Chapter 1

Visual Surface Representation: A Critical Link between Lower-level and Higher-level Vision

Ken Nakayama, Zijiang J. He, and Shinsuke Shimojo

One of the most striking things about our visual experience is how dramatically it differs from our retinal image. Retinal images are formed on the back of our eyeballs, upside down; they are very unstable, abruptly shifting two to four times a second according to the movements of the eyes. Moreover, retinal images are sampled very selectively; the optic-nerve fibers that send information to the brain sample more densely from the central area than from peripheral portions of our retinae. Yet, the visual scene appears to us as upright, stable, and homogeneous. Our perception is closely tied to surfaces and objects in the real world; it does not seem tightly related to our retinal images.

The goal in this chapter is to illuminate some of the most elementary aspects of perception as a way of arguing that an indispensable part of perception is the encoding of surfaces. We believe that a surface representation forms a critical intermediate stage of vision poised between the earliest pickup of image information and later stages, such as object recognition. In addition, it is probably the first stage of neural information processing, the results of which are available to us as conscious perceivers.

Why do we think surfaces are so important? The visual part of our brain is not an abstract or neutral information transmission system but one that must capture significant and recurring aspects of our visual environment. Early stages of our visual brain must begin to encode what is most general about our visual environment, providing information about diverse scenes, many of which will differ greatly one from another in terms of specific objects and their layout.

The surface-representation level may provide this necessary intermediate stage for the development of more complex visual processing—for

Thanks are due to the Life Sciences Directorate of AFOSR for support of the research reported here, to Dr. Charles Stromeyer for a painstaking review and criticism of the manuscript, and to Dr. Nava Rubin, Satoru Suzuki, and Emre Yilmaz for their constructive comments.

locomotion across a world of surfaces and for manipulation and recognition of objects that are defined by surfaces.

One of the most important characteristics of a world defined by surfaces is that it is three dimensional; ordinarily it has a ground plane below and is accompanied by other assorted surfaces, many of which occlude each other. This means that we cannot expect to see just one surface at a time along any given direction of gaze. Often we see multiple surfaces in local regions of visual space, with closer objects at least partially covering those behind. Thus many surface regions have no counterpart in the retinal image. Yet, remarkably, we do not feel much loss of information when part of a surface is rendered invisible by occlusion; we do not see invisible surface regions as nonexistent. This suggests that we are making "unconscious inferences" (Helmholtz 1910) about literally invisible entities. In the two-dimensional drawing shown in Figure 1.1, we encounter a small set of closed forms that are almost impossible for us to perceive as simply two dimensional. Even without recognizing the lines or patches as parts of familiar objects, we automatically see the configuration as part of a scene in depth and infer that patch **x** is in front of patches **y** and **z**. More important, we infer that patches **y** and **z** make up the same surface and that this surface continues behind surface **x**.

Where in the brain are such inferences made? If we use the word inference, of course, we invite all kinds of possibilities. Is it the kind of inference that we associate with ordinary thinking? Or is it something more visual, linked more specifically to the visual system? We are persuaded by the latter view and shall argue that such inferences are tightly and exclusively tied to visual processing. Our view is that such inferences are embedded in the visual system and can occur at surprisingly early stages, almost independent of our knowledge about familiar objects.

Before continuing our description of surfaces and surface representation, however, we pause to outline briefly what is generally understood about lower-level and higher-level vision as a general context for our results.

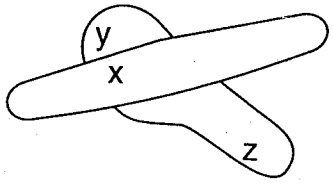

Figure 1.1
Three two-dimensional image areas, labeled **x**, **y**, **z**. These do not combine to form an obviously recognizable object. Nevertheless, region **x** is perceived to be in front, covering regions **y** and **z**, which are perceived to be part of the same surface completing behind **x**.

First, consider the anatomy of the brain, especially the brains of primates. One of the most startling conclusions to emerge over the past forty years is that approximately 50 percent of the cerebral cortex of the macaque monkey is devoted to vision; the estimated percentage in humans is only slightly smaller (Zeki 1978; Allman and Kaas 1974; Van Essen et al. 1990). At first glance this might seem disproportionate, given the apparent ease and simplicity of seeing, in comparison to, say, thinking, language, or problem solving. Seeing seems so automatic that it might lead us to assume that it requires much less processing. Yet, again, the past forty years of brain research have begun to indicate otherwise, that vision is an extremely complex process, so complex that it is now conceivable that it occupies a sizable fraction of our brains.

Let us look at some specific details. The visual system of the macaque monkey, an animal whose low-level visual capacities are remarkably similar to our own (DeValois et al. 1974), is increasingly understood as an elaborate hierarchical system subserving diverse ultimate functions. The animal's retina contains over one hundred million photoreceptors that send over a million axons to the cerebral cortex via the thalamus in fairly precise register; thus, different parts of the visual field have their exact counterparts in the striate cortex, the first visual cortical receiving area. Surprisingly, more than twenty such separate maps of the retina are projected onto the cortex (Maunsell and Newsome 1987).

What might all these additional visual areas be used for? Little is known. Yet, there is now some evidence that these higher-level visual areas can be divided into at least two streams that serve different higher-order visual functions. In a provocative theoretical speculation—based primarily on anatomy and the results of lesion studies in monkeys and clinical cases in humans—Ungerleider and Mishkin (1982) have suggested that these many cortical areas can be roughly categorized into several substreams that point to important sets of disparate functions for vision. A ventral stream is important for object recognition; damage here leads to an inability to recognize objects in monkeys and to severe losses of object recognition in human patients. A dorsal stream is more specialized for determining the position of points in space or the spatial relations between them. Others (e.g., Goodale et al., this volume) have suggested that this dorsal system might best be described as relating to spatially guided motor behavior, for example, reaching and grasping. These two streams are depicted in Figure 1.2. For the moment, we consider them in order to characterize the major higher-order functions of vision and their anatomical substrates. (See also chapters 3, 4, 5, and 7, this volume.)

These higher-level functions must have input from lower-order visual processes, which, in turn, must receive inputs from the retina and striate cortex. What kinds of information are necessary to serve as useful input for such diverse higher-order functions?

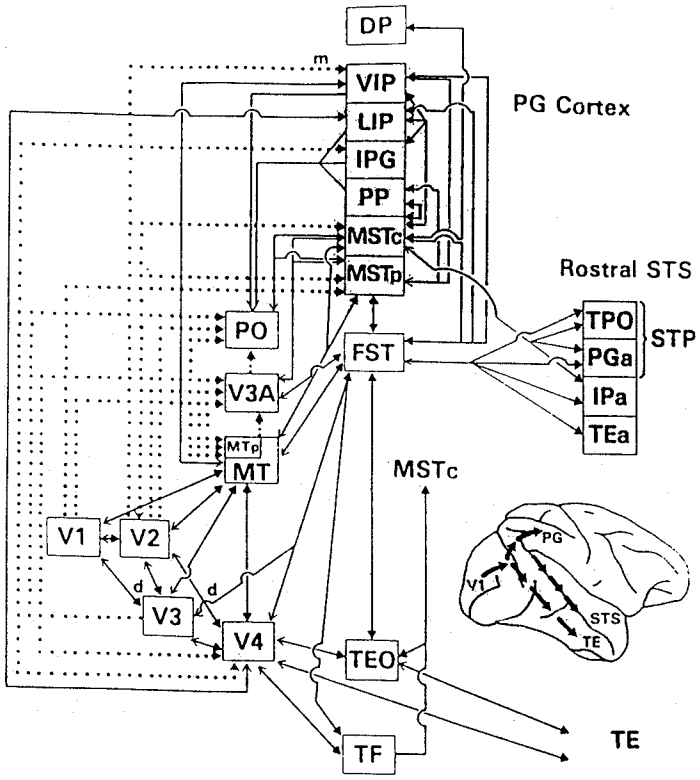

Figure 1.2
Schematic diagram of the connectivity of some of the known cortical areas, grouped into a dorsal and ventral stream. (Reprinted by permission from L. G. Ungerleider and M. Mishkin, Two cortical visual systems. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, eds., *Analysis of visual behavior*, 1982. Copyright 1982 by MIT Press.)

As a start, we might think of the signals arising from well-known classes of visual neurons in the early visual pathway. Work in this area was pioneered by electrophysiological recordings, starting about forty years ago (Barlow 1953; Lettvin et al. 1959; Hubel and Wiesel 1959). By painstakingly recording from one nerve cell at a time, neurophysiologists have pieced together an unusually detailed account of how neurons respond to visual input at various stages of the visual pathway—from the photoreceptors, to the visual cortex, and beyond. Not surprisingly, photoreceptors are sensitive to just a small localized part of the visual field, responding only when light is presented in this very small region. Additional procession is evident, however, when we record the responses of the ganglion cells, the cells that make up the fibers of the optic nerve and

convey information from the eye to the brain. Instead of simply responding to luminance in a given region, these cells respond to luminance differences. In general, they respond best when light is flashed in a small circular region of the visual field and not in its immediate local surround.

As one records from the primary or striate visual cortex, additional selectivity becomes evident. Cells here respond to a more specific local geometrical-image property, that of orientation. For example, one class of visual cortical cells responds best to an oriented blob, at, say, 45 degrees, another to an oriented edge of the same orientation, while others respond to other edges, bars, or blobs at other orientations. Thus, if we think of the visual cortex as a whole, there appears to be a complete set of analyzers for each retinal location, each one sensitive to different orientations (and sizes of image regions). The region of the visual field that can influence the firing rate of a cell is called its receptive field. It is clear from analysis of cells and their receptive fields that different aspects of the visual image are coded in different sets of visual neurons.

Although most cells of the striate visual cortex respond more vigorously to one eye or the other, some are binocular. These cells have separate yet highly similar receptive fields mediated through each eye and have the same orientation preference and position in the visual field. Yet careful measurement reveals that for some binocular cells the relative position of the receptive fields in each eye is slightly offset (Barlow et al. 1967; Poggio and Fischer 1979). This is an important discovery because we have known for many years (Wheatstone 1838) that the small difference between image points in two fused photographs or line drawings is the basis of stereoscopic vision. This means that if an animal fixates on a given point in space, different cells will respond differentially to the relative depth of a given visual stimulus, suggesting that a population of disparity-sensitive binocular cells can provide the visual system with a method of encoding stereoscopic depth in a scene.

From what we have said so far, it is evident that the properties of single cells as embedded in the visual system are remarkable; they are selective to complex visual patterns and, even more specifically, to the depth of visual stimuli. In fact, much of the modern work on visual perception assumes that we can understand perception in terms of the properties of these cells; Barlow (1972) has espoused an explicit neuron doctrine for perception. The example of neurons with differing binocular separations seems to go a long way toward explaining how we see stereoscopic depth in natural scenes.

Motion perception seems to be another area in which single-cell recording would be explanatory. In all species studied, cells have been found that are highly sensitive to the direction of image motion. Such cells respond to movement in one direction but not to movement in the opposite direction (Barlow and Levick 1965; Nakayama 1985).

One might conclude from these very impressive findings that perception is simply the working out of the firing patterns of single cells. To understand how we see things, all we need do is continue to explore the response properties of visual neurons. We might think that this level of processing machinery could deliver an adequate representation to the higher functions of object recognition and visuo-motor control. Yet, although we do not deny that some aspects of perception are illuminated by understanding the properties of these cells, they do not adequately explain the specific aspects of perception we shall describe in this chapter. *Our view is that higher functions require, as an input, a data format that explicitly represents the scene as a set of surfaces.*

We have, therefore, divided the remaining portion of this chapter into three sections: Part 1.1 surveys the phenomenology of surface perception; Part 1.2 examines experimental studies showing the importance of surfaces; and Part 1.3 presents our theoretical understanding of the mechanisms of surface perception. In Part 1.1, we consider certain perceptual demonstrations, some of them familiar to the reader, which show how the viewing of very simple patterns is surprisingly revealing of the underlying properties of surface perception. We show that surface perception requires an inferential process residing largely "within" the visual system. These inferences do not require higher-level cognitive processing based on the knowledge of familiar objections. In Part 1.2, our goal is twofold. First, we report on experiments that confirm that phenomenological descriptions, thus adding weight to our previous analysis. Second, we show that the role of surface representation is crucial in a wide variety of visual functions, even those that have been traditionally thought to be directly mediated by the properties of early cortical neurons. The visual functions we studied include visual search, visual object recognition, visual motion perception, and visual texture perception. These studies indicate that seemingly primitive visual functions require, as a prerequisite, the analysis of visual surfaces. We also demonstrate that space perception and visual attention cannot be understood independent of an explicit consideration of a surface representation. In Part 1.3, we suggest a possible site in the brain where surface representation might begin and conclude in a more theoretical vein, suggesting a framework for understanding the perceptual learning of surfaces from images.

1.1 Phenomenological Studies

Experimental phenomenology is a valuable tool for studying perception. It requires the discerning characterization of a person's visual experience in response to well-defined stimuli. Although it is somewhat unusual in a scientific field to dwell on the details of private conscious experiences, it is

an essential step to understanding perception. It also gives the study of perception its particular immediacy. Contemporary scientific research in most fields requires complex measuring devices, extensive data collection, and statistical analysis—all of which distance us from the primary data. The study of perception affords the student and researcher alike the opportunity to experience at first hand some of the basic facts of vision. If well conceived, perceptual demonstrations provide viewers with unusually direct access to the nature of their own perceptual machinery.

Some of these demonstrations may be familiar to the reader. They have been marvelled at and endlessly reproduced, gracing textbooks and popular works alike. Yet, despite their wide exposure, some of these demonstrations are often misunderstood, even by experts. Furthermore, they have not been used as part of an overall argument for the existence of a separate stage of visual surface representation.

We start with one of the most famous demonstrations, the Rubin (1921) face-vase phenomenon (see Figure 1.3). Sometimes we see a pair of faces, and sometimes we see a single vase. Additional reflection on what we are seeing leads to several important conclusions. First, the perception is bistable, meaning that we see either the vase as a figure or faces as figures. Second, when one portion of the picture becomes the figure, the other portion degenerates. Yet it doesn't just become less visible; it becomes the background, continuing behind. Third, with each perceptual reversal, there is also a reordering of depth. Whichever portion is seen as the figure always appears to be closer.

Before attempting to explain this demonstration in terms of surface perception, we need to deal with an obvious objection. Maybe the face-vase reversal has nothing to do with surface representation but is mediated at a higher cognitive level, say at the level of object representation. Its bistability may rely on the fact that we all know what faces and vases look like and that we alternate between the two because one can only look at one recognizable object at a time.

This concern is addressed in Figure 1.4, which was also introduced by Rubin. Even though none of the patches on the left or the right are familiar or easily identifiable, the same reversal occurs; and the basic phenomenological effects described for the face-vase figure can be confirmed. This suggests that figure-ground reversal does not depend on such higher levels of processing as object recognition. However, this demonstration, as well as the original Rubin face-vase demonstration, is very different from other well-known classes of ambiguous figures, such as the famous Jastrow Rabbit-Duck illusion in Figure 1.5. Of course, they share the same bistable reversing quality, which suggests a similarity. Yet, there is a fundamental difference. In Figures 1.3 and 1.4, what switches is the patch that is seen as either foreground or background; as described earlier, this also involves a

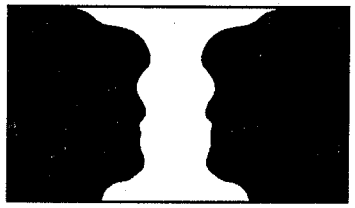

Figure 1.3
Face-Vase reversing figure. (Adapted from E. Rubin, *Visuall wahrgenommene Figuren* [Copenhagen, 1921].)

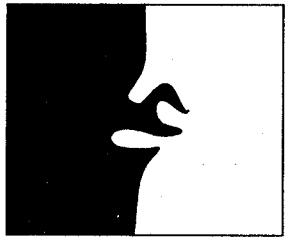

Figure 1.4
Reversing figure without familiar objects. (Adapted from E. Rubin, *Visuall wahrgenommene Figuren* [Copenhagen, 1921].)

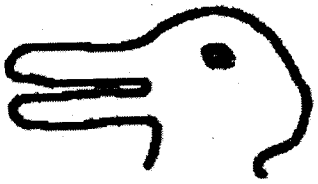

Figure 1.5
Rabbit-Duck reversing figure. In contrast to the previous two demonstrations, foreground and background do not reverse when the perception reverses, suggesting that this is a different class of figural reversal, one mediated by object-level processes.

reversal of depth perception. With the Rabbit-Duck, however, no such reversal of foreground-background occurs. The figure is always seen in the foreground. What varies is the object perceived. Unlike the two other cases, the Rabbit-Duck involves a reversal at the level of object recognition, requiring object knowledge.

Based on this discussion, we suggest that the figure-ground reversal reflects a more basic, autonomously driven mechanism that is relatively free from top-down, object-level knowledge. In other words, we see evidence of a level of perceptual analysis that is interposed between cells with particular receptive fields, say in the striate visual cortex (as studied by neurophysiologists), and such later stages of visual representation as object recognition.

Figure 1.6 schematizes our placement of the level of visual surface representation as an independent, explicit stage of visual analysis in relation to the overall scheme outlined in Figure 1.2. It is a general purpose, intermediate representation in that it codes enduring aspects of our physical world yet is not concerned with detailed specifics. This surface level

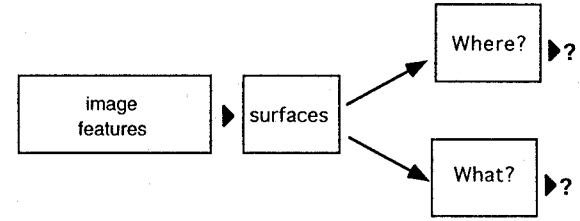

Figure 1.6
Presumed placement of surface representation in relation to lower-level and higher-level visual functions.

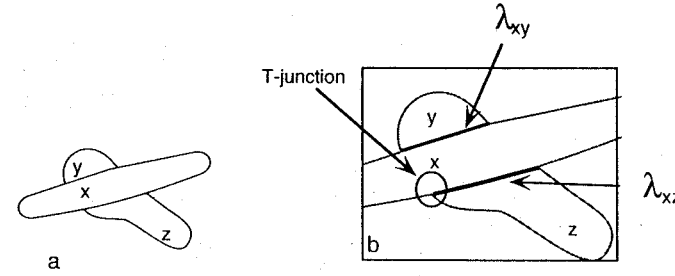

Figure 1.7
Inset of Figure 1.1, detailing the common border λ_{xy} between region **x** and **y** and the common border λ_{xz} between region **x** and **z**. In the parsing of surfaces, the visual system needs to determine which surface regions "own" these common or shared borders. Note existence of T-junction (in circle), which helps to establish depth ordering.

determines whether surfaces are seen as connected or disconnected; folded, straight, or curved; whether they pass in front of or behind; whether they are transparent or opaque. Again, we see this level as distinct from object-level processing, which requires knowledge of specific object or object classes.

1.1.1 Amodal Completion of Occluded Surfaces

Our research has shown that adopting a few simple rules makes surface representation much more comprehensible. For clarity, we initially outline these rules semi-dogmatically, illustrating them with the example presented in Figure 1.7.

Rule 1. *When image regions corresponding to different surfaces meet, only one region can "own" the border between them.* Thus in Figure 1.7, it is important for the visual system to assign ownership to contours λ_{xy} and λ_{xz}. For example, it needs to decide which image region, **x** or **y**, owns the contour λ_{xy}.

Rule 2. *Under conditions of surface opacity, a border is owned by the region that is coded as being in front.* In Figure 1.7, this means that region x "owns" the border λ_{xy}.

Rule 3. *A region that does not own a border is effectively unbounded. Unbounded regions can connect to other unbounded regions to form larger surfaces completing behind.* We call such completion *amodal* completion after Michotte (1964) and Kanizsa (1979).

To see how these rules might play out in actual practice, consider the border between region **x** and region **y** as well as the border between region **x** and **z**. In Figure 1.7, Rule 2 states that the border is owned by the region that is coded as in front. How does the visual system know a region is in front? In this case, the information is supplied by what are known as T-junctions, one of which is circled in Figure 1.7b. This is a junction where three lines meet. Two of the lines are collinear, forming the top of a T; the other line forms the stem of the T. In many natural scenes, such T-junctions are good (but not entirely infallible) clues to depth and occlusion. The top of the T is usually the occluding contour, occluding the stem of the T presumed to continue behind.

Now consider the image patches **y** and **z**. Note that the borders shared with patch **x**, λ_{xy} and λ_{xz}, belong to patch **x**. This means that at this border, regions **y** and **z** are essentially unbounded. Then, according to Rule 3, region **y** and **z** can become connected behind the occluder.

To illustrate these points in a different way, we generate a stimulus (Figure 1.8) in which border ownership changes with the introduction of an occluding figure. Thus, when the individual fragments of the letter *B*s

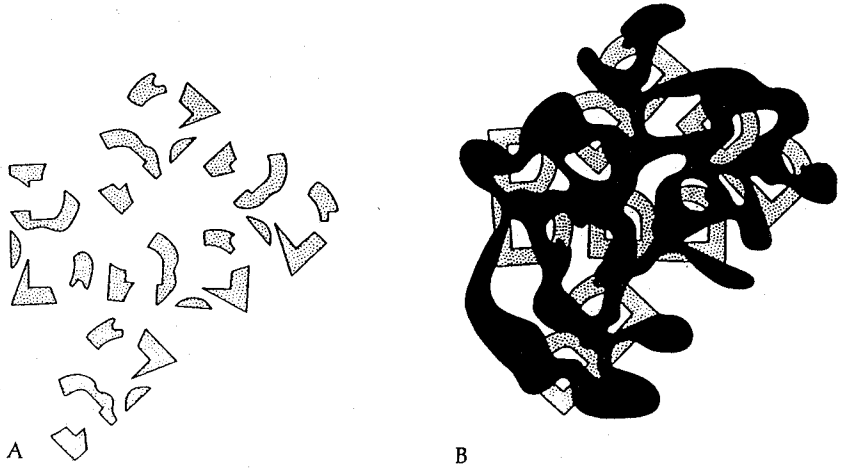

Figure 1.8
Effect of occluder facilitating the recognition of an object behind. (A) Fragments of the letter B. (B) Same fragments plus the occluder, which makes recognition of the letter much easier. Also, note the existence of T-junctions indicating depth and occlusion. (Reproduced with permission from A. L. Bregman, Asking the "what for" question in auditory perception. In M. Kubovy and J. R. Pomerantz, eds., *Perceptual Organization*, 1981. Copyright 1981 by L. Erlbaum Associates.

are presented, we cannot see the *B*s. Only when the occluder is present can we discern the letters. Again, the presence of T-junctions in Figure 1.8a and not in Figure 1.8b justifies the rules we have outlined.

At this point the reader may feel uncomfortable. Sure, the basic ideas are reasonable, but isn't there a kind of logical circularity, particularly because we said that T-junctions indicate occlusion and that they provide information for the stem of the T-junction to continue behind? We have T-junctions in both Figure 1.7a and 1.8b. Isn't there another way of defining depth without T-junctions?

In the remainder of the chapter, we will rely strongly on a fairly obvious and effective method of introducing depth—binocular disparity. For those unfamiliar with stereograms, we include an appendix to the chapter describing various ways of gaining proficiency in the perception of three-dimensional scenes from fused image pairs without using glasses or optical aids. We use stereograms, not because we are interested in binocular disparity or stereopsis itself, but because of the unusual advantages inherent in this method of creating depth. What is particularly useful about binocular disparity is that dramatic changes in depth can be created by tiny shifts in image position. Furthermore, by switching left and right images, we can reverse depth without changing the total amount of information

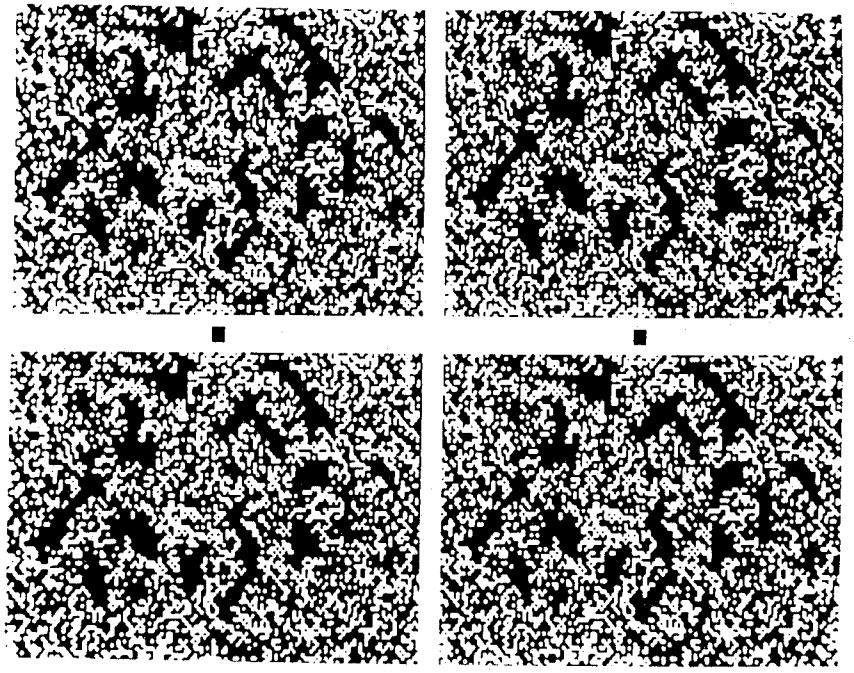

Figure 1.9
Fragments of the letter *B* revealed by stereoscopic depth without T-junctions. First, notice that without stereoscopic fusion, it is essentially impossible to perceive the fragments as comprising parts of *B*s. Crossed fusion of the top two half-images or parallel fusion of the bottom two half-images will show a snake-like figure in front connecting the fragments without the benefit of T-junctions. (Reprinted by permission from K. Nakayama, S. Shimojo, and G. H. Silverman, Stereoscopic depth: Its relation to image segmentation, grouping, and the recognition of occluded objects, 1989, *Perception* 18, 55–68.)

given to the two eyes. So, if our depth hypothesis (i.e., Rule 2) is correct, we should be able to make radical changes in the perceived layout of surfaces with otherwise imperceptible changes in the monocular image.

We should, therefore, be able to show the efficacy of rules 1, 2, and 3 without the benefit of T-junctions. Figure 1.9 is a stereogram showing fragments of *B*s lying in a background plane with a snake-like occluder in front of it. Note that the *B*s are essentially invisible if the pattern is not fused; they are not seen if there is no depth ordering. With stereoscopic fusion, however, something dramatic happens. The *B*s in the background are now clearly visible as individual fragments join to complete the letters behind other surfaces defined stereoscopically. This demonstrates that monocularly defined T-junctions alone do not control the selective completion of surfaces behind occluders. In the next demonstration, in Figure

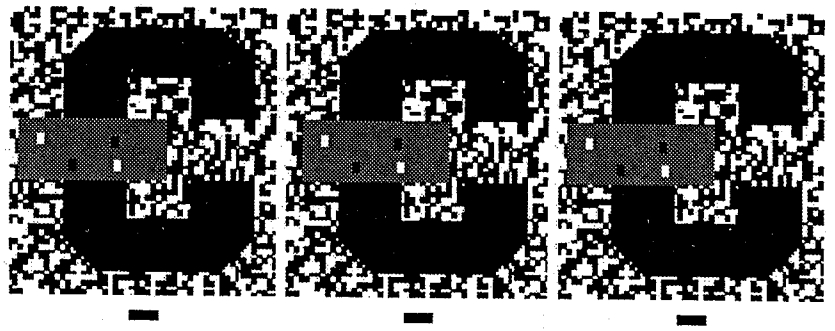

Figure 1.10
Perception of the letter *C* as influenced by depth. When the figure is normally viewed as a stereogram, we see a *C*, amodally completing behind a small gray rectangular occluder in front. When viewed in the reverse configuration such that the occluder is seen as behind, we perceive two disconnected U-shaped fragments and no longer perceive the fragments as part of a *C*. (Reprinted by permission from K. Nakayama and S. Shimojo, and G. H. Silverman, Stereoscopic depth: Its relation to image segmentation, grouping, and the recognition of occluded objects, 1989, *Perception* 18, 55–68.

1.10, we make the point even more forcefully, by showing that stereoscopic depth can easily overrule existing T-junctions. Without stereoscopic fusion, we see a complete large letter *C* behind a gray rectangular occluder. This is not surprising in light of the arguments presented so far. Because the gray patch (via T-junctions) is perceived to be in front, ownership of the common border (according to Rule 2) is ceded to the rectangle and the remaining image fragments are unbounded, thus completing amodally behind (according to Rule 3).

When stereoscopically fused, no perceived change is expected, because the depth defined by binocular disparity and by the T-junction are in agreement. Both are compatible with interpreting the gray patch as in front, allowing the *C* to remain as highly visible, completing behind the occluder. The reader can verify this by either cross fusing the two left images or parallel fusing the two right images (as described in the Appendix). Perception is very different, however, when the images of the two eyes are reversed such that the gray patch is seen as behind. When this happens, the pieces of the *C* break up into isolated fragments, forming two *U*s—one upright, one inverted, separate and ungrouped. The *C* is no longer visible.

At this point, it should be clear that our perception of recognizable objects can be dramatically influenced by visual surface representation. In addition, we see that perceived depth is extremely important in the perception of objects, although not in the sense usually assumed. Rather than being used to represent the internal three-dimensional structure of the

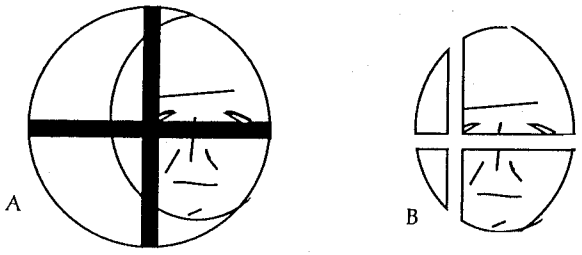

Figure 1.11
(A) Schematic face seen through a window. (B) Face fragments alone (see text).

objects themselves, depth has a more important role: it determines what pieces of an image actually comprise the object to the recognized. Depth is needed to parse objects into wholes or parts, to determine what in an image actually constitutes the parts and boundaries of a single object (Nakayama, Shimojo, and Silverman 1989). In other words, depth dictates perceptual grouping and perceptual segmentation.

Next, we need to deal more specifically with the issue of object recognition. Many contemporary theorists suggest that object recognition requires the matching of stored templates to portions of an image (Biederman 1987; Marr 1982; Nakayama 1990). Examples of ordinary occlusion suggest that there is a profound problem in determining what part of an image will be used for the template matching, a process presumed to occur in object recognition. We cannot simply match templates with the raw, or even filtered, image; because some very spurious matches would be made, preventing the operation of any reasonable recognition mechanism. This problem can be perhaps illustrated by the cartoon shown in Figure 1.11. In *A* we see a face through a circular, paned window, whereas in *B*, we see only the visible face fragments. Face recognition has often been seen as a holistic process (see Chapter 3, this volume). An important consideration for this recognition is presumed to be the overall outline of the face and the exact spatial relations between its various parts—not just recognition of the parts themselves. For example, in Figure 1.11b the face is spuriously elongated because the boundary of the window is interpreted as the boundary of the face. How is it then, that we can recognize a face even when it is broken up into pieces and when the outline of the pieces no longer conform to the outline of the face?

This type of problem reinforces our conviction that before the process of object recognition can begin, an object must be separated from the rest of the image and made available to the mechanisms of pattern recognition. This realization further justifies the flow chart outlined in Figure 1.6 and

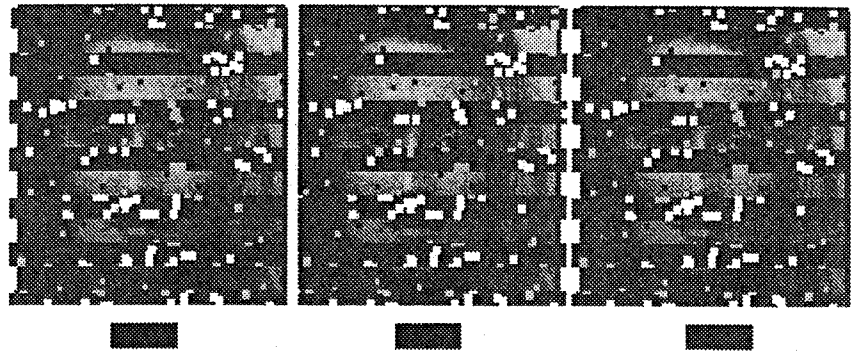

Figure 1.12
Stereogram of a face either in front of or behind occluding strips. Note that the face is more easily perceived when it is behind. (Reprinted by permission from K. Nakayama, S. Shimojo, and G. H. Silverman, Stereoscopic depth: Its relation to image segmentation, grouping, and the recognition of occluded objects, 1989, *Perception* 18, 55–68.)

suggests that we cannot think of object recognition as proceeding from image properties such as those captured by early cortical receptive fields; there needs to be an explicit parsing of the image into surfaces. Without such parsing of surfaces, object recognition cannot occur.

Figure 1.12 is a striking illustration of the importance of depth and surface parsing for object recognition. It shows interrupted strips containing partial images of a face that, viewed stereoscopically, can be seen as either in front of or behind the other interlaced strips. The information available about the hidden face is identical for both depth conditions. Yet, there is an obvious difference in our ability to perceive the face. It is very difficult when the face fragments are in front; but when they are in back, perception is hardly disrupted. It is almost as if all the face is there behind the occluding strips (Nakayama et al. 1989). Again, we see this large difference in the face's visibility as the specific playing out of Rules 1, 2 and 3. When the face fragments are in front, each face strip owns the border between the face and non-face; linkage between the strips does not occur. With the face fragments in back, the common border is owned by the occluding strips in front and the face fragments in back are unbounded, leading to surface completion.

Completion of objects behind nearer objects is ubiquitous in our daily lives. Our demonstrations show that the completion of image fragments behind occluders is not arbitrary but acts according to very specific and highly adaptive rules. It depends on depth and, as a consequence, border ownership, which in turn dictates which image fragments are grouped or segregated.

1.1.2 Completion of Surfaces in Front (Modal Completion of Subjective Surfaces)

Although the need to complete surfaces behind occluders is very frequent in everyday life, occasionally we also need to infer the existence of contours and surfaces in front of other surfaces. This occurs when the luminance difference between a foreground and a background surface is not evident, due to poor illumination or to the chance identity of foreground and background luminance. This situation raises the issue of subjective contours and subjective surfaces.

Thanks to the well-crafted demonstrations of Kanizsa (1979), we are well aware that our brain can create a contour where none exists in the image (see Figure 1.13). Kanizsa describes such contours as examples of *modal*, or *visible* completions. He notes that modal contours and surfaces must complete in front of other surfaces, in contrast to amodal, or invisible completion, which indicates a completion behind other surfaces. Although modal completion has received far greater attention than amodal completion, they have much in common (Kellman and Shipley 1991). Most important, they both qualify as inferences, testimony that our visual system can determine the presence of an edge or surface from incomplete information.

We can ask the same questions about modal completion as about amodal completion. Where do such inferences occur? Are these perceived contours inferences of the sort we make in our daily life or are they inferences made within the confines of the visual system? In the past, these contours were also dubbed *cognitive contours*, implying that thinking or problem solving is involved (Gregory 1972; Rock 1984). Nowadays, the term cognitive contour is little used, and the reasons are important. From what we have said so far, we might argue that the contours seen in Figure 1.13 could have been constructed by some type of top-down inference; that is, we could say that we could reason that a triangle could have covered the adjacent region, thus justifying the term cognitive contour. The same might hold for the sinusoidal contour seen in Figure 1.13b. The

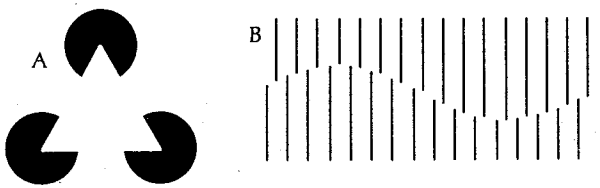

Figure 1.13
Subjective or illusory contours. (A) Kanizsa triangle. (B) Subjective sinusoidal contour formed by offset abutting lines.

almost palpable sense that we *see* the contour certainly argues against this notion of a higher-level inference. But better evidence is needed. In this regard, two additional classes of subjective contours, those driven by binocular disparity and those occasioned by DaVinci stereopsis, are relevant.

If we look at the fish-like silhouettes in Figure 1.14 (adapted from Kanizsa 1979) before fusing them stereoscopically, we can sometimes imagine subjective contours or cognitive contours, with the heads or tails of one "fish" occluding the other "fish." Although there is a tendency to see the broader "heads" covering the narrower "tails," it can reverse. So, our cognitive knowledge or imagination can influence the perception of such contours, particularly when the scene is very impoverished. When fused as a stereogram, however, the specific layout of the subjective contours are immediately apparent. Cross fusing the two half-images on the left and center, we see the tails in front and, automatically, their boundaries as subjective occluding contours. In the opposite stereo case, the heads are seen in front, and we immediately see visual subjective contours bounding them. Furthermore, the perception of these contours is stable and unchanging; one is hard-pressed to argue for some form of deliberate top-down inference in this case. The subjective contours appear to be formed by an efficient, adaptive, and autonomous process driven, in this case, by binocular disparity signals, which overcome higher-order knowledge or expectations about objects.

Even more telling is the case of DaVinci stereopsis. In an earlier study, we created subjective contours in a situation where higher-order inference cannot occur; that is, where no information is available at a conscious level (Nakayama and Shimojo 1990). To understand DaVinci stereopsis, it is necessary to appreciate that some regions of most real-world scenes are

Figure 1.14
Stereo version of Kanizsa "fish." Although this is a flat two-dimensional figure, observers generally see depth even when pairs of images are not fused as a stereogram. Unfused, two different surface arrangements are apparent. Usually one sees the broader "heads" in front with visible subjective contours completing in front of the narrower "tails." At other times, one sees the narrower "tails" in front, bounded by their corresponding subjective contours. When fused stereoscopically, binocular disparity determines the depth placement of the heads and tails accompanied by subjective contours.

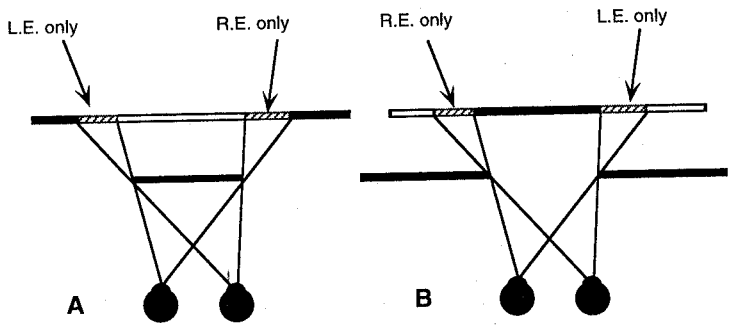

Figure 1.15
Top view of two scenes outlining the geometrical foundations of DaVinci stereopsis. (A) An opaque square in front of a frontoparallel surface. (B) A more distant surface seen through an aperture or window. In each case the differential binocular optical consequences of occlusion are characteristic and invariant. Regions seen by the left eye only are on the farther surface just to the left of a closer occluding surface. Regions seen by the right eye only are on the farther surface just to the right of a closer occluding surface.

visible to one eye or the other but not to both. This can be understood by referring to Figure 1.15. This top-view diagram shows two situations: Figure 1.15a shows a square occluding a wall behind; Figure 1.15b shows a distant wall viewed through a window. Because the closer surface is opaque, there are regions (depicted as the hatched area) that are visible only to the right eye or the left eye.

Such half-occlusions, or unpaired points, arise almost constantly in our everyday life because we are inevitably exposed to the edges of objects at different depths. These unpaired regions lie on more distant surfaces adjacent to the image boundaries of nearer surfaces. What is important for our discussion is the highly constrained nature of this binocular unpairedness. First note the obvious fact that our eyes are horizontally aligned and thus have different viewpoints along a single horizontal dimension. This means, in general, that such half-occlusions occur only when there is a vertical component to an occluding contour; they do not occur for purely horizontal edges. Even more important is the fact that there is an obligatory, nonarbitrary relation between a given unpaired point and the placement of the occluding contour that causes it to be unpaired. Unpaired right-eye-only points are seen only next to occluding contours to their immediate left. Unpaired left-eye-only points can be seen next to occluding contours to their immediate right (see Nakayama and Shimojo 1990; Shimojo and Nakayama 1990).

One might ask the following question. Given that such pairing is ubiquitous in everyday life, what would happen if we were able to insert a few unpaired points in an otherwise identical pair of images? Would this call

forth the perception of an illusory subjective contour, and would such contours assume the exact placement dictated by the geometrical considerations just outlined? With these general considerations in mind, we created a stereogram that, although it contains no binocular disparity, is able to create the impression of a sharply defined occluding surface in depth (see Figure 1.16). To understand what is occurring in this stereogram, refer to Figure 1.17, which depicts the surfaces perceived in terms of the exact placement of the right-eye-only, left-eye-only, and binocular points.

Case A, shown in Figure 1.17a is a control condition. To view this case, one simply needs to fuse any of the identical images in the top row of Figure 1.16. Since the images are the same, all points are seen binocularly and there is no binocular disparity. Not surprisingly, one sees only a single flat surface in the picture plane, with no depth. Case B, the main DaVinci demonstration, is exactly the same as case A, except that four half-points have been removed from the binocular image. The physical-stimulus situation is explained in Figure 1.17b, showing the dots remaining. Note that there are two left-eye-only points (depicted by the open symbols) and two right-eye-only points (depicted by the gray symbols) in addition to the rest of the points, which are binocular. This pattern of binocular stimulation simulates a condition in which an invisible opaque surface is placed in front of a surface containing dots, a condition similar to the top view depicted in Figure 1.15a.

To view the DaVinci case the reader must fuse images on the bottom row of Figure 1.16, either cross fusing the left and center images or parallel fusing the center and right image. If fusion is successful, the perceptual consequences should be apparent and dramatic. One sees a phantom black square in the stereogram, the borders of which are exactly depicted in Figure 1.17b. The square, which appears even blacker than the background, lies in front of the rest of the pattern and is bounded by very sharp vertical edges. It is necessary to scrutinize the stereogram carefully to see that the relationship between the perceived square in front is exactly as depicted in Figure 1.17b. Thus, the left edge of the phantom occluding square is perceived to the immediate right of the left-eye-only unpaired points, and the right edge is perceived to the immediate left of right-eye-only unpaired points. When viewing the stereogram the exact placement of the phantom square and the unpaired points can be checked by alternately closing each eye.

Case C, the reversed-eye pattern, takes a little more exposure and practice but is well worth the effort. It simulates what is seen in Figure 1.15b. The proper stereoscopic stimulation can be accomplished by fusing the alternative pair of images in the bottom row of Figure 1.16. A methodological hint: Gazing directly at the location of the presumed window in Figure 1.16 may break fusion because the system may make an inappro-

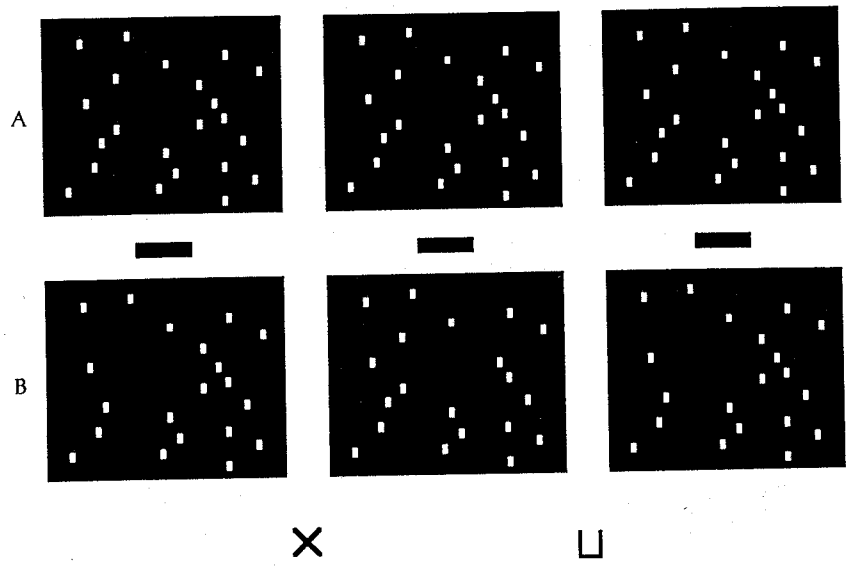

Figure 1.16
DaVinci stereopsis stereogram. (A) The control case: each half-image for binocular fusion is identical. As such the fused image should appear as flat. (B) The identical stimulus as in (A), except that four dots (two from each eye) have been removed. The observer sees a subjective square in front bounded by the unpaired points (as illustrated in Figures 1.15a and 1.17b). When fused in the reversed-eye configuration, observers see an aperture through which is seen a distant surface (as illustrated in Figures 1.15b and 1.17c).

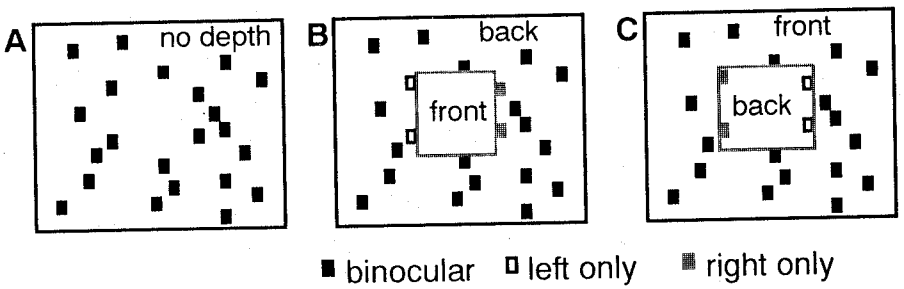

Figure 1.17
Explanation of the DaVinci stereogram, illustrating the relations between perceived surfaces in depth and the exact placement of paired and unpaired points shown in Figure 1.16. (A) There are no unpaired dots, and no depth is seen. In (B) two left-eye-only dots on the left side and two right-eye-only dots on the right side lead to perception of a central square in front. In (C) the configurations of the left-eye, right-eye stimuli are reversed, leading to the perception of a window through which one sees the unpaired dots in back. In all cases, left and right side of the occluding surface is bounded by right-eye-only and left-eye-only points, respectively.

priate convergence eye movement, attempting to fuse the two pairs of unpaired points. This tendency can be overcome by fixating on the horizontal bars between the upper and lower rows and attending to the area of interest.

The perceived configuration in relation to the physical stimulus is outlined in Figure 1.17c. In this situation, instead of a square, one sees a subjective window, revealing the unpaired points, which are now seen as far back, and define the edges of the window. Note that the window is significantly wider than the occluding square seen when the eyes are reversed.

We cannot overemphasize that there is a very specific and unvarying rule about where the subjective contour will lie. According to the diagrams in Figures 1.15 and 1.17, subjective contours should arise to the immediate left of right-eye-only points and to the immediate right of left-eye-only points. Careful examination of the lower stereograms in Figure 1.16 shows that this simple relationship holds for all four situations. This fact is easiest to appreciate when the subjective square is in front but is also apparent to those who can see a subjective window with the unpaired points in back. In each case, the position of the subjective contours in relation to the unpaired point is predictable and determinate and arises from the optical and geometrical constraints imposed by viewing a scene from different vantage points.

The main point is that these findings indicate clearly that vivid subjective contours can be created by information unavailable to conscious experience. We have no awareness of which eye is receiving the unpaired right-eye-only or left-eye-only stimulation, and we are unaware of the geometrical relations depicted in Figure 1.15; yet we see the results of our perceptual machinery—subjective occluding contours at very specific and predicted loci in the display. We believe this demonstration, in particular, lays to rest any view that subjective contours are the result of higher-order, nonvisual inferences. In Part 1.3, we make the point that such visual inferences occur very early in the visual pathway, perhaps as early as the striate cortex (area V1, as shown in Figure 1.2).

1.2 Experimental Studies

Phenomenology, the method used in the studies described so far, is often viewed suspiciously by those unfamiliar with its contributions. In part, this is due to the demand for an objective, not subjective, methodology in psychology and cognitive science. In part, it is due to worries about observer and/or experimenter bias and a lack of quantitative or statistical measurement. That said, however, one must also add that despite all these

seemingly valid misgivings, phenomenology survives and even flourishes among a small group of practitioners. Moreover, its results and conclusions often enjoy wide circulation in the scientific and lay community at large.

Why is this so? We discern several possible reasons. First, the results are actually much more objective than one might suppose. With well-crafted demonstrations, perceptual agreement between observers is actually far greater than that obtained in many psychological experiments, which often require statistical analyses of results from large numbers of subjects. Second, of course, is the immediacy and verifiability of the demonstration. All practitioners and interested parties can see the phenomenon for themselves and need not be concerned that the scientific reports are, as they sometimes are, mistaken. Third, the advent of good and cheap media technology, precise printing, and computer graphics technology enables many excellent demonstrations to be widely disseminated. Overall, phenomenology furnishes us with a surprisingly large, rich, personal (yet shared) data base from which to draw systematic theoretical conclusions.

There is, nevertheless, a great need for more objective verification of the sorts of phenomena we describe, not only to validate the method but also to reach out to other areas of knowledge, particularly the brain sciences. Because, for example, we cannot similarly characterize the visual perception of other species, we need to develop more objective experiments that do not rely on the subtle details of perception obtained from verbal reports. We cannot limit ourselves to phenomenology to obtain, for example, a satisfactory neurophysiological explanation of surface perception.

How then, do we convert a phenomenological observation of visual surfaces into one that can be substantiated by objective experiment, one that might also be conducted, if desired, on a laboratory animal? In the studies we describe here we use an indirect route. Instead of asking for a description of the experience of a surface as practiced in the section above, we ask an observer to perform a task we presume depends on surface encoding. In this way, we can evaluate the consequences of a surface representation without relying on the observer's subjective phenomenological judgment.

1.2.1 Two Views of Intermediate Visual Processing

Our goal, however, is more than simply the verification of phenomenological observations. The nature of our results allows us to challenge some widely held beliefs about intermediate visual processing and to replace them with an alternative conception. We argue that many seemingly early visual tasks are actually performed on a surface representation rather than

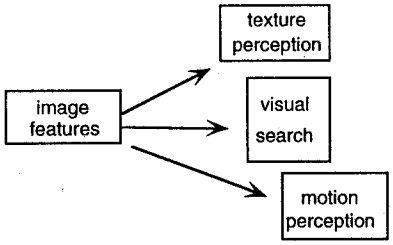

Figure 1.18
Current views regarding the dependence of rapid visual processing (texture perception, visual search, motion perception) on feature processing as mediated by receptive fields of cortical neurons in early cortical areas.

on the image. This position requires us to review briefly the widely held views we oppose. Some of the latter are described pictorially in Figure 1.18.

According to this way of thinking, there are a number of intermediate visual processes that do not require object knowledge, but can perform various rapid visual functions, including texture segregation, visual search, and motion perception. It is generally assumed that these functions operate at the level of simple features or filters in a retinotopic space. What characterizes this approach is the belief that the kinds of operations conventionally thought to be involved in the wiring of receptive fields are also likely to be explanatory in dealing with these perceptual functions.

To review this point, we need to step back and describe how vision scientists conceive of features and filtering and how these processes might be understood in terms of receptive fields. The basic form of the explanation proposed is extremely simple. Receptive fields of retinal ganglion cells, for example, can be understood if we simply assume that they are fed by two classes of convergent yet antagonistic inputs that are spatially delineated: an excitatory center and an inhibitory surround. Light falling on a center region alone will excite the cell. Light falling on the center and surround region, however, will not. We can conceive of these cells as sensitive to local differences in luminance or, more technically, contrast. A similar straightforward convergence is assumed to explain the properties of cortical cells. Converging and excitatory inputs from only correctly located ganglion cells would provide a cortical cell with orientation selectivity (Hubel and Wiesel 1959). Other simple schemes can account for motion sensitivity (Barlow and Levick 1965), as well as more complex receptive field specification such as end stopping (Hubel and Wiesel 1965). As mentioned earlier, convergent input to a cell from similar receptive fields at slightly different offsets in the two eyes for different cells would

provide a system whereby stereoscopic depth could be coded by comparing the inputs to different sets of cells. From these general findings Barlow (1972) made a strong conjectural case showing how a system might plausibly code important properties of a visual scene.

1.2.2 Visual Search and Visual Texture Segregation

The psychological/perceptual functions of texture segregation and visual search have been similarly conceived, although at a somewhat higher level of complexity. Thus, for texture segregation, it is assumed that by an analogous summation, and then differencing, of the outputs of cells with oriented receptive fields, a later stage should be able to signal texture boundaries. A similar conception suggests how an odd target in a popout task is identified. A strong indicator that the basis for easy texture segregation and popout must be fairly primitive—and can perhaps be accounted for by these simple mechanisms—is seen by examining the relative ease with which a segregated figure emerges in Figure 1.19a where the texture is defined by differently oriented elements, in contrast to the greater difficulty of seeing the emergence of texture in Figure 1.19b, where the texture elements are defined by the letters *T* and *L*. Although each *L* is easily distinguishable from each *T* in Figure 1.19b, it is apparent that the difference is not sufficient for rapid texture discrimination.

This distinction between simple and more complex features is also apparent in experiments on visual search (see Figure 1.20). Here the observer is asked to find the odd target among distractors, with the number of distractors varied. With only a few distractors, search reaction times for the two target/distractor types are comparable. When many more distracting elements are added, however, performance is degraded only for the *L* among *T*s case, in which reaction times increase markedly. With the

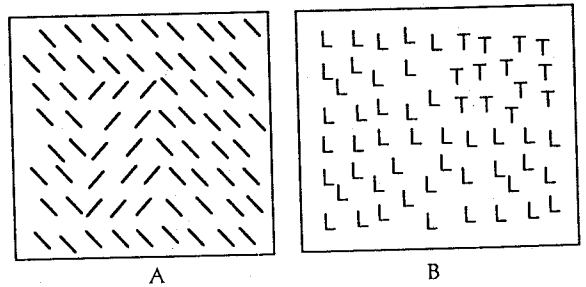

Figure 1.19
Texture segregation displays. (A) Texture difference determined by oriented lines. (B) Texture difference determined by different letters. Note that the emergence of a differently textured area is more prominent in A than in B.

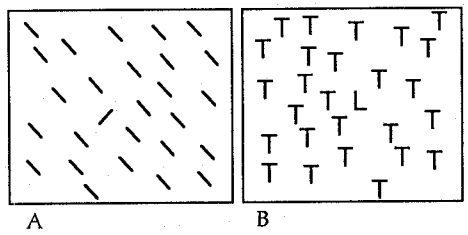

Figure 1.20
Visual Search. (A) For an odd orientation. (B) For an odd letter.

simple oriented lines no such increase in search time is apparent. Thus, performance seems to be independent of the number of distractors.

These findings on texture segregation and visual search lead to similar conclusions. The basis of rapid visual processing is presumed to lie in spatial primitives at a very rudimentary level of pattern recognition (Julesz 1986; Treisman 1982; Beck et al. 1983). Because of the orientation tuning of receptive fields and the oriented nature of stimuli that are easy to segregate, it might seem natural to see these receptive fields as prime candidates for mediating the very primitive type of pattern recognition required. A number of models of texture segregation and visual search make relatively appropriate predictions of a wide range of phenomenon (e.g., Malik and Perona 1990; Fogel and Sagi 1989). These models, for example, assume hypothetical units that pool the activity of classes of receptive-field types, then take differences in the outputs from these units, effectively obtaining differential excitation if a boundary exists between, say, regions of one orientation and another (as in Figure 1.19). Thus, rather than showing a sensitivity to simple luminance differences (as described for ganglion cells in the retina), these hypothetical units would be sensitive to differences in the density of particular texture elements, thus enabling a system to be selectively responsive to texture boundaries. Models of visual search suggest that a related mechanism can account for the emergence of the odd target among a field of distractors (Julesz 1986; Koch and Ullman 1985).

Even though such models explain much of the data described so far, they cannot explain the classes of phenomena we will describe below. Our motive in doing so is to suggest that surface representation is a necessary intermediate form of perceptual representation, one that forms an appropriate foundation for other visual functions—object recognition, object manipulation, and locomotion. We propose, as an alternative hypothesis to primitive receptive-field outputs, that perceptual function must funnel through a surface representation and that the most rapid visual functions we can measure must also pass through this required stage (as illustrated in

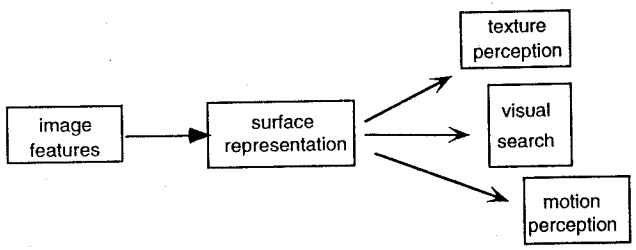

Figure 1.21
Our proposed view, showing that surface representation must precede such perceptual functions as texture perception, visual search, and visual motion (in contrast with the view outlined in Figure 1.18).

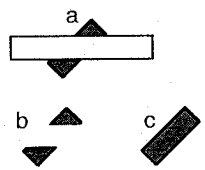

Figure 1.22
Occlusion and grouping: (A) scene in which gray region is seen as rectangle behind (as in C) and not as two disconnected fragments (as in B).

Figure 1.21). Thus the primitives that have been assumed to govern texture segregation—that is, receptive-field outputs—are not the ones responsible for the perceptual phenomenon under study.

We hypothesize that, because we cannot easily perceive the results of operations prior to the stage of surface representation, the latter is the first stage to which we have immediate access as perceivers (He and Nakayama 1992, 1994c). Such a hypothesis provides some strong predictions. It means that because of amodal completion (in accordance with Rules 1, 2, 3), the gray regions depicted in Figure 1.22a, instead of being encoded as two small separate polygonal figures (Figure 1.22b), are likely to be seen as an oriented rectangle in back (Figure 1.22c).

Such reasoning leads to the following question: What level of visual processing governs performance in rapid visual tasks presumed to be important for everyday vision? Is it the shapes of the image pieces themselves or the surface shape as defined by amodal or modal surface completion? Our surface hypothesis, of course, predicts that completed surfaces will be found most important. Earlier views of these processes, on the other hand, predict that the fragmented shapes of the image will dominate. To evaluate the merits of these competing hypotheses, our strategy was to

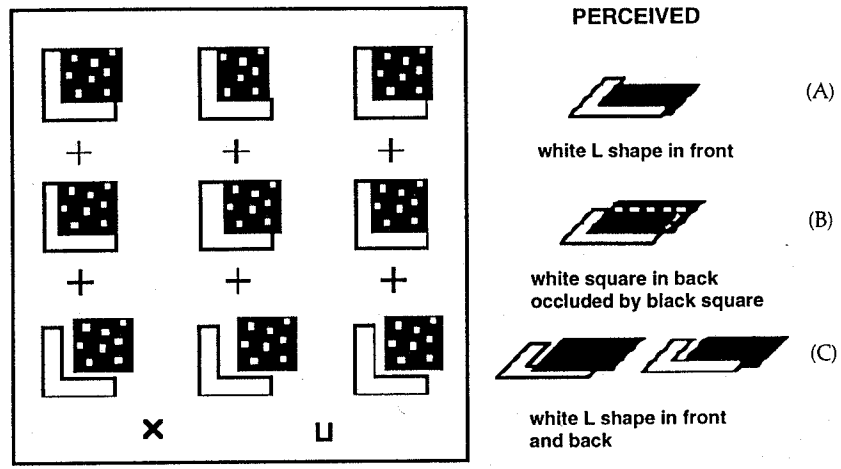

Figure 1.23
Elements used in a visual search display in which an observer is to find normal *L*s among reversed *L*s, or vice versa (reversed *L*s among normal *L*s). The individual L-shaped elements are presented adjacent to black textured squares and coded stereoscopically either in front of or in back of these squares. Each of the different conditions is depicted in one of three rows, labeled A, B, and C. A stereoscopic view of an individual element is presented in the box on the left, accompanied by a pictorial description of the perceived surface layout of each element on the right (under column labeled Perceived). (A) The *L* is in front of the squares and appears as an L. (B) The *L* appears in back and appears as part of a larger figure completing amodally behind. (C) Control condition, in which the L-shape is separated from the black square. Here the *L* is seen as an *L*, no matter what the relative depth of the black square or the *L*. (Modified by permission from Z. J. He and K. Nakayama, 1992.)

conduct experiments in which the stimulation of early cortical receptive fields is more or less unchanged but, by subtle shifts in binocular disparity, we altered depth relations in the display. This change in depth relation can lead, in turn, to the dramatic shifts in surface representation we described in the previous section. Our experiments show how this leads to a large difference in visual performance.

1.2.3 Surface Shape in Visual Search

We start by describing experiments on visual search (He and Nakayama 1992). In Figure 1.23 we show a stereogram of *L*s adjacent to a black textured square, which is present in all the displays. When the *L*s are in front, it should be clear, they look like *L*s (Figure 1.23a). When they are in back, however, they look very different (Figure 1.23b). Stereoscopic depth (in accordance with Rule 2) ensures that the border is owned by the square in front. As a consequence (and in accordance with Rule 3), the *L* becomes

part of an amodally completing surface, continuing behind the black squares. As such, it becomes less *L*-like and looks almost like a square in back.

If we set up a visual search experiment in which the observer was to find an *L* among reversed *L*s (or vice versa), we would expect that binocular disparity would have little effect on the result if simple image features are important in determining the outcome. If, on the other hand, completed surface shape is important, we would expect the visual search to become more difficult when the *L*s are behind. In this situation, both regular *L*s and reversed *L*s would become part of larger, more indistinguishable surfaces completing behind the rectangular occluder, each appearing as "almost a square." As a consequence, the inverted *L*s would become much less distinguishable from the regular *L*s in the visual search task.

This prediction is borne out by studies on search reaction times in which we varied the number of distractors for the *L*s-in-back versus the *L*s-in-front cases. In Figure 1.24, we show that when the *L*s are in front, search times are more or less constant for increasing numbers of distractors. For *L*s in back, however, it is very different. Search times increase dramatically with greater numbers of distractors.

One might argue, however, that it is easier to see targets when they are in front because there is a perceptual salience for closer targets. Control experiments in which small gaps are placed between the *L*s and the squares (as in Figure 1.23c) indicate that the fact of the *L*s being in front cannot alone account for the results shown in Figure 1.24 and that the involvement of surface completion is crucial (He and Nakayama 1992).

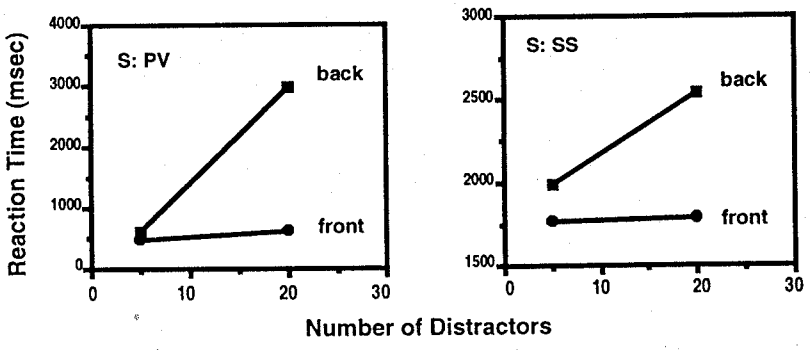

Figure 1.24
Reaction time to see the odd target (and *L* among reversed *L*s, or vice versa), showing dependence of distractor number for *L*s-in-front versus *L*s-in-back cases. Note that reaction times increase only for the *L*s-in-back case. (Reproduced by permission from Z. J. He and K. Nakayama, Surfaces versus features in visual search, 1992, *Nature* 359, 231–233.)

1.2.4 Surface Shape as Primitives for Texture Segregation?

As mentioned earlier, texture segregation is another area in which researchers have generally thought that performance is determined by differences in receptive-field outputs in early visual processing. According to this conception, primitive shape differences are sensed by such postulated mechanisms, and texture boundaries are computed automatically by the filtering properties of early cortical neurons.

Yet, our surface hypothesis might apply here as well. Perhaps it is not primitive shape, as determined by early receptive-field mechanisms, but surface shape, determined after the process of surface formation and surface completion. To test this hypothesis, we arranged an experiment in which the observer is presented with a very brief visual display followed by a mask. The observer's task is to report whether the differently textured region is a rectangle oriented horizontally or vertically (He and Nakayama 1994b).

The texture displays are similar to that shown in Figure 1.25, where the textured central rectangle differs from its background by being either *I*s among *L*s or vice versa. Here too the observer must report whether the

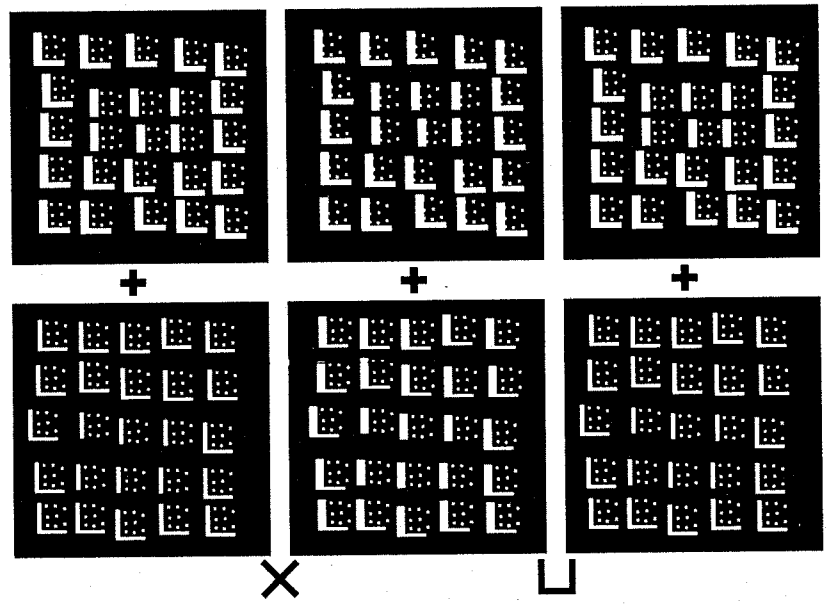

Figure 1.25
Stereograms showing texture segregation. Top row: texture elements are in front and texture segregation is easy. Bottom row: texture elements are in back, leading to amodal completion and more difficult texture segregation. (Reprinted by permission from Z. J. He and K. Nakayama, Perceiving textures: Beyond filtering, 1994, *Vision Research* 34, 151–162.)

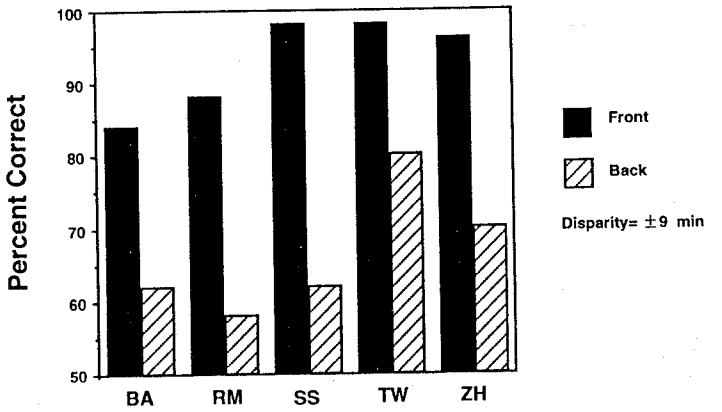

Figure 1.26
Percent correct identification of the orientation of a rectangular texture area defined by *Is* versus *Ls* (or vice versa). (Reproduced by permission from Z. J. He and K. Nakayama, Perceiving textures: Beyond filtering, 1994, *Vision Research* 34, 151–162.)

textured region is elongated horizontally or vertically. From the stereogram shown in this figure, it should be clear that it is much easier to discern the region of distinctive texture when the elements are in front than when they are in back. Compare the upper and lower stereogram. This difference is confirmed by the graph (Figure 1.26), which reports the percentage of correct scores of five observers for front versus the back cases.

In both visual search and visual-texture segregation experiments, we were able to change the surface representation so as to leave the image and, thus, the feature representation largely intact. This change in surface representation had a major effect. It was decisive in determining performance in very rapid visual tasks. This rapidity further underscores the importance of surface representation for immediate vision. It suggests that when we are confronted with an image under time constraints, we cannot respond to the shapes of the image fragments themselves. Our first impression is that of a surface representation.

1.2.5 Perception of Motion

Motion perception has generally been regarded as a fairly automatic and early visual function not dependent on higher-order visual input or top-down processing. As mentioned earlier, there are neurons selective to local motion in the striate and extrastriate cortices of primates, suggesting that at least some aspects of human motion perception are mediated by such cells (Hubel and Wiesel 1968; Britten et al. 1992). Yet there are a number

of indications that motion perception cannot be determined simply by the outputs of motion-sensitive neurons with localized receptive fields. One of these is the aperture problem; another is the phenomenon of long-range apparent motion. We consider each of them in turn.

The Aperture Problem

Figure 1.27 illustrates the aperture problem. In Cases A and B, we show two directions of motion taken by different elongated surfaces textured with diagonal lines. Despite the large differences in motion, vertical for Case A and horizontal for Case B, the motion is indistinguishable when considered locally. When informed about these very different global motions through a circular aperture, our visual system defaults and sees neither horizontal or vertical motion but a diagonal motion of stripes, its direction being orthogonal to the orientation of the local oblique contours (equivalent to that depicted in Case C). This perception indicates that local measurements of motion (as accomplished by orientation-selective, motion-sensitive neurons) by themselves are insufficient to specify true motion direction. Wallach's famous barber pole illusion (1935) shows that if we change the shape of the moving stimulus aperture, perceived direction of motion changes dramatically. If the aperture is oriented horizontally (as

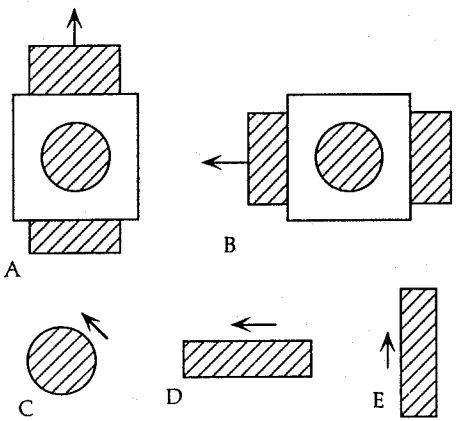

Figure 1.27
Moving oblique gratings viewed through various apertures. In (A) and (B) we show two very different motions of surfaces containing oblique lines, upward and leftward, respectively. Even though the true motion is different in each case, the local motion, as viewed through an aperture (as in C) and as it would be coded by motion-sensitive neurons, is the same, diagonal motion, up and to the left. (D and E) Wallach's barber pole illusion: perceived motion of the oblique lines is determined by the orientation of the elongated aperture.

in D), horizontal motion is seen to be moving horizontally. Similarly, vertical motion is seen in the vertical aperture (as in E).

Hildreth (1984), attempting to explain such perceptual phenomenon in terms of receptive-field-like entities, postulated a subsequent stage of analysis in which unambiguous motion of line terminators at the T-junctions (as in Figure 1.27) can propagate and overcome the ambiguity of such local motion (see also Nakayama and Silverman 1988; Yuille and Grzywacz 1988). According to this postulate, the "solution" to the aperture problem is an encapsulated one, "solved" exclusively within a motion module and operating only on a two-dimensional representation.

In line with our own understanding of visual surfaces, however, we analyze the problem very differently (Shimojo, Silverman, and Nakayama 1989). We ask whether the visual system regards the surface composed of stripes as continuing behind the aperture, in the same plane, or in front. From what we have said so far, one might expect that the moving stripes would be regarded as moving behind a rectangular aperture. This follows from the numerous T-junctions, which might indicate that the surface was behind and not bounded by the aperture. This depth cue, however, is in conflict with binocular disparity, which indicates that the diagonal stripes and the outline of the stripes are in the same depth plane.

Interestingly, if we look at the vertical barber pole illusion (as depicted in Case E, Figure 1.27) with only one eye; the bias toward motion along the aperture length is attenuated; that is, the illusion is weakened (Shimojo et al. 1989). This is not predicted by any receptive-field mechanism accompanied by velocity propagation from terminators. It can be explained at a surface level, however, when we realize that with monocular viewing the T-junctions denoting occlusion are no longer in conflict with the binocular cue of flatness. The surface itself is no longer seen as elongated but as boundless, appearing to extend beyond the aperture through which it is viewed. Not surprisingly, and for the same reason, the barber pole illusion is weakened further if we, by manipulating binocular disparity, arrange it so that the stripes are seen in back.

Figure 1.28 shows even more dramatically the importance of amodal surface completion behind occluders. In this experiment we changed the horizontal motion ordinarily seen in three horizontally oriented barber poles to vertical motion, simply by manipulating binocular disparity so that the configuration appears as a single, large vertical barber pole continuing behind nearer occluding stripes in front. This is accomplished by manipulating the binocular disparity of the two small stipled strips sandwiched between the three horizontal rows of oblique lines (Shimojo et al. 1989).

All these findings indicate that we cannot understand the perception of motion solely in terms of low-level motion signals. Even for the simple

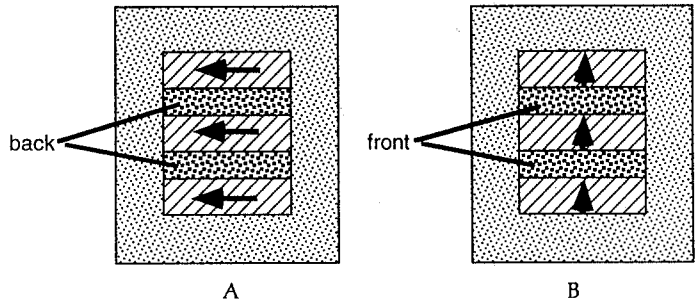

Figure 1.28
(A) Three small horizontal barber pole illusions showing movement to the left occurs if the stippled panels separating them are in back. (B) Putting the stippled panels in front, allows for the completion of all diagonal regions behind and leads to the perception of a large vertical rectangle. In this case, motion is seen as vertical. (Reprinted by permission from S. Shimojo, G. H. Silverman, and K. Nakayama, Occlusion and the solution to the aperture problem for motion, 1989, *Vision Research* 29, 619–626.)

coding of motion direction, the visual system needs information about the layout of surfaces in three-dimensional space.

Apparent Motion

The illusion of apparent motion was identified almost a hundred years ago by Gestalt psychologists (Wertheimer 1912). It occurs when stationary stimuli are flashed on and off in succession—one at time t_1, followed by the other at time t_2. This illusion is schematized in Figure 1.29.

It is interesting that the range of distances over which apparent motion can be seen is very large, spanning many degrees of visual angle. This poses problems for an explanation of motion perception in terms of motion-sensitive neurons in the visual cortex; such neurons are directionally selective but only over a very local area as small as a fraction of a degree in the striate cortex. Furthermore, the duration over which apparent motion is seen is long in terms of the measured properties of directionally selective cortical neurons. For these reasons, the processing of apparent motion has been classified separately from the processing of continuous motion and has been designated a long-range (as opposed to a short-range) motion process (Braddick 1974; Anstis 1980).

Still more interesting properties emerge with just the small addition of complexity to the usual apparent motion configuration. Ramachandran and Anstis (1983), for example, employed a 2 × 2 competitive-motion paradigm in which two pairs of stimuli occupying opposite corners of an imaginary rectangle flash alternately (see Figure 1.30). Note the potential ambiguity in this display. The element or token in, say, the upper-left

Figure 1.29
The simplest case of apparent motion. A time T_1 a small stationary square is flashed, followed at time T_2 by another flash.

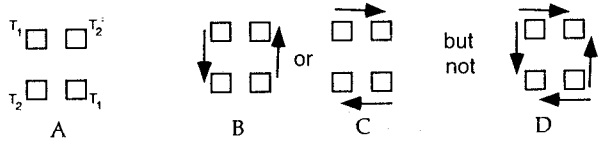

Figure 1.30
A. Bistable, competitive, apparent-motion paradigm where at time T_1 flashes at two stationary squares in opposite diagonal corners of an imaginary rectangle are followed at time T_2 by flashes at the complementary diagonal squares. Vertical (as in B) or horizontal motion (as in C) is perceived. Ambiguous motion (as in D) is not perceived.

corner flashing at time t_1 can be paired with other identical tokens flashing either at the upper-right or lower-left corners at time t_2. Despite this ambiguity, the perception of motion is pronounced. However it is also bistable. If the horizontal and the vertical distances are approximately the same, one sees either vertical or horizontal motion with approximately equal probability (as in Figure 1.30b, c). Surprisingly, we rarely, if ever, see a transitional motion perception in which, for example, the two targets split off to become the other two (as depicted in Figure 1.30d). This phenomenon illustrates what has been called the *correspondence problem* and its solution for apparent motion: Our visual system appears to make a binary decision, linking a token in frame 1 to a token in frame 2; there is no in-between solution or blending resolution.

What determines this correspondence? First, and most important in terms of establishing our experimental method, is the relative proximity between tokens. Correspondence is preferentially established between closer rather than distant tokens. Second is token shape. We consider each in turn.

The importance of token proximity can be easily demonstrated. If the relative vertical distance between tokens is decreased, motion will be predominantly vertical, whereas if the relative horizontal distance is decreased, horizontal motion will win out. Thus, if we keep vertical distance constant and gradually increase the horizontal distance in small steps, we can measure a motion-dominance function (see Figure 1.31) that summarizes the

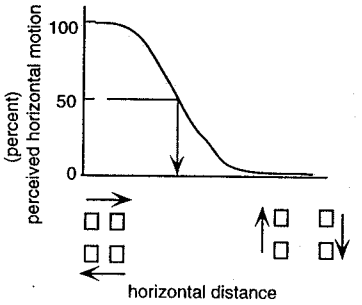

Figure 1.31
Motion-dominance function illustrating the proximity tendency for apparent motion. Keeping the vertical distance constant and increasing the horizontal distance decreases the likelihood of seeing horizontal motion. Horizontal affinity corresponds to the distance (marked by the arrow) where this function exceeds 50 percent (see text).

amount of horizontal motion seen as a function of horizontal distance. This function reflects the proximity tendency, which shows that matches are more likely to be made with nearest neighbors.

Less powerful but more pertinent for our immediate discussion is the role of shape in determining apparent-motion correspondence. Although shape matching is weak and can be easily overwhelmed by small differences in proximity, its existence is clearly revealed in a competitive-motion situation in which the various proximity tendencies between possible matches are more or less balanced. If, for example, we arrange our apparent-motion configuration so that one pair of identical shapes is presented sequentially in the top row alternately with a different pair of shapes in a bottom row, we see a preference for horizontal motion at intermediate horizontal distances, where the proximity tendency is more or less balanced for each possible match. In terms of the motion-dominance function seen in Figure 1.31, a bias toward matching identical tokens in the same horizontal row would shift the motion-dominance function to the right.

We have discussed the role of both shape and position in determining apparent motion correspondence strength, but we have not as yet linked these findings to the central theme of the chapter. In the context of visual surface representation, we need to define more precisely what is meant by *shape* and *position*. Is it shape as it might be defined in an image or image fragment or is it shape as defined after the processes of surface representation have been completed? Similarly, with position. Is it the position of the image patch narrowly defined, or after a surface representation has been established?

Shape Similarity in Apparent Motion

First, we deal with the issue of shape, showing that it is not image shape that determines correspondence but surface shape. To reveal its importance, we bias the competitive paradigm toward horizontal motion by increasing the relative similarity between potential horizontal matches, selecting the same shape between elements of the upper and lower pairs, respectively (see Figure 1.32). The upper row consists of oriented +45-degree bars, the lower row of oriented −45-degree bars. In each case, the pairs of white diagonal bars flash in opposite diagonal corners (as described earlier for Figure 1.30a), and flanking stationary nonflashing oblique rectangles are always present in all four positions.

