The Problem of Depth Perception

Two of the three spatial dimensions of the environment are explicitly present in the 2-D images on the two retinae. But the third dimension—the distance of the surface from the observer, which is often called depth—is lost in the process of optical projection from surfaces in a 3-D world to the 2-D retinae. Once this information is lost, it can never be regained with absolute. But the fact that people are very good at perceiving their 3-D environment demonstrates that surfaces in depth can indeed be accurately recovered from 2-D images under the vast majority of naturally occurring circumstances.

There are actually two closely related problems that must be solved in perceiving the spatial arrangement of surfaces with respect to the observer. One is determining depth: the distance of the surface from the observer in the 3-D environment. The other is perceiving surface orientation: the slant and tilt of the surface with respect to the viewer's line of sight. Slant and tilt, although often used as synonyms, technically refer to two different parameters of surface orientation in depth. Slant refers to the size of the angle between the observer's line of sight and the surface normal (the virtual line sticking perpendicularly out of the surface at that point). The larger this angle, the greater the surface slant. In the circular gauge figures illustrated in Figure 1, slant corresponds to the elongation of the projected ellipses, greater elongation resulting from greater slant relative to the frontal plane (which is zero slant). Slant also corresponds to the length of the projection of the unit surface normal (the surface normal of unit length) onto the frontal plane, longer projections resulting from greater slants. Tilt refers to the direction of the depth gradient relative to the frontal plane. In Figure 1, tilt corresponds to the direction of the surface normal projected onto the frontal plane.

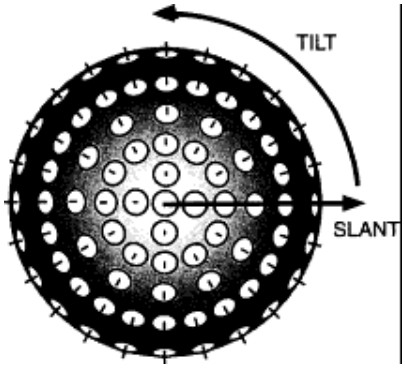

Figