical circuit when it touches the contact. The vertical position of the contact is changed by a user adjusting the setting of the thermostat; raising the contact increases the set-point and lowering it decreases the set-point. Whether or not the switch closes thus depends on two factors: the height of the end of the bimetallic strip (as determined by the temperature) and the height of the contact (as determined by the setting).

This device implements the first thermostat algorithm as follows: The continuous temperature variable is implemented by the vertical position of the end of the strip. The continuous setting variable is implemented by the vertical position of the contact. The comparison of temperature and setting is made directly by the relative positions of the end of the strip and the contact. If they are touching, the circuit is completed, switching the furnace on. If they are not touching, the circuit is broken, switching the furnace off. This device does the job, but it is just one of many ways to build a thermostat based on this algorithm.

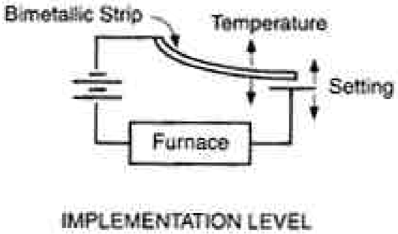

Figure 1.6 Implementational-level description of a thermostat.

1.3.3 Four Stages of Visual Perception

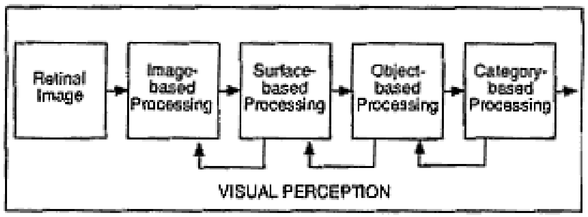

Figure 1.7 Four stages of visual processing. Visual processing can be divided into four main stages beyond the retinal image itself: image-based, surface-based, object-based, and category-based processing. (See text for details.)

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, as illustrated in Figure 1.7. Although different theorists refer to these stages by different names, we will use a generic labeling scheme in which each stage is named for the kind of information it represents explicitly: the image-based, surface-based, object-based, and category-based stages of perception. Much of the rationale for this theoretical framework came from the influential writings of David Marr (1982) and his colleagues at M.I.T.

The Retinal Image

The input stimulus for vision is the pair of 2-D images projected from the environment to the viewpoint of the observer's eyes. Figure 1.8 shows one such image of an extremely simple scene consisting of a ceramic cup resting on a flat, white surface in front of a dark wall. The optical image that strikes the retina is completely continuous, but its registration by the mosaic of retinal receptors is discrete. The complete set of firing rates in all receptors of both eyes therefore constitutes the first representation of optical information within the visual system. This retinal representation is complicated by the distribution of receptors. Receptors are more densely packed in the fovea than in the periphery.

In most computational theories of vision, the retinal representation is almost always simplified and regularized by approximating it as a homogeneous, two-dimensional array of receptors called pixels. 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.

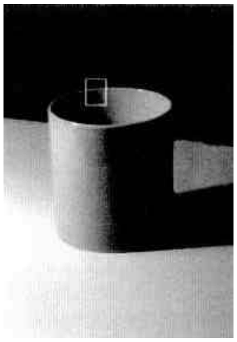

Figure 1.8 An image of a simple scene. People effortlessly perceive a ceramic cup resting on a table top, but all that is present to the visual system is an array of light whose intensity varies continuously over space.

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 retinally defined horizontal and vertical. Receptor positions are specified relative to this retinal frame of reference.

The Image-Based Stage

The image-based stage 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 1.9A indicates the locations of local edges that would constitute part of the image-based representation for the cup image shown in Figure 1.8.

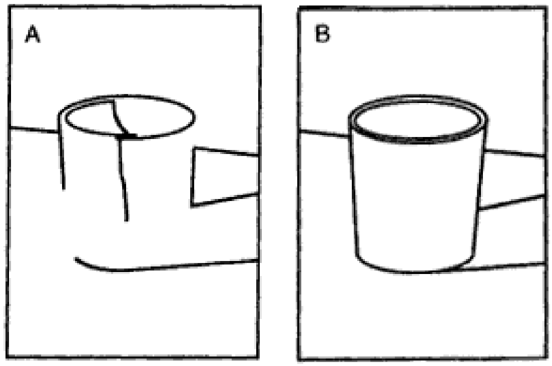

Figure 1.9 Edges in the cup image. Local intensity edges can be detected using computerized edge-finding algorithms, as illustrated in part A. The set of edges thus identified are not the same as those in a clean line drawing of the image (part B), however.

Notice that the luminance edges that have been detected in Figure 1.9A are not exactly the same as the edges most people readily identify in the same image, as shown in Figure 1.9B. 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 1.9A. Otherwise life would be 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 edge s, bars, blobs, 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.

Such image-based primitives 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 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. Marr named his surface-based representation the **2.5-D sketch** to emphasize the fact that it lies somewhere between the true 2-D structure of image-based representations and the true 3-D structure of object-based representations.

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 fiat over a sufficiently small region, just as the spherical earth seems fiat on the scale at which people experience it. This 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. Figure 1.10 illustrates what such a surface-based representation would look like for the ceramic cup in Figure 1.8 by showing circles lying on the local surface patches and vectors sticking perpendicularly out of them representing the normal directions.