Once again we use binocular disparity to manipulate the depth relations between the gray textured rectangle and the pairs of white bars. When the flashing white bars are in front no amodal completion between them can occur. They will be seen as two distinct diagonal bars and, because of the shape identity within a horizontal row, we expect to find the greatest horizontal affinity between the tokens, which should shift the motion-dominance function to the right. It should be very different, however, when binocular depth is reversed. Not only will the parallel bars be seen as behind but, more importantly, they will become part of a single surface

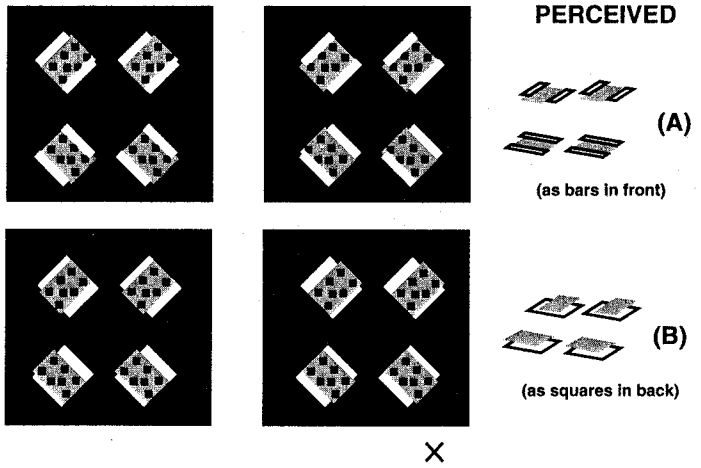

Figure 1.32
Stimulus showing that surface shape, not image shape, biases motion correspondence: flashing stimuli presented in the 2 × 2 paradigm. Note that the textured "occluders" are always present; only the parallel bars flash. (Stereograms use crossed fusion only.) (A) Parallel bars are coded in front and are therefore seen as parallel bars. (B) When parallel bars are coded in back, they are seen as parts of squares, completing behind.

completing *amodally* behind the rectangle, which now becomes an occluder. This shift, in turn, abolishes the preferential affinity between horizontal tokens because the white bars are now seen as part of completed squares in back.

This configuration should offer no opportunity for a shape bias and, as a consequence, less horizontal motion should be seen. The motion-dominance curve should shift to the left. Note that this particular expectation is not predicted by an image-based matching scheme, in which horizontal preference should be equal, whatever the depth relations. The results of this experiment are clear (see Figure 1.33). Greater horizontal motion bias is seen only for the bars-in-front case.

A final phenomenological observation confirms the critical role of surface encoding in a particularly revealing way. Because of the strength of the proximity tendency in relation to the weakness of the shape tendency, proximity can force vertical matches even if shape favors horizontal matches. For example, this can happen in the white-bars-in-front case, where the white bars are seen as distinct oriented bars. Because the orientation of the upper and lower tokens is different, the perceived-motion trajectory is no longer a simple vertical translation. The pair of bars is perceived both to rotate and to translate in the picture plan (as diagrammed in Figure 1.34a). If we think of what edges in each frame are

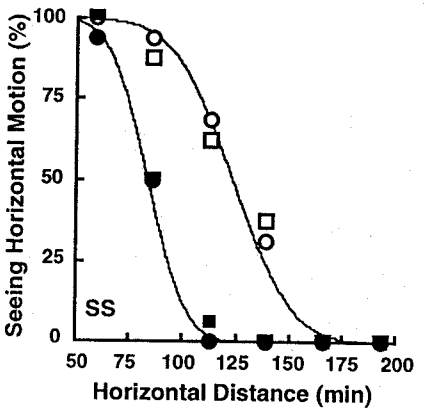

Figure 1.33
Surface shape, not image shape, biases motion correspondence (data from two observers). Open symbols refer to the case in which motion tokens are coded by binocular disparity to be in front. No amodal completion is expected; and, because the shapes match in the horizontal but not the vertical direction (as in Figure 1.32), there should be increased horizontal bias. Different times refer to two stimulus durations in the apparent-motion paradigm. (Reprinted by permission from Z. J. He and K. Nakayama, Surface shape not features determine motion correspondence, 1994, *Vision Research* 34, 2125–2136.)

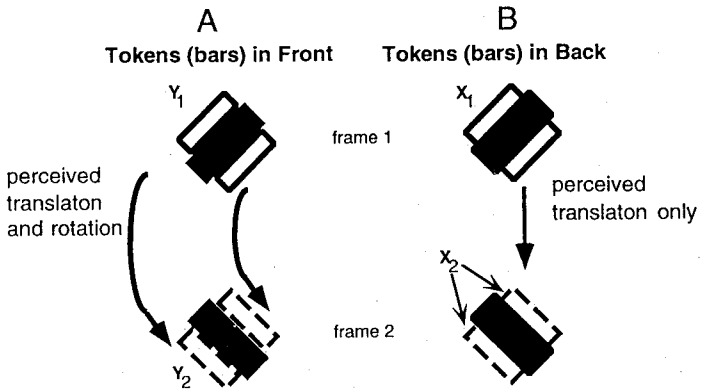

Figure 1.34
Phenomenology of rotational and vertical motion, showing that matches must be surface based, not image based (see text).

matched from one frame to the other, it is clear that the upper-left edge (labeled Y_1) of the oblique bar matches the lower-left edge of the orthogonally oriented bar in the next frame (labeled Y_2).

Contrast this to the case in which vertical matches are made in the bars-in-back case (as shown in Figure 1.34b) such that the bars are now seen as part of a square surface. Here the perceived motion is that of a simple vertical translation; no rotational motion component is seen. If we analyze what edges are matched, the answer is telling. Note the image edge X_1 labeled in Figure 1.34b for frame 1; there is *no* counterpart in the image after the apparent motion in frame 2. Edge matching only occurs at a level of surface representation between X_1, which has a visible counterpart, and X_2, much of which is an amodal contour hidden from view. This phenomenon alone argues strongly for matching at a surface level.

Surface Position Changes Mediated via Amodal "Leakage"

So far we have shown the importance of surface shape relative to image shape in determining correspondence in apparent motion. In this section, we address the issue of surface position, as opposed to image position, by selectively allowing a surface to amodally "leak" behind another (Shimojo and Nakayama 1990). Consider the seemingly innocuous visual situation, depicted in Figure 1.35a.

Frame 1 illustrates the same apparent-motion situation as before, but with an added feature—a large stationary rectangle (marked with an X) that can act as a potential occluder. We accentuate this role by altering its depth so that is perceived to be in front of the flashing tokens (illustrated

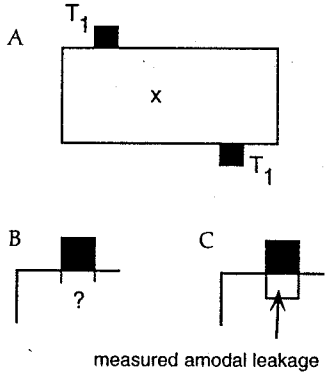

Figure 1.35
Amodal leakage. (A) Motion tokens in back are flashed adjacent to a stationary rectangle (marked *x*) in front. (B) Hypothetical amodal leakage of a flashing target behind an occluder. (C) Measured estimate of amodal leakage obtained from the motion competition paradigm, expressed as the size of an equivalent visible surface. (See Shimojo, and Nakayama 1990.)

as black tabs in Figure 1.35a). Consider the encoding of the small token in the upper left. Because of stereoscopic depth plus the T-junction, the border between the small token and the larger rectangle "belongs" to the large rectangle. This means, of course, that the bottom boundary of this flashing tab is essentially unbounded and thus has the potential to complete behind the occluder. However, there is no similar unbounded surface nearby to which it can link (as in the conditions shown earlier, e.g., in Figure 1.1). Yet it seems unlikely to think that the lower boundary of the tab stops exactly at the occluder. Might not the visual system infer that the tab continued for some short distance behind the occluder? If so, how far would it extend (see Figure 1.35b)?

If we confine ourselves to simple phenomenology, such a question seems very imprecise and uncomfortably subjective. We would be hard-pressed to accept our own answers, let alone those of others. Fortunately, from the perspective of motion correspondence, a precise answer can be obtained. The motion-dominance function introduced in Figure 1.31 indicates that horizontal motion perceived 50 percent of the time denotes an indifference point, one at which horizontal and vertical motion affinity is equivalent. We can therefore calculate the relative horizontal distance that yields this indifference. Predictably, the motion-dominance function shifts toward vertical motion when the central closer rectangle (marked X) is placed just along the edges of the flashing tokens. Consequently, the indifferent point moves toward shorter horizontal motion. From the

measured size of this shift, we can calculate the amount of amodal leakage behind the occluder and express it as the size of an equivalent visible surface that would cause such a shift, assuming a center of mass representation for a token position. This surface (as estimated from the data of six observers) is surprisingly large in relation to the size of the visible flashing tab. It is depicted as the white outlined area labeled amodal leakage in Figure 1.35c. (For further details, see Shimojo, Silverman, and Nakayama 1989.)

Taken together, these studies on motion indicate that apparent motion correspondence is dictated at a surface level of representation rather than one based on image shape or position. Why is this so? Our reasoning again rests on the computational problem posed by occlusion. Moving objects, no less than stationary objects, can be occluded by other objects. This means that motion-encoding schemes based on images alone are too unreliable; everyday perceptions of motion cannot be effectively mediated by the motion-sensitive neurons that respond to motion at the image level. As various parts of a surface become occluded or unoccluded, an image-based motion system would tend to sense spurious or nonrigid motion. For example, in examining the perception outlined in Figure 1.34b, an image-based matching system would perceive a rotary motion of individual bars; but because we see the two bars as part of a larger surface completing behind, such a spurious motion does not occur. Instead we see pure translational motion. The visual system codes, and we see, motion of a surface, not the motion of isolated image fragments.

1.2.6 Motion and Attention Dependent on Perceived Surfaces, not Three-Dimensional Geometry

In this final section reviewing experiments on surface perception, we address briefly a largely unexplored issue. Again, we challenge what we feel to be some implicity, yet wrongly held views, those concerning the nature of space perception and spatial representation. Because we have a two-dimensional retina and because we live in a three-dimensional world, many have seen the problem of space perception as the recovery of the third dimension.

As such, conventional studies of visual space perception start with a spatial description of our environment inherited from geometry, in particular coordinate or Euclidian geometry. Perhaps it seems especially rigorous and scientific to think of space in terms of the XYZ Cartesian axes and of space perception as the recovery of the Z dimension—usually via binocular disparity—with X and Y being supplied by the retinal image. Distance, according to this view, is represented by the length of a straight line joining two environmentally localized points.

Yet there are reasons to think that this is not the manner in which spatial distance is encoded in the visual system. Perceptual psychologist J. J. Gibson (1966) argues that space is not perceived in this way but in terms of the surfaces that fill space. The most important and ecologically relevant surface is the ground plane. In Gibson's view, Euclidian distances between arbitrary points in three-dimensional space are not biologically relevant (see also Nakayama 1994). We see our world in terms of surfaces and plan our actions accordingly. Locomotion (except for flying animals or airplanes) is usually confined to surfaces.

To begin to understand how distance might be encoded in the visual system and to evaluate the role of surfaces, we have exploited the proximity tendency in apparent motion (He and Nakayama 1994a). You will recall that the motion-dominance functions shown in figures 1.31 and 1.33 reflect a strong tendency for the visual system to make matches between tokens having greater proximity, that is, shorter distances. But, as pointed out above, the exact definition of distance in defining proximity has not yet been fully elaborated. This is particularly true if we think of potential motion tokens as occupying positions on perceived surfaces, not as arbitrary points in three-dimensional space.

If simple distances in space are important, we would expect increased matches between horizontal tokens as we introduce binocular disparity between the upper and lower motion tokens (as in Figure 1.36). As binocular disparity increases, perceived three-dimensional distance between upper

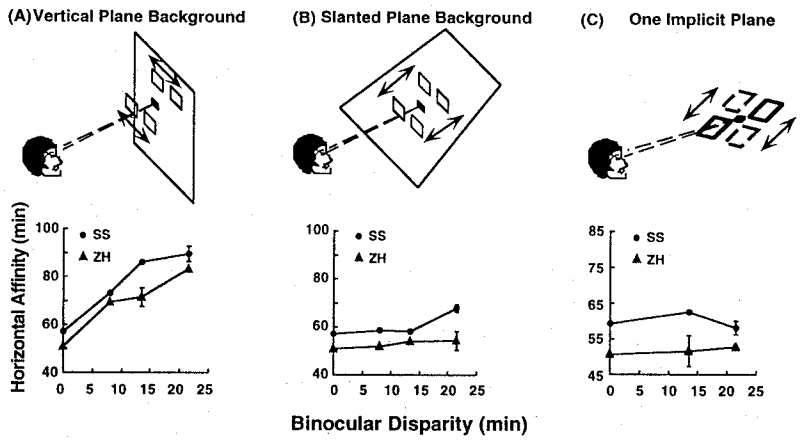

Figure 1.36
Apparent motion on different versus single plane receding in depth. (Reprinted by permission from Z. J. He and K. Nakayama, Apparent motion determined by surface layout not by disparity or 3-dimensional distance, 1994, *Nature* 367, 173–175.)

and lower tokens also increases; and by a proximity principle, motion matches should then be more prevalent between horizontal tokens. As Figure 1.36a illustrates, this is exactly what happens. The 50 percent point or indifferent point of the motion-dominance function (which we call the *horizontal affinity* because it reflects the strength of horizontal matches) increases with increasing binocular disparity. This experiment essentially replicates one originally reported by Green and Odum (1986), who demonstrated that matches were preferred between tokens having the same perceived depth.

Such an experiment does not, however, distinguish between this 3-D Cartesian view and one based on surfaces. Clearly, the outcome was predicted by a simple three-dimensional distance hypothesis. Yet, it is also the predicted outcome of a surface-binding hypothesis. If we hypothesize that perceived motion is preferentially bound to surfaces, it should be apparent that the two lower and the two upper tokens in Figure 1.36a define two implicit surfaces, which become increasingly distinguishable as binocular disparity is increased. If we also suppose that motion matches *within* a surface are preferred, we would also predict that horizontal matches would increase with increasing binocular disparity.

To differentiate a purely Cartesian depth hypothesis from our surface-binding hypothesis, we conducted two additional experiments. In each case, we varied the mean binocular separation between the upper and lower set of tokens, thus preserving the increase in perceived three-dimensional distances. However we also made subtle manipulations to accentuate the connections between the upper and lower tokens in terms of a surface representation. In the first case, we used exactly the same tokens as employed in the previous experiment, except that we added a stereoscopic receding plane composed of random dots upon which the tokens could "rest" (Figure 1.36b). In a second case, we increased the slant of each individual token so that if all four tokens were visible at the same time, they would be co-planar, lying in the same receding plane (Figure 1.36c).

We predicted that if motion is tied to surfaces rather than to three-dimensional depth per se, these two manipulations would greatly reduce the effect of binocular disparity—because such an increase would not be accompanied by a surface segregation. This is exactly what we found. The graphs in figures 1.36b and 1.36c show that binocular disparity in these situations does not increase the strength of horizontal matches. The results, therefore, emphasize the importance of surfaces. The preferential horizontal matches seen in the first experiment (shown in Figure 1.36a) were not due to increasing three-dimensional distance between vertically adjacent tokens. The same increase in three-dimensional distance had no effect when the increase in depth separation was accompanied by perception of a common surface.

Hypothesis: Motion is Tied to Surfaces Because Attention is Tied to Surfaces

In this section, we provide additional reasons to explain why motion is so closely tied to surfaces. We do so by advancing and testing a novel idea, namely that apparent motion is tied to surfaces because attention is also tied to surfaces. What is the basis of such a view? To start, let us return to the discussion of apparent motion and reexamine Figures 1.29 and 1.30. We noted there that apparent motion operates over a very large set of spatial intervals, comprising much larger distances than could be accounted for by known motion mechanisms in the striate and extrastriate cortex. Such neurons respond selectively to motion direction, but the distances between targets on successive frames are too small to explain apparent motion. For the central part of the visual field, they are approximately 0.5 degrees for striate cortex and from 2 to 4 degrees in area MT, the extrastriate cortical area specialized for motion. Apparent motion, however, can be seen over many tens of degrees.

Recently Cavanagh (1992) conducted an important set of experiments that strongly indicate that perceived motion is closely linked to attention. His findings show that if our attention is directed to one identifiable pattern and then to a similar one in a different position, we perceive apparent motion. Like the perception of motion obtained by following an actual moving target with eye movements, tracked attention provides the perceptual system with information about motion.

From Cavanagh's attentive motion, it is only a short step to the possibility that apparent motion is tied to surfaces because of its dependence on attention. We hypothesize that surfaces are also very important for the deployment of visual attention, arguing that attention cannot be arbitrarily directed to points or volumes in abstract space but is bound to perceived surfaces. Knowing that motion is so closely tied to attention helps explains why motion is also so closely bound to surfaces. To test such a view, however, we need to examine the deployment of attention in relation to surfaces more directly.

Our approach was to study directed focal attention in a cueing paradigm similar to that introduced by Posner (1980). The observer is presented with a cue at a site that is predictive of the target location in 80 percent of the trials (cue valid cases). The target appears at the uncued site on only 20 percent of the trials (cue invalid cases). We measured reaction times as a function of increasing binocular disparity separately for targets in the cue-valid and the cue-invalid trials. Our display was similar to that used in the apparent motion studies shown in Figure 1.36. Figure 1.37 shows the three stimulus conditions in which cued and uncued targets were presented: (A) in separate frontoparallel planes; (B) in separate frontoparallel planes resting on a common stereoscopic plane receding back;

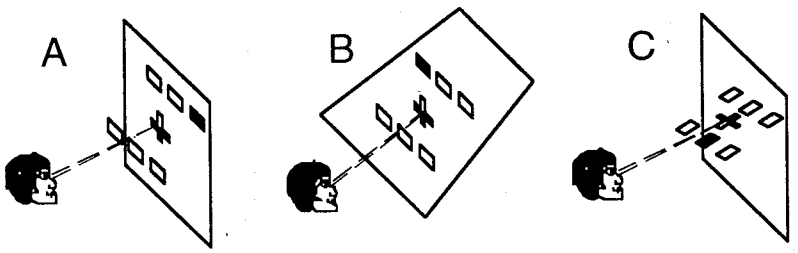

Figure 1.37
Cued attention to an upper or lower row examined as a function of stereoscopic depth separation. Three different configurations: (A) Tokens are frontoparallel seen against a background frontoparallel plane. (B) Tokens are frontoparallel and appear to rest on a receding plane, defined stereoscopically. (C) Tokens slant back according to the disparity difference in the upper and lower rows so that they form a single implicit plane, slanting back. Increased depth separation aids selective attention to a row only for case (A).

and (C) within a single stereoscopic plane receding back. In line with our attentional hypothesis, we predicted that only in condition A would there be an increased difference between cue-valid and cue-invalid cases as binocular disparity increased. In the other two conditions, we predicted, attentional focusing would not be as effective because attention would be automatically spread within surfaces (Case C) or spread evenly between separate surfaces lying on a common surface (Case B).

To begin the experiment, the observer fixated on a central cross flanked by an upper and lower row of three elements. At the start of each trial either an upper or lower limb of the cross would brighten, thus pointing to the upper or lower row of gray squares. The observer was instructed to attend to this cued row. Approximately one second later all six squares were presented colored either red or green—with five squares in one color and the remaining, target square in a different color. The task of the observer was to release a button when he or she saw the odd colored target. On 80 percent of the trials (cue-valid), the odd colored target was in the row that had been cued. On 20 percent of the trials (cue-invalid case), the odd colored target was in the other row. The strength of attentional focusing was determined by noting the difference in reaction times between the cue-invalid and cue-valid cases.

As predicted, in Case A we found that increasing disparity aids the observer to maintain attention on the cued row. But, as in the case of apparent motion, the result does not distinguish between a pure depth hypothesis for attentional segregation and a surface-binding hypothesis. This was why we added conditions B and C. In these two situations, although there is the same increase in binocular disparity, the stimuli are more closely related to a common surface; they either rest on a common surface or comprise one. As the surface hypothesis predicts, under these

conditions there is little difference in the ability of the observer to maintain attention as disparity increases.

This result shows that it is easier to confine attention to distinguishable surfaces than to confine one's attention to particular regions within a surface. This new idea, confirmed here, provides a mediating explanation for why apparent motion is confined to surfaces. Because attention is confined to surfaces, and because apparent motion is dictated by the mobility of attention (Cavanagh 1992), apparent motion is preferentially tied to surfaces.

1.2.7 The Perceptual and Phenomenological Primacy of Surfaces: A Critical Explicit Link?

In the first part of this chapter relying primarily on phenomenological observations, we described how small changes in binocular disparity can have dramatic effects on surface completion. In this second section we have basically confirmed these phenomenological observations by using objective methods that show very strong evidence of the processes underlying surface perception. Surface shape, not image shape, determines whether we see texture as segregated, whether single targets pop out of a display of distractors, whether motion is seen to conform to the aperture it is enclosed in, whether motion correspondence will occur, and so on. Surface properties rather than image properties are decisive. It appears that all higher visual processes must have, as a data format, a surface representation. We think it justified, therefore, to consider surface representation an indispensable link between low-level and higher-level vision.

Our perception cannot be conceptualized as a simple combination of image properties without understanding what specific visual entities must be coded. Because of occlusion, one of the prime candidates for explicit encoding is visual surfaces, stable, enduring aspects of the world that provide appropriate inputs for higher-order visual functions.

1.3 Possible Mechanisms of Visual Surface Representation

So far, we have mainly stressed the functional aspects of a surface representation, emphasizing the importance of surfaces in mediating very rapid visual processes and underscoring the idea that surface representation is a relatively primitive bottom-up process. Yet it is also one that appears to be governed by functional interactions not obviously related to the known properties of neuronal receptive fields. The coding of surfaces is better understood in terms of more macroscopic concepts: border ownership, depth, modal and amodal completion, and so on. To help bridge the gap between what appear to be qualitatively distinct levels of processing, we

would like to specify an anatomical locus, a probable cortical site where visual surface representation might begin.

1.3.1 Surface Representation May Begin as Early as the Striate Cortex (V1)

If we look at the diagram of the known extent of the visual brain, as shown in Figure 1.2, we note the large number of topographic maps, all of which are specialized for seeing. Where in this complex hierarchy of projections might surface representation begin?

A number of converging lines of evidence suggests that it must begin fairly early. Because surface representation seems to require little in the way of object-specific knowledge, it is likely to be antecedent to cortical areas in which object knowledge is stored. Therefore, it would probably not be postponed until, say, the infero-temporal cortex, the "what" or object-recognition system to Ungerleider and Mishkin (1982). In fact, we have argued elsewhere that it must be prior to object recognition to be of any use (Nakayama et al. 1989) as it must parse images into the appropriate surface units upon which object recognition can act.

The need for a visual surface representation in mediating rapid visual processes was outlines in Part 1.2. It plays a critical role in the perception of motion, the segregation of texture, and the processes required for rapid visual search. That such a broad range of functions not requiring object knowledge are so critically dependent on a surface representation argues strongly for an early rather than a late anatomical site for such processing.

The strongest piece of evidence that at least some part of surface representation must begin very early is the phenomenon of DaVinci stereopsis (Nakayama and Shimojo 1990; Anderson 1994). In our discussion of this phenomenon we noted that the placement of subjective contours and surfaces was critically dependent on which eye received the unpaired information. We found that right-eye-only points elicited subjective contours to their immediate left, whereas left-eye-only stimulation elicited subjective contours and surfaces to their immediate right (see Figures 1.16, 1.17). This indicates that critical aspects of surface perception are determined by very unusual sorts of information, of a class not generally available to us as conscious perceivers. To appreciate this fact, look around you with the right eye covered, then the left. No obvious difference exists in our perceptions unless there is some gross interocular anomaly. At a conscious level of perception, explicit eye-of-origin information is unavailable to us as perceivers. Interestingly, this information also appears to be hidden from most of the higher visual system as well. Cortical neurons from V2 and beyond respond more or less equally to stimulation delivered to one eye or the other (Maunsell and Newsome 1987; Burkhalter and Van Essen 1986). These neurons do not, therefore, carry eye-of-origin information. Only earlier, in the striate cortex, where the inputs from the two eyes are physically segregated into ocular-dominance columns, is explicit eye-of-origin information preserved (Hubel and Wiesel 1968). This suggests that cells in the striate cortex are the only ones available to signal the presence of subjective contours from unpaired points and inform us of whether an occluding contour lies to the left or the right of a given dot. The implications of this line of thought are potentially far-reaching. They mean that at least some aspects of surface representation must begin very early and must rely on information coming directly from cortical area V1, the striate cortex.

This very early cortical site for the beginnings of surface processing is also broadly consistent with findings that neurons in cortical area V2, the next stage of visual processing after V1, are responsive to subjective contours (von der Heydt et al. 1984, 1989). In these studies, receptive fields cells in area V2 were localized and an orientation preference determined (see Figure 1.38a for the position of a receptive field). It is significant that a stimulus similar to that shown in Figure 1.38b also excites V2 cells. These patterns elicit perception of a subjective contour yet have no luminance boundaries within the measured receptive field. As an important control, von Heydt and his colleagues showed that very small changes in the configuration (as shown in Figure 1.38c) abolish both the impression of an illusory contour and the neuronal response of these cells.

These two independent sources of evidence point to an explicit surface representation that is likely to begin somewhere in the neighborhood of cortical areas V1 and V2. In the broader picture of visual processing outlined in Figure 1.2, this suggests that visual surface representation is strategically placed just before the branching points of different functional visual streams—the "what" and "where" systems of Ungerleider and Mishkin (1982) or, alternatively, Goodale's framework for conscious perception versus visuomotor action (Chapter 5, this volume). Although we acknowledge the need for additional evidence, the view that surface processing is occurring at such early anatomical stages is appealing for a number of reasons.

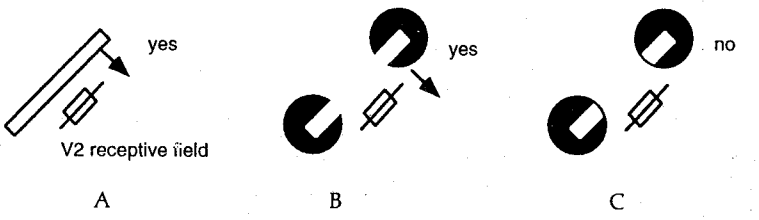

Figure 1.38
Schematic diagram showing cells responsive to illusory contours.

First, it is broadly consistent with the multiple higher-order functions just mentioned. All such functions are likely to need a visual surface representation as an input data format. This is evident in the area of object recognition (see Nakayama et al. 1990; Biederman 1987; Biederman, Chapter 4, this volume), but it is also likely for visuomotor control. Goodale's patient DF, for example, can reach for and grasp visual objects appropriately, even though she cannot report on their identity and shows little or no conscious awareness of their spatial properties. We suggest that DF may have preserved mechanisms of surface representation, adequate for the visuomotor processing. Second, surface processing at an early stage is consistent with the very broad range of findings reported in Part 1.2. We have shown that many primitive visual tasks, such as motion perception, visual search, and visual textural segregation depend on a surface representation. This dependency suggests that visual surface representation must be one of the earliest visual functions beyond simple coding of image properties by low-level cortical receptive fields. Third, such a view invites us to think more mechanistically. Areas V1 and V2 are some of the most completely characterized portions of the cortical visual system; their inputs and outputs are more clearly identified than any other cortical visual structure. Furthermore, in comparison to other cortical areas, the receptive-field properties of cells here have been well characterized.

Can we, therefore, begin to envision how these cells might account for the surface properties described here? Perhaps. Yet, impressed as we are with this evidence regarding cortical localization, it does not follow that a reductionistic understanding of surfaces in terms of neural circuitry is imminent. Localization is only the very first step in understanding a function mechanistically, that is, in terms of specific classes of neural connections. Several of the most difficult and challenging questions lie ahead.

First, there is no indication of how amodal completion of surfaces and contours behind occluders is encoded in the firing patterns of visual neurons. So far, the only connection we know of between neuronal properties and surface perception is through the phenomenon of modal completion, that is, subjective contours. Although this provides striking confirmation of our early visual system's ability to make important inferences about surfaces, subjective contours represent only a small fraction of the occasions on which we need surface-completion phenomena in our daily lives. Except for silhouettes and cases of very low illumination, the real boundaries of objects are almost always accompanied by physical "visible" luminance changes in the image. Such zero-contrast boundaries become even rarer for longer contours. Amodal completion, the completion of boundaries behind occluders, is much more common. There is rarely a scene in which the need for such completion is absent. Furthermore, the image distance over which such completions are necessary are often substantial, subtending many degrees of visual angle. Thus amodal completion, one of the most important aspects of surface completion, has as yet no known neurophysiological counterpart or correlate.

The second issue is border ownership. In thinking about surface completion, this has been an important concept; it determines whether or not surface fragments group and, if so, whether they do so in front of or in back of other surfaces. An attempt to account for at least part of surface completion in terms of end-stopped and binocular-disparity-specific cells was suggested in an earlier paper (Shimojo, Silverman and Nakayama 1989); this finding, however, accounts for only cases in which surfaces are covered or bounded by lines, not those created by general textures. It also is not explicit about how a boundary gets assigned to one region or another. As yet, we are lacking in a plausible neuronal explanation of how border belongingness is attached to a given image region.

What we are saying is that an as-yet-unbridged conceptual gap lies between the coding of image properties and the coding of surfaces. So, despite the success in relating some aspects of surface representation to, say, the striate cortex (through DaVinci stereopsis) and the important discovery that V2 neurons respond to subjective contours, a satisfactory scientific explanation of the coding of surfaces in terms of specific neural properties and neural circuitry remains elusive.

How then should we proceed? How can we begin to understand how image properties as measured by neuronal receptive fields are related to the more inferred representation of surfaces? Given the difficulty of the problem, it might be advantageous to step back, to think more broadly about how surface perception might emerge through development and the process of perceptual learning.

1.3.2 A View from Developmental Neurobiology: The Critical Role of Associative Learning

We mention recent work on visual development because it provides a strong argument for the importance of learning and plasticity at a cellular level. This work, which has emerged over the past ten to fifteen years, demonstrates the profound influence of neural activity in shaping neuronal connections, both prenatally and postnatally. Even the gross features of central nervous system topography, such as the lamination of geniculate nucleus and the ocular-dominance structure of visual cortex, are determined by activity-dependent cellular learning mechanisms. The interplay of Hebbian learning and the statistical pattern of correlation between neighboring afferent inputs from the two eyes to higher centers accounts for much of the observed gross structure and connections at a millimeter scale (see Kandel and Jessell 1991). The firing patterns of retinal ganglion

cells coding the same general direction of the visual field are correlated within the same eye but not between the two eyes. Research has shown that this correlation is decisive in forming the selective associations between the particular connections at the lateral geniculate nucleus and striate cortex, leading to the characteristic pattern of eye-of-origin lamination of the lateral geniculate nucleus and the ocular-dominance columns of the striate cortex (Stryker and Harris 1986; Shatz 1990). The results imply that, aside from the topographic maps established by the mechanisms of neuronal growth and guidance (Jessell 1991), the exact connections a given neuron makes with its neighbors is profoundly dependent on experience, that is, the past history of its inputs from other neurons. Similar mechanisms are also likely to be responsible for the formation of binocular connections needed for stereopsis (Hubel and Wiesel 1965), motion sensitivity (Daw and Wyatt 1976), and even the refined retino-topographic map itself (Schmidt 1985). The pervasiveness of learning at a cellular level to fashion the most dominant, well-documented connections of the visual cortex indicates that we need a similar understanding of the role of learning for other visual functions that develop through visual experience.

We also need, however, a conceptual framework for understanding which aspects of visual experience may be relevant—a means of identifying the visual and environmental events that must be functionally associated that is analogous to the statistical correlation of ganglion-cell discharges that occur prenatally. Because one of the most important challenges in understanding the visual coding of surfaces is establishing a relationship between image-based and surface-based representations, we start here. We hypothesize that the visual experience of the young, mobile observer, sampling images of surfaces from varied vantage points, provides the defining context within which to understand the learning of a visual surface representation.

1.3.3 Surface Transparency, a Proposed Example of the Associative Learning between an Image and a Surface

In Nakayama and Shimojo (1992), we developed a theoretical framework to explain the learning of a surface representation. Here, we condense this argument by resting our case on a single example, showing how analysis of the association between an image and a surface representation can explain an otherwise bizarre perceptual phenomenon, the emergence of perceived transparency in stereograms.

In the stereogram shown in Figure 1.39a, we created a stimulus made up of four repeats of a simple pattern of bipartite bars set against a black background frame. Each bipartite bar is divided into a gray and white region. Stereoscopic information is sparse, consisting of only two discrete

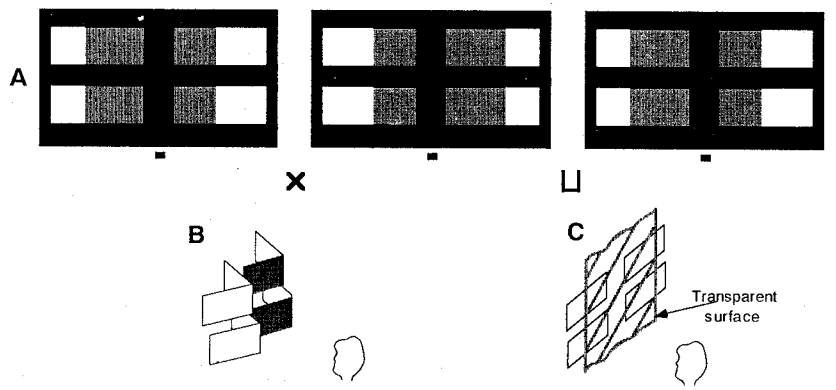

Figure 1.39
(A) Stereogram showing transparency. The gray-white border is coded in front; all other features are seen in back, in the plane of the frame. (B) Folded cards, the expected perception given depth interpolation. (C) Perceived transparency seen for the configuration in (A). (Modified by permission from K. Nakayama and S. Shimojo, Experiencing and perceiving visual surfaces, 1992, *Science* 257, 1357–1363.)

disparity values, front and back. The vertical contour dividing each of the bars is stereoscopically coded as in front. Every other contour, including that of the frame, is coded in back. It should be clear that only vertically oriented contours can supply binocular depth information in this stereogram. All other depth values, such as those along the horizontal contours or within the interior of the figure, are indeterminate. This occurs because there is no image variation in the horizontal direction from which to make a binocular match in the two eyes. Although the ordinary rules of stereopsis do not explicitly predict the depth of these indeterminate regions, we can justifiably assume that the perceived depth of these regions can be obtained by interpolation. The perception of sparse, yet textured stereograms with curved surfaces is consistent with this view. When observing such stereograms viewers see continuous curved surfaces interpolating appropriately in local regions where no texture exists (Julesz 1971; Ninio 1981).

Such an interpolation would predict a set of perceived surfaces like those in Figure 1.39b, a set of folded cards whose the convex edges face the observer. This configuration represents the simplest form of linear interpolation between points of defined binocular disparity. Surprisingly, this simplest interpretation is rarely, if ever, seen. Instead of seeing surfaces slanting in depth and connected at a fold, our perception is qualitatively different (Figure 1.39c). We see two disconnected sets of surfaces, one in front of the other, each frontoparallel with respect to the observer. More

striking, we also see a *material* change. The closer surface appears as transparent and partially occluding a white surface in back (Figure 1.39c). The perceived transparency is so potent that it makes our visual system see a filled surface over the whole transparent region perceived. Thus the gray transparent material actually appears to invade the black region and to be enclosed by a subjective contour that bounds the spreading. This is even more clearly seen in studies using colored stimuli, in which red areas spread into the otherwise black regions (Nakayama and Shimojo 1990; 1992). The importance of depth in eliciting transparency is clearly demonstrated when we view the stereogram in the reverse configuration, with the gray-white contour coded in back. Here we see no transparency, no color spreading, and no subjective contours. Instead, we perceive an opaque surface that is behind and seen through rectangular apertures.

Why, returning to the perception of transparency, does the visual system opt for what seems like an unusual interpretation, seeing a large global transparent surface instead of the folded cards? Why does the visual system avoid what would seem to arise naturally from depth interpolation? Our answer considers the question in terms of perceptual learning. We argue that through learning some critical feature of the binocular image shown in Figure 1.39a becomes associated with a transparent surface.

The essence of the general argument is simple. Those images most strongly associated with given surfaces determine which surfaces are perceived. What remains, then, is to develop a principled approach for estimating the association between images and surfaces. We take as a starting point the work on aspect graphs, a concept popular in machine vision pioneered by the mathematical insights of Koenderink and van Doorn (1976). In their seminal paper, Koenderink and van Doorn outline the characteristic pattern of topological stability and change of images sampled during shifts in viewer position.

Following their lead, we consider all of the various images (views) that can be associated with a given surface configuration as an observer takes all possible positions around a given surface. If we assume essentially random motions of an observer with respect to surfaces, then the determination of the probabilistic association between images and surfaces becomes an exercise in solid geometry. We need to estimate the volume in space from which a given images can be sampled. To illustrate this analysis, we consider the sampling of images from a familiar set of surfaces, a cube, from various positions in space.

To simplify the analysis we looked at potential vantage points in terms of a set of regions on a "viewing sphere" (see Figure 1.40a). The totality of such spheres of varying radii constitute all the possible vantage points that can be taken relative to a surface. It should be clear that a number of possible topologically defined classes of image can be sampled from positions on this sphere. From most vantage points, we see the image in the usual three-quarter, generic view, with three faces visible. We can also see it, however, from unusual vantage points so that we see accidental views in which just one or two faces are visible. Thus the vantage points that can give rise to such accidental views are very much more restricted than those giving rise to generic views. An image sample of just one face, is seen from just six vantage points or discrete loci on the sphere, as defined by the intersections of the circles. The number of vantage points yielding an image sampling from two faces is somewhat larger but still very limited, along a line defined by the three circles. From all other positions on the sphere, we obtain the generic view in which the image has three faces.

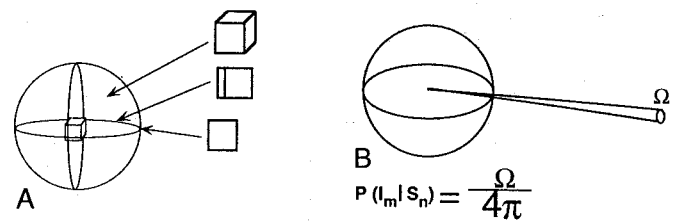

Figure 1.40
(A) Cube viewed from different positions on a "viewing sphere." Note that three topological classes of image can be sampled: one-faced, two-faced, and three-faced. One-faced and two-faced images can be sampled only from very restricted positions on the sphere, from points at the intersections of the circles and on the circles, respectively. Three-faced images can be sampled from all other positions on the viewing sphere. (B) Diagram indicating that the probability of seeing a given view from a given surface is related to the ratio of two solid angles.

If we assume random locomotion around the cube, the probability of an association between a surface representation and a given image reduces to the quotient of two solid angles. More formally, the conditional probability of obtaining a particular image I_m, given surface S_n, can be approximated by the following ratio:

$$p(I_m|S_n) = \Omega/4\pi \quad (1)$$

where Ω is the solid angle over which we can sample a particular image class, and 4π is the solid angle comprising the total set of vantage points from which the surface can be sampled. From this analysis, it should be clear that the probable association between image I_m, three faces, approaches unity as the distance from the cube increases. The probability of the other accidental images, two faces and one face, approaches zero.

In general, we can conceive of the totality visual experience (images) and sets of possible surfaces as depicted by the matrix in Figure 1.41. Each cell in the matrix represents the value of $p(I_m|S_n)$ that, as outlined above,

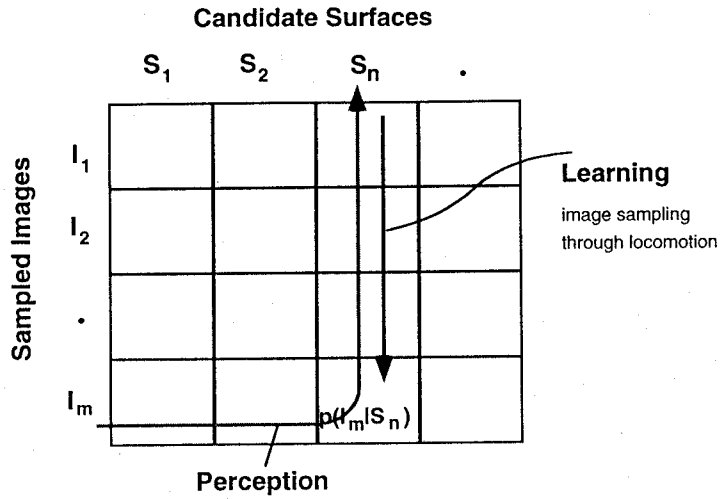

Figure 1.41
Generalized associative matrix denoting the probability of obtaining an image I_m, given surface S_n. We suggest that through locomotion—which places the viewer at random positions with respect to given surfaces—various images are sampled with determinate probabilities. This is shown as the conditional probability of sampling image I_m given surface S_n. We assume that these probabilities can be learned and represented in the connection strengths between an image and a surface representation. Learning of these probabilities through experience is expected to proceed along the downward arrow. Assuming that associative learning between various surfaces and images has occurred, the act of perception, denoted by the bent arrow connecting a particular I_m to a particular S_n, is hypothesized to depend the on strongest connection strength (conditional probability) for a given row, that is, for a given image.

can be plausibly estimated from geometry. It summarizes the visual experience of a mobile observer in terms of the images sampled from a surface. We hypothesize that these probabilities can be encoded in the nervous system as simple connection strengths between representations of images and representations of surfaces. Given these assumptions, the task for perception is clear. When confronted with an image, I_m, it must come up with the perceived surface representation, S_n, most closely associated with the image. In terms of the matrix, for any given row the perceptual system must find the cell having the highest associative strength [i.e., the highest $p(I_m S_n)$], which in turn defines the perceived surface. In Figure 1.41, therefore, we can envision the route from the image to perception. It is depicted as the bent arrow, starting from the image to the strongest connection for the image, thus pointing to the associated surface representation.