The flow diagram in Figure 1.11 indicates that the representation of surfaces is constructed from several different sources: stereopsis (the small difference between the lateral position of objects in the images of the left and right eyes), motion parallax (differences in velocity of points at various distances due to motion of the observer or object), shading and shadows, and various other pictorial properties such as texture, size, shape, and occlusion. We will have more to say about stereo and motion later.

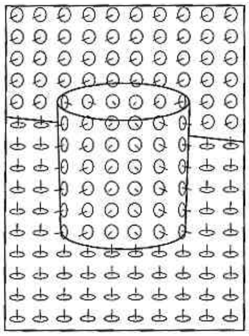

Figure 1.10 A surface-based representation of the cup scene. The surfaces visible in Figure 1.8 are represented as a set of local estimates of surface orientation (slant and tilt) and depth with respect to the viewer. Surface orientation is depicted by a set of imaginary circles on the surface and "needles" pointing perpendicularly out of them at a sampling of image locations.

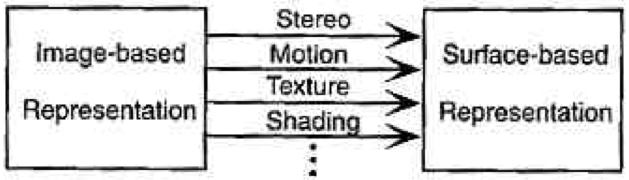

Figure 1.11 A flowchart showing how the surface-based representation might be derived from the image-based representation.

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 1.8 simply did not exist or that it had some quite different shape from the smooth cylindrical one everyone perceives so effortlessly. 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 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 1.12 shows as dashed lines the hidden edges that everyone perceives in Figure 1.8. 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.

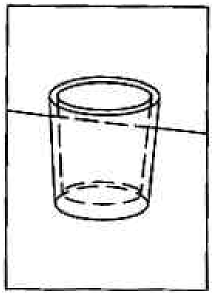

Figure 1.12 An example of an object-based representation. The cup of Figure 1.8 is shown with its occluded edges represented by dashed lines, indicating how people typically perceive this scene as being composed of 3-D volumes.

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 1.13 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.

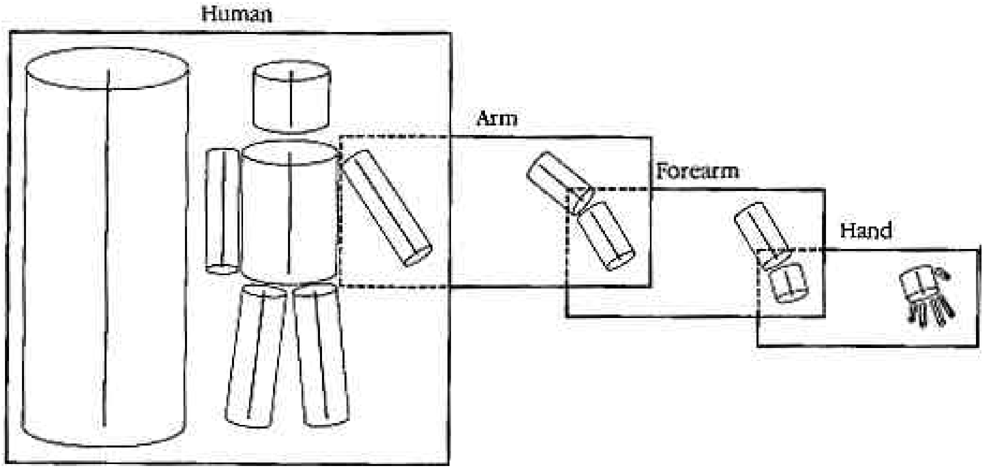

Figure 1.13 Using shape primitives in an object-based representation. The shape of a person's body as a 3-D volume is roughly represented here as a set of cylinders of the appropriate size, shape, orientation, and connectedness. Each box shows a perceptual object at a coarse, global level as a single cylinder and at a finer, more local level as a configuration of several cylinders.

The Category-Based Stage

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 1.8 as showing a cup is thus the result of 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 1.8 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. Koffka 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'" (Koffka, 1935, p. 7). For a more recent account of what categories reveal about the mind, you can refer to (Lakoff 1987). 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 its function 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 t hat 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.

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. Later processes may feed back to influence earlier ones. In the later part of this course, we will only examine mainly stage 2 and stage 3.

References:

Agin, G. J., & Binford, T. O. (1976). Computer description of curved objects. IEEE Transactions on Computers, C-25, 439-449.

Aloimonos, Y (1993). Active Perception. Aloimonos, Y. (Ed.). Lawrence Erlbaum Assoc. Pub.

Ballard, D. and Brown, C (1992). Principles of animate vision. CVGIP: Image Understanding, 45:3--21, Special Issue on Purposive, Qualitative, ActiveVision, Aloimonos, Y. (Ed.).

Edelman, G. M. (1987). Neural Darwinism: The theory of neuronal group selection. New York: Basic Books.

Edelman, G. M. (1989). The remembered present: A biological theory of consciousness. New York: Basic Books.

Lakoff, G (1987). Women, fire, and dangerous things : what categories reveal about the mind.
Chicago : University of Chicago Press.

Koffka, K. (1935). Principles of Gestalt psychology. New York: Harcourt, Brace.

Kuhn, T. S. (1962). The structure of scientific revolutions. Chicago: University of Chicago Press.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. San Francisco: W. H. Freeman.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. Proceedings of the Royal Society of London, 200, 269-294.