Let us apply this analysis to the cube, and consider what alternative surface interpretations might plausibly be evoked by various images of a

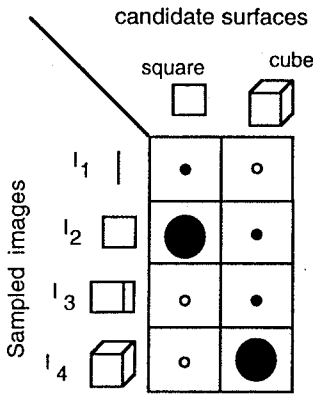

Figure 1.42
Associative matrix for cubes and squares with the same representation as in Figure 1.41, except that the probabilities of sampling are denoted by symbols (see text).

cube. For this purpose, we consider the exhaustive set of images that could arise from a cube and a square described in Figure 1.42. We illustrate the image sampling probabilities schematically: high probabilities or generic views are represented by the symbol ●; low probabilities or accidental views by •; and zero probabilities by o.

According to this scheme, when presented with, say, image I_4, we see the cube because it corresponds to the column having the largest associative strength. Interesting, when presented with image I_2, we do not see a cube, even though seeing a cube is compatible with this image class. Instead we see a square because of the greater associative strength between I_2 and the square.

We can now turn to the case of the stereogram shown in Figure 1.43 to explain why we see a transparent surface instead of the folded cards. First, consider again the relative rate of sampling of the various images from each surface type. In the case of the folded cards, the two classes of image that could be sampled, I_x (straight) and I_y (bent) are shown in Figure 1.43a. This diagram demonstrates that image I_y can be sampled from many vantage points; it is thus an generic view of a fold. The image presented in stereogram I_x is sampled only from very special vantage points, where the observer is at exactly the same height as the configuration. As such, I_x is an accidental view of the folded card and has very low probability. It is quite otherwise when sampling the same image I_x from a transparent occluding surface (Figure 1.43b). Any changes in the viewer's position, including motions up and down, preserves the same image. Image I_x is a generic view of this surface. Thus, for each of the associative strengths outlined in our image/surface matrix for the cube, we can see an analogous

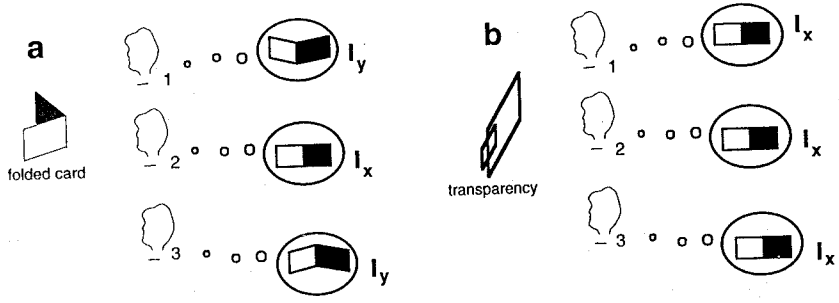

Figure 1.43
Generic and accidental views of candidate surfaces. (A) The sampling of image I_x from folded cards, which can occur only under a restricted set of vantage points at just the right height. (B) In contrast, the sampling of image I_x from the transparent surface can occur over a wide range of observer viewpoints.

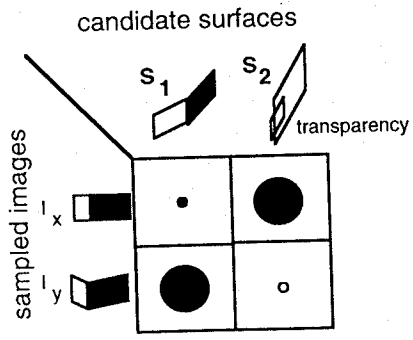

Figure 1.44
Associative matrix for folded cards and transparent surfaces.

situation for the folded card and the transparency. When confronted with image I_x, the visual system has the greatest associative strength in relation to the transparent surface, not the folded card. Consequently, image I_x as shown in Figure 1.44, elicits the perception of transparency, not a folded card.

Local Mechanisms of Inference: From Image Fragment to Surface Properties

So far, our analysis relating images to surfaces has consisted of extended images and surfaces (cubes, squares, folded cards, transparent surfaces, and so on). Although pedagogically clear, these examples are too complex to provide a credible way in which to envision how local image features might be used to build the fragments of a local surface representation. Yet such a simpler process might be expected, or at least hoped for. Early

cortical neurons, by virtue of their retinotopically organized receptive fields, analyze the image locally, providing information about the patterning of the image in a small, limited region of the visual field. It seems plausible that surface representation too might be built up by an inferential and mechanistic associative process similar to that outlined but at a more local level that links image fragments to surface properties.

How might such a local process occur? First, we can analyze the perceived transparency in Figure 1.39a at a slightly more microscopic level. A reasonable clue is the existence of critical T-junctions. Recall, however, that ordinary T-junctions accompany occluding contours, with the top of the T occluding the stem of the T. But our configuration does not contain the usual T-junction. Rather, the stem of the T, as dictated by binocular disparity, is coded as closer than the top of the T. This configuration is incompatible with the usual case, in which the top of the T occludes the stem. As we cannot appeal to the properties of ordinary T-junctions, therefore, we must analyze this configuration at a more essential level.

Figure 1.45, represents two stereoscopic image fragments, I_1 and I_2, from the image-sampling matrix outlined earlier. In both cases the vertical contour is coded binocularly as in front. It should be clear that while image fragment I_1 could have arisen from the folded convex corner, that likelihood is very low and requires an accidental vantage point. On the other hand, image fragment I_2, which does not contain any accidental collinear

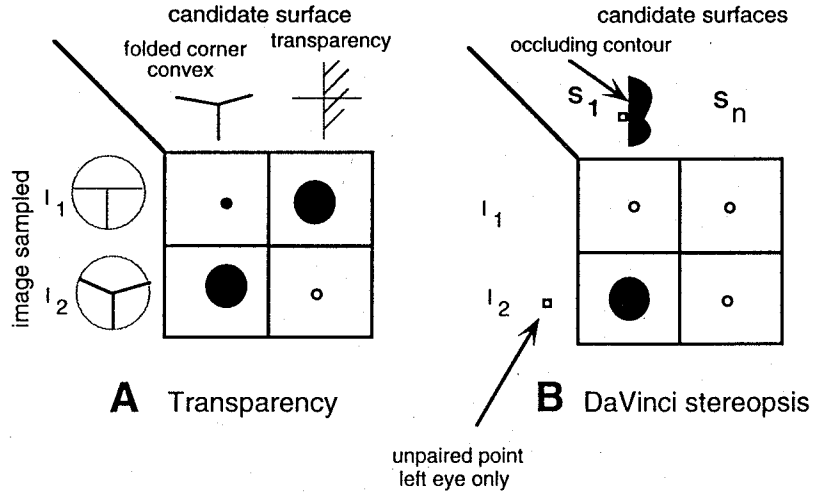

Figure 1.45
From image fragment to surface properties. Associative matrices explain transparency junctions (A) and subjective occluding contours from unpaired points (B).

lines, could have arisen from the corner from a very wide range of vantage points. Employing reasoning more or less identical to that used for whole surfaces, we can predict which image configurations will lead to the appropriate local surface properties; the system needs simply to make the connection of the image fragment to the most likely surface property. Thus, even at a very local level, when confronted with an image, say fragment I_1, the system, thanks to associative learning, is equipped to signal transparency, a surface property.

A similar analysis can be extended to DaVinci stereopsis. Here, we should note that image I_2, an unpaired left-eye-only point, is always paired with a closer occluding edge to its immediate right. Such a configuration containing unpaired points serves to reinforce the occluding status of the edge, that it is not, for example, a surface marking, i.e., paint on a surface. It is interesting to consider what happens when single unpaired dots are presented alone, as in Figure 1.16, illustrating DaVinci stereopsis. Here, we interpret the appearance of subjective contours as a ghostly pale reflection of this process of associative pairing.

1.3.4 The Generic View Principle: An As-If Heuristic

In the preceding section we sketched out the beginnings of a low-level mechanistic approach to surface perception, arguing that associative pairing of image fragments to a surface representation provides a plausible framework within which to understand relations between image data and a surface representation.

In this final section, we continue this argument, but more broadly. Our line of thinking has three parts. First, we state what is emerging across many disciplines as an important principle of visual analysis, the generic-view principle. Second, we argue that this principle can be conceived of as an automatic and passive consequence of the type of perceptual learning we have outlined. Third, we show that the generic-view principle has broad explanatory power, helping us understand at a deeper level some of the macroscopic concepts of surface representation we initially formulated as ad hoc rules.

Simply stated, the generic view principle asserts that *when faced with more than one surface interpretation of an image, the visual system assumes it is viewing the scene from a generic, not an accidental, vantage point* (Nakayama and Shimojo 1992).

Thinking back over our foregoing discussions of associative learning, we can see that the generic-view principle is a more formal and succinct assertion of the more extended arguments we have already made. By treating the probabilistic assertions we have made as discrete entities, the principle dichotomizes images or views into two categories. Stated in this

form, the principle is not new. It has been one of the core assumptions of machine-vision algorithms (Guzman 1969), as well as a key insight for Biederman's (1987) theory of human object recognition. However, with the exception of Rock's (1984) important work, it has rarely been formulated and applied as a general explanation to perceptual phenomena, to the human perception of depth and surface layout in simple scenes, both binocular and monocular (Nakayama and Shimojo 1990; 1992).

One of the advantages of the generic-view principle is that it codifies an approach to thinking about perception, providing a simple as-if rule stripped of any mechanism. As such it becomes a justifiable shorthand by which to apply the ideas just outlined without having to invoke a detailed discussion of probabilities, associations, and so forth. It is also becoming a more widely accepted view, independent of our own attempt to explain it in terms of perceptual learning (see, e.g., Freeman 1994; Albert and Hoffman 1995). To show some of its broad explanatory power, we pick two well-known but surprisingly misunderstood phenomenon: (1) the so-called impossible objects, and (2) the so-called figure-ground phenomenon.

First, consider what have been called impossible objects, objects that, shown pictorially, do not correspond to any real object we have ever encountered or could imagine. In Figure 1.46, we show the famous Penrose triangle (Penrose and Penrose 1958). At first glance, it appears to be a line drawing of a three-dimensional object, but it becomes almost immediately obvious that there is an inconsistency. We cannot conceive of how the various parts fit together as a real three-dimensional object.

An important face regarding this drawing is, however, underappreciated. The Penrose figure does *not* depict an impossible object but a real, *misperceived* object. Furthermore, it is misperceived in a way that provides strong support for the distinct level of surface representation we have

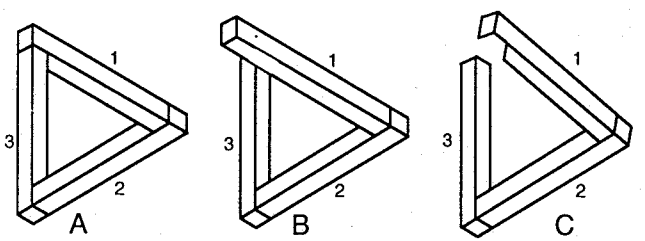

Figure 1.46
(A) The Penrose "impossible" triangle. (Redrawn by permission from L. S. Penrose and R. Penrose, Impossible triangles: A special type of illusion, 1958, *British Journal of Psychology*, 49, 31.) (B and C) Demonstration that the Penrose triangle is in fact a physically realizable object consisting of three bars that when skillfully notched (as in C) and viewed from an accidental vantage point, is perceived as an impossible triangle (A).

been advocating, an autonomous level of representation independent of a coding of known objects.

Gregory (1970) and later Ernst (1992) constructed such real objects, showing that, viewed from a specific accidental vantage point, they look like the Penrose "impossible" triangle. Each researcher produced a physical object, the tri-bar, and photographed it from various angles. The tri-bar is on display at various science museums, including the Exploratorium in San Francisco. It consists of an extended wooden figure in three dimensions with three arms or bars. For the sake of clarity, we show it first in a somewhat unfinished form (Figure 1.46b). The three arms are joined at approximately right angles and two of them extend in space. Arms 1, 2, and 3 are arranged so that each one is farther away from the viewer. Next, by using a precise woodworking techniques, the closer arm (1) was cut down to size and notched to form the three-dimensional object seen in Figure 1.46c. It is important for readers to understand that we are still discussing the same three-dimensional object and positioning it so that the arms have the same depth ordering as before. Readers must now imagine themselves as just slightly changing their vantage point, moving upward and to the right around the object, imagining that they are now viewing the tri-bar from an accidental vantage point that results in the image shown as shown in Figure 1.46a. Because of the skillful cut shown in Figure 1.46c, which permits the exact optical alignment of the near and far ends of the tri-bar from this vantage point, viewers will sample an image and see the so-called impossible triangle.

The results are the same and equally dramatic for viewers moving around the real tri-bar. Although they know full well that it is a three-dimensional object and not a triangle, when in the critical position, viewers see an impossible triangle. They cannot visualize the object in front of their eyes as anything else. Why do they see the impossible triangle and not bar 1 in front of bar 2, that is, the tri-bar in its correct spatial configuration?

We suggest that what is happening is perhaps the strongest, most powerful, and most dramatic example of the generic-view principle at work. Recall the principle again: When faced with more than one surface interpretation of an image, the visual system assumes it is viewing the scene from a generic, not an accidental, vantage point. Note that the new T-junction formed by the accidental alignment of arm 3 with the notch in arm 1 strongly indicates that that the lower surface of arm 1 is occluded by the front surface of arm 3 (Figure 1.46a). In accord with the generic-view principle, our visual system assumes that we are viewing this T-junction from a generic vantage point and thus recover the expected set of surfaces. What is so striking is that the generic-view principle is so strong at a local level that it recovers a surface representation of an image that is literally

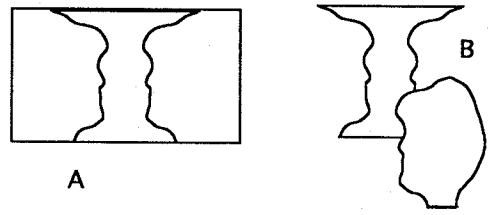

Figure 1.47
Face-vase revisited. (A) The classical face-vase illusion. (B) Same face, partially occluding the vase. Both figures are seen simultaneously with no difficulty despite the partial occlusion of the vase.

impossible from the point of view of object knowledge. Our surface processing is so powerful and autonomous that it generates an object we can't even conceive of. Furthermore, even when we have just seen and touched the tri-bar, our recent experience with and knowledge of the real object is of no help in resisting the generic-view interpretation.

This single example is the strongest piece of evidence in favor of a distinct level of visual representation independent of object knowledge or object representation. It provides the generic-view principle with a most unusual opportunity to demonstrate its predictive power.

The generic-view principle also shows its explanatory power in understanding another important perceptual demonstration, the figure-ground reversing configuration. When we presented the famous Rubin face-vase image in an earlier discussion, we invoked the need for an autonomous process at a pre-object-recognition level. Various portions of even unfamiliar figures can flip so that a surface that was the figure becomes the ground. The term *figure-ground*, in fact, gives the impression that there is a *figure* existing prior to recognition, a hypothetical proto-object.

It is possible that such a level does exist. Yet we should not be blind to the potential role of even more primitive and local processes of surface formation to explain the figure-ground reversal. In fact, we argue that local processes of border ownership and surface completion are more explanatory than processes at a hypothetical figural level. The generic-view principle again clarifies the important role of these more primitive processes.

To remind ourselves that we cannot so easily appeal to a prototypical figure to explain the figure-versus-ground reversal, in Figure 1.47b we show the same two components, but in the more usual, real-world arrangement. Here the face and the vase share a common border in the more common everyday case in which one partially occludes the other. In distinction to the ceaseless competition between the two in the Rubin configuration, we easily see the face and vase as concurrent figures, even when

part of one figure, the vase, completes amodally behind the face. Our claim is that the critical and obvious event in figure-ground reversal is not the reversal of objects or figures but the local reversal of border ownership. Each time we experience the perceptual flip, the common border between the two image regions changes. The reader should confirm this for selected instances of figure-ground reversals, recalling our Rule 1, which states that different surface patches cannot share a border and that the visual system must decide which region owns any common border between surface regions.

Here again, the generic-view principle provides strong explanatory power. Consider two separate surfaces at varying distances from an observer. What is the probability that a boundary between the two surfaces will be viewed in such a way that the edge of one surface coincides exactly with the edge of another? Even if they were to have the same shape (which is already unlikely), the chance that we would be just at the correct vantage point to see them as aligned is vanishingly small. Thus, the probability of two objects aligning to form a common border owned by both surfaces is essentially zero. The visual system does not assume an accidental alignment of surface boundaries in an image. This implies that borders between different surface regions cannot be shared, a strong rationale for the border ownership principle (Rule 1.)

1.4 Concluding Comments

We live in a three-dimensional world, full of objects resting on various surfaces. As visual creatures we rely on reflected light to obtain information from the world around us. Such reflections, of course, arise only from the boundaries of objects, the interface between various states of matter, usually solids and gases (air) but also water and air, water and solids. (Gibson 1966).

Thus, in general, surfaces constitute the only visually accessible aspect of our world. We cannot, for example, obtain visual information from the interior parts of ordinary objects. Yet even the surfaces of objects are not fully accessible to us as observers. Surfaces occlude other surfaces. Furthermore, "the amount of occlusion varies greatly, depending on seemingly arbitrary factors—the relative positions of the distant surface, the closer surface, the viewing eye. Yet, various aspects of visual perception remain remarkably unimpaired. Because animals, including ourselves, seem to see so well under such conditions and since this fact of occlusion is always with us, it would seem that many problems associated with occlusion would have been solved by visual systems throughout the course of evolution" (Nakayama and Shimojo 1990).

We argue that because of occlusion, the visual system is forced to adopt certain strategies for dealing with optical information. Most important, we claim, the visual system must develop as a prerequisite for further analysis a representation of the scene as a set of surfaces. Image-based representations, although indispensable first steps in capturing optical information, are insufficient bases for higher-level vision.

We have pursued a number of goals in the present chapter. First, we wished to establish the existence of a surface representation and to argue that it is a legitimate domain of inquiry and a definite stage of visual processing vital to other higher-order processing. Second, we sought to show some of the properties of this representation and to indicate that it seems to have rules distinct from those of neuronal receptive fields but also distinct from processes involved in the coding of familiar objects. Finally, and at a more mechanistic level, we were concerned with the issue of implementation, the possible manner by which surface properties might be derived from image information. Our goal here was to emphasize the powerful role of learning and to outline a possible low-level associative mechanism linking image samples to surface representations.

Appendix: Free Fusion of Stereograms without Optical Aids

Ever since the invention of the stereoscope by Wheatstone (1838), it has been recognized that binocular vision is important for depth perception. By finding a way for an observer to fuse two pictures, Wheatstone demonstrated that our brains are able to synthesize the perception of three dimensions from two flat images.

To appreciate how stereopsis works, we need first to understand binocular image sampling from visual scenes. Figure 1.48a shows a very simple diagram of physically realizable objects consisting of two rods, the left one closer than the right. By comparing left- and right-eye views, it is evident that the angular distance between the two rods is different in the two views, being smaller on the left. Wheatstone discovered that this difference, called binocular disparity, is sufficient to elicit the perception of depth and demonstrated it with his stereoscope (Figure 1.48b), which consists of two mirrors (labeled *m* and *n*) that supply the eyes separate images. An example of two such separate images are the center and right images in Figure 1.48c. The perceived depth relations of the two rods, which appear in front of the observer, is labeled by points 1 and 2 in Figure 1.48b.

More recently, large numbers of observers have been taught to fuse two separate images without a stereoscope by training them to misalign their eyes so that each views separate pictures. This has spawned a popular

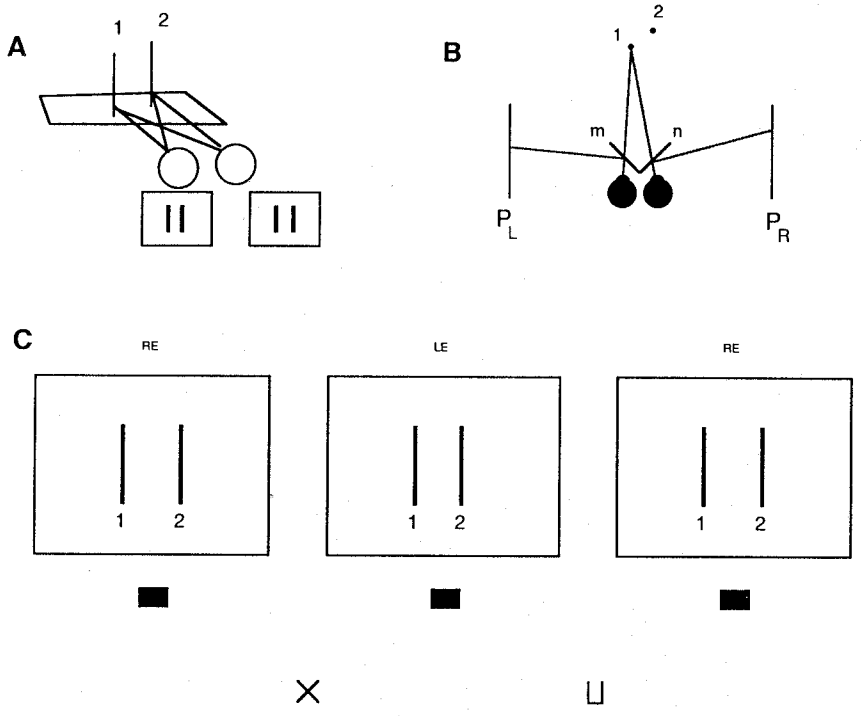

Figure 1.48
(A) Binocular viewing of a real-world scene showing two vertical rods, the left of which is closer than the right. Note that in the image of the two rods, the angular separation in the left eye is smaller than that on the right. (B) Schematic diagram of Wheatstone's mirror stereoscope, illustrating the position of the perceived rods (numbered 1 and 2) when viewing pictures P_L and P_R of the rods taken from two different vantage points. (C) Separate left-eye and right-eye images of the rods. Note that this is a three-panel stereogram in which the image to the left eye is presented in the middle and the image for the right eye is presented twice, at the extreme right and left. For this stereogram, the observer should cross fuse the left and center image or parallel fuse the center and right image (as described in Figure 1.49 and in text).

new art form in which observers peer into large pictures with repetitive patterns called autostereograms. To their surprise, they see otherwise invisible figures emerging in depth. A number of popular books using such displays is available. We recommend that readers who have difficulty fusing the stereograms in this chapter purchase several of these books for additional instruction and practice. Learning how to fuse stereograms sometimes takes a bit of practice. Like learning to ride a bicycle or to swim, some pick it up immediately, while others have a harder time. Once mastered, however, it is almost impossible to forget. If you know someone

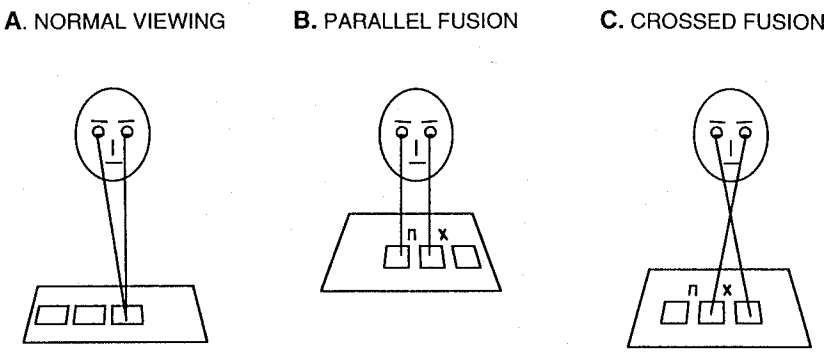

Figure 1.49
(A) Normal viewing of a book containing stereograms. (B) Parallel fusion of a stereogram, showing the viewing of the center and right images. (C) Cross fusion of the stereogram, showing the viewing of the left and center images.

who has mastered the technique, you may find it easier to learn with his or her help.

To help you understand the various techniques, we show the usual method of viewing a page (see Figure 1.49a). The eyes are not parallel but converge on the point of interest on the page. Each eye sees the same panel, and no stereopsis is obtained because there is no fusion of disparate images. There are two ways to misalign the eyes to obtain fusion. First, as shown in Figure 1.49b, we can diverge them abnormally in such way that the eye alignment is roughly parallel (parallel fusion). Alternatively, we can overconverge our eyes to obtain crossed fusion (Figure 1.49c).

When using the method of parallel fusion the observer imagines he or she is peering through the picture to a distant point beyond the picture. Doubling and blurring of vision is a good sign. Try not to focus the eyes immediately (to remove the blur) and concentrate on obtaining the perception of three images; the center image will eventually become the fused three-dimensional image. Eventually, the blur should disappear. To view these patterns (and to try the crossed fusion technique), experiment with holding the book closer and farther away or with removing or putting on your glasses. It is also useful to make sure the picture is viewed straight on (not at an oblique angle), at eye level, and in even illumination. Right-eye and left-eye images should be of the same brightness, and there should be no shadows on the page. For parallel fusion, most observers find that starting with close observation works better than viewing from a great distance.

Figure 1.49c shows the method for crossed fusion. This technique can usually be started by holding the book somewhat farther away than for

parallel fusion. Holding the tip of a pencil in front of the eyes, close each eye alternately and move the pencil tip so that it lines up with a visible feature on "crossed" pictures. The pencil tip should be inserted just where the lines of sight shown in Figure 1.49c cross. Thus, the right eye should be viewing an image to the left of the image presented to the left eye. It is most appropriate to try this on the two left-most images of the three-paneled stereogram shown as Figure 1.49c.

By now the reader has probably noticed that, with few exceptions, the stereograms in this chapter consist of three, instead of two pictures. This feature allows the observer to fuse them by either the crossed or parallel fusion method. For crossed fusion, the symbol **X** placed between the left and center image serves as a reminder that these are the two images to be fused. For parallel fusion, the center and right images should be fused. These two images are marked with the symbol **U**.

At times the text of this chapter asks the reader to view the reversed stereogram, or reversed-eye configuration. Under these circumstances, he or she should look at the wrong pattern, the one marked by the **U** for crossed fusion or **X** for uncrossed fusion.

For readers who have yet to master one of these two techniques, there is one additional method of viewing the stereograms in this chapter. One can purchase a very inexpensive stereo-viewer, which consists of two lenses (and prisms) that aid in attaining parallel fusion. Here, the observer should put the viewer over the image pair marked U and otherwise follow instructions.

Suggestions for Further Reading

The present chapter presents, in the spirit of this volume, a case history of scientific inquiry. Much of the research effort has rested on the technique of binocular stereopsis. Although studies in the past have used binocular vision, they have used it to understand the encoding of depth. Our endeavour, on the other hand, has been mainly concerned with the consequences rather than the causes of depth. We employ binocular vision only as a powerful tool to create depth. In interpreting our results, we were fortunate to be able to draw on a rich tradition of well-written, important books on perception and vision. In particular we recommend the work of J. J. Gibson (1950; 1966; 1979) whose series of books outlines an evolving theoretical context for understanding visual perception. Strongly influenced by Gibson is David Marr (1982), who attempts to synthesize information from many fields, including visual psychophysics, visual physiology, and computer vision. Although the details of Marr's work need revision, his broad conception of the enterprise of understanding vision remains as a guideline for future work. Also important, is the book by Kanizsa (1979), who provides a particularly clear example of perceptual phenomenology in the service of a deeper inquiry into the representation of surfaces. Additional insight regarding perception and inference can be obtained by referring to Rock (1984), Gregory (1970), and Hochberg (1978). We also recommend the work of Julesz (1971), a pioneer in the study of binocular stereopsis whose book is full or stimulating stereoscopic demonstrations and comments. For a wider exposure to current approaches to vision and perception, we also recommend basic textbooks, including those by Blake and Sekuler (1994), and Bruce and Green (1992).

Problems

1.1 What is wrong with problem-solving approach to the study of vision that presents visual perception as a problem to be solved by reasoning?
1.2 What is wrong with a physiological reductionistic approach to the study of visual perception? Comment on ways in which these problems may be overcome.
1.3 Argue for or against the statement that we do not "see" light but the surfaces of objects.
1.4 If you were a neurophysiologist, how might you find out whether neurons are coding attributes of surfaces?

References

Albert, Marc K., and Donald D. Hoffman (1995). Genericity in spatial vision. In D. Luce, ed., *Geometric representations of perceptual phenomena: Articles in honor of Tarow Indow's 70th birthday*, 95–112. New York: L. Erlbaum.

Allman, J. M., and J. H. Kaas (1974). The organization of the second visual area (VII) in the owl monkey: A second-order transformation of the visual hemifield. *Brain Research* 76, 247–265.

Anderson, B. L. (1994). The role of partial occlusion in stereopsis. *Nature* 367, 365–368.

Anstis, S. (1980). The perception of apparent movement. *Philosophical Transactions of the Royal Society of London* B290, 153–168.

Barlow, H. B. (1953). Summation and inhibition in the frog's retina. *J. Physiol.*, 119, 69–88.

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception* 1, 371–394.

Barlow, H. B., C. Blakemore, and J. D. Pettigrew (1967). The neural mechanism of binocular depth discrimination. *Journal of Physiology* 193, 327–342.

Barlow, H. B., and W. R. Levick (1965). The mechanism of directionally selective units in rabbit's retina. *Journal of Physiology* 178, 477–504.

Beck, J., K. Prazdny, and A. Rosenfeld (1983). A theory of textural segmentation. In J. Beck and A. Rosenfeld, eds., *Human and machine vision*. New York: Academic Press.

Biederman, I. (1987). Recognition by components: A theory of human image understanding. *Psychological Review* 94, 115–117.

Blake, R., and R. Sekuler (1994). *Perception*. New York: McGraw-Hill.

Boring, E. G. (1942). *Sensation and perception in the history of experimental psychology*. New York: Appleton-Century-Crofts.

Braddick, O. (1974). A short-range process in apparent motion. *Vision Research* 14, 519–528.

Bregman, A. L. (1981). Asking the "what for" question in auditory perception. In M. Kubovy and J. R. Pomerantz, eds., *Perceptual organization*. Hillsdale, NJ: L. Erlbaum.

Britten K. H., M. N. Shadlen, W. T. Newsome, and J. A. Movshon (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience* 12, 4745–4765.

Bruce, V., and P. R. Green (1990). *Visual perception: Physiology, psychology, ecology*. Hillsdale, NJ: L. Erlbaum.

Burkhalter, A., D. J. Felleman, W. T. Newsome, and D. C. Van Essen (1986). Anatomical and physiological asymmetries related to visual areas V3 and VP in macaque extrastriate cortex. *Vision Res.*, 26, 63.

Cavanagh, P. (1992). Attention-based motion perception. *Science* 257, 1563–1565.

Crick, F., and C. Koch (1992). The problem of consciousness. *Scientific American* 267, 152–159.

Daw, N. W., and H. J. Wyatt (1976). Kittens reared in a uni-directional environment: Evidence for a critical period. *Journal of Physiology* 257, 155–170.

DeValois, R. L., H. Morgan, and D. M. Snodderly (1974). Psychophysical studies of monkey vision: III. Spatial luminance contrast sensitivity tests of macaque and human observers. *Vision Research* 14, 75–81.

Ernst, W. (1992). *The eye beguiled.* Cologne: Benedikt Taschen.

Fogel, I., and D. Sagi (1989). Gabor filters as texture discriminator. *Biological Cybernetics* 61, 103–113.

Freeman, W. T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature* 368, 542–545.

Gibson, J. J. (1950). *Perception of the visual world.* Boston: Houghton Mifflin.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Green, M. (1986). What determines correspondence strength in apparent motion? *Vision Research* 26, 599–607.

Green M., and J. V. Odom (1986). Correspondence matching in apparent motion: Evidence for three-dimensional spatial representation. *Science* 233, 1427–1429.

Gregory, R. L. (1970). *The intelligent eye.* New York: McGraw Hill.

Gregory, R. L. (1972). Cognitive contours. *Nature* 238, 51–52.

Guzman, A. (1969). Decomposition of a visual scene into three-dimensional bodies. In A. Graselli, ed., *Automatic interpretation and classification of images.* New York: Academic Press.

He, Z. J., and K. Nakayama (1992). Surfaces versus features in visual search. *Nature* 359, 231–233.

He, Z. J., and K. Nakayama (1994a). Apparent motion determined by surface layout not by disparity or 3-dimensional distance. *Nature* 367, 173–175.

He, Z. J., and K. Nakayama (1994b). Perceiving textures: Beyond filtering. *Vision Research* 34, 151–162.

He, Z. J., and K. Nakayama (1994c). Surface shape not features determines apparent motion correspondence. *Vision Research* 34, 2125–2136.

He, Z. J., and K. Nakayama. Deployment and spread of attention to perceived surfaces in 3-dimensional space. Proceedings National Academy of Science. (in press).

Helmholtz, H. von (1867). *Treatise on physiological optics,* Vol. III. Trans. from the 3rd German edition, J. P. C. Southall, ed. New York: Dover Publications, 1962. First published in the Handbuch der physiologischen Optik, 1867, Voss.

Hildreth, E. C. (1984). *The measurement of visual motion.* Cambridge, MA: MIT Press.

Hochberg, J. (1978). *Perception.* Engelwood Cliffs, NJ: Prentice-Hall.

Hubel, D. H., and T. N. Wiesel (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.,* 148, 574–591.

Hubel, D. H., and T. N. Wiesel (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) in the cat's visual cortex. *J. Neurophysiol.,* 148, 229–289.

Hubel, D. H., and T. N. Wiesel (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology* 195, 215–243.

Jessell, T. M. (1991). Cell migration and axon guidance. In E. R. Kandel, J. H. Schwartz, and T. M. Jessell, eds., *Principles of neural science.* New York: Elsevier.

Julesz, B. (1971). *Foundations of cyclopean perception.* Chicago: University of Chicago Press.

Julesz, B. (1986). Texton gradients: The texton theory revisited. *Biological Cybernetics* 54, 245–251.

Kandel, E. R., and T. Jessell (1991). Early experience and the fine tuning of synaptic connections. In E. R. Kandel, J. H. Schwartz, and T. M. Jessell, eds., *Principles of neural science.* New York: Elsevier.

Kanizsa, G. (1979). *Organization in vision: Essays on gestalt perception.* New York: Praeger.

Kellman, P. J., and T. F. Shipley (1991). A theory of visual interpolation in object perception. *Cognitive Psychology* 23, 141–221.

Koch, C., and S. Ullman (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology* 4, 219–227.

Koenderink, J. J., and A. J. van Doorn (1976). The singularities of the visual mapping. *Biol. Cybernetics,* 24, 51–59.

Lettvin, J. Y., H. R. Maturana, W. S. McCulloch, and W. H. Pitts (1959). What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers* 47, 1940–1951.

Malik, J., and P. Perona (1990). Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America* A7, 923–932.

Marr, D. (1982). *Vision.* New York: W. H. Freeman.

Maunsell, J. H. R., and W. T. Newsome (1987). Visual processing in monkey extrastriate cortex. *Annual Review in Neuroscience* 10, 363–401.

Metelli, F. (1974). The perception of transparency. *Scientific American* 230, 90–98.

Michotte, A. (1964). *La perception de la causalité.* Louvain: Publications Universitaires.

Nakayama, K. (1985). Biological image motion processing. A review. *Vision Research* 25, 625–660.

Nakayama, K. (1990). The iconic bottleneck and the tenuous link between early visual processing and perception. In C. Blakemore, ed., *Vision: Coding and efficiency.* Cambridge, Eng.: Cambridge University Press.

Nakayama, K. (1994) James J. Gibson—An appreciation. *Psychological Review* 101, 329–335.

Nakayama, K., S. Shimojo, and G. H. Silverman (1989). Stereoscopic depth: Its relation to image segmentation, grouping and the recognition of occluded objects. *Perception* 18, 55–68.

Nakayama, K., and S. Shimojo (1990). DaVinci stereopsis: Depth and subjective contours from unpaired monocular points. *Vision Research* 30, 1811–1825.

Nakayama, K., S. Shimojo, and V. S. Ramachandran (1990). Transparency: Relation to depth, subjective contours and color spreading. *Perception* 19, 497–513.

Nakayama, K., and S. Shimojo (1990). Toward a neural understanding of visual surface representation. In T. Sejnowski, E. R. Kandel, C. F. Stevens and J. D. Watson, eds., *The Brain* 55, 911–924. Cold Spring Harbor, NY: Cold Spring Harbor Symposium on Quantitative Biology.

Nakayama, K., and S. Shimojo (1992). Experiencing and perceiving visual surfaces. *Science* 257, 1357–1363.

Nakayama, K., and G. H. Silverman (1988). The aperture problem II: Spatial integration of velocity information along contours. *Vision Research* 28, 747–753.

Ninio, J. (1981). Random-curve stereograms: A flexible tool for the study of binocular vision. *Perception* 10, 403–410.

Penrose, L. S. and R. Penrose (1958). Impossible objects: A special type of illusion. *British Journal of Psychology* 49, 31.

Poggio, G. F., and B. Fischer (1977). Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *Journal of Neurophysiology* 40, 1392–1405.

Posner, M. I., C. R. R. Snyder, and B. J. Davidson (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General* 109, 160–174.

Ramachandran, V. S., and S. M. Anstis (1983). Perceptual organization in moving patterns. *Nature* 304, 529–531.

Rock, I. (1984). *The logic of perception.* Cambridge, MA: MIT Press.

Rubin, E. (1921). *Visuall wahrgenommene Figuren.* Copenhagen: Gylden Kalske Boghandel.

Sagi, D. (1990). Detection of an orientation singularity in gabor textures: Effect of signal density and spatial-frequency. *Vision Research* 30, 1377–1388.

Schmidt, J. T., and D. L. Edwards (1983). Activity sharpens the map during the regeneration of the retinotectal projection in goldfish. *Brain Res.,* 269, 29–39.

Shatz, C. J. (1990). Impulse activity and the patterning of connections during CNS development. *Neuron* 5, 745–756.

Shimojo, S., G. H. Silverman, and K. Nakayama (1989). Occlusion and the solution to the aperture problem for motion. *Vision Research* 29, 619–626.

Shimojo, S., and K. Nakayama (1990). Amodal presence of partially occluded surfaces determines apparent motion. *Perception* 19, 285–299.

Shimojo, S., and K. Nakayama (1990). Real world occlusion constraints and binocular rivalry interaction. *Vision Research* 30, 69–80.

Schmidt, J. T. (1985). Formation of retinotopic connections: Selective stabilization by an activity-dependent mechanism. *Cellular Molecular Neurobiology* 5, 65–84.

Stryker, M. P., and W. A. Harris (1986). Binocular impulse blockade prevents the formation of ocular dominance columns in cat visual cortex. *Journal of Neuroscience* 6, 2117–2133.

Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology: Human Perception and Performance* 8, 194–214.

Ungerleider, L. G., and M. Mishkin (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, eds., *Analysis of visual behavior*. Cambridge, MA: MIT Press.

von der Heydt, R., E. Peterhans, and G. Baumgartner (1984). Illusory contours and cortical neuron responses. *Science* 224, 1260–1261.

von der Heydt, R., E. Peterhans, and G. Baumgartner (1989). Mechanisms of contour perception in monkey visual cortex: I. Lines of pattern discontinuity. *Journal of Neuroscience* 9, 1731–1748.

Van Essen, D. C., D. J. Felleman, E. A. DeYoe, J. Olavarria, and J. Knierim (1990). Modular and hierarchical organization of extrastriate visual cortex in the macaque monkey. *Cold Spring Harbor Symposium of Quantitative Biology* 55, 679–696.

Wallach, H. (1935). Uber visuell wahrgenommene Bewegungrichtung. *Psychologische Forschung* 20, 325–380.

Wertheimer, M. (1912). Experimentelle Studien über das Sehen von Bewegung. *Zeitschrift für Psychologie* 61, 161–265.

Wheatstone, C. (1838). On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London* B128, 371.

Yuille, A. L., and N. M. Grzywacz (1988). A computational theory for the perception of coherent visual motion. *Nature* 333, 71–74.

Zeki, S. (1978). Functional specialization in the visual cortex of the rhesus monkey. *Nature* 274, 423–428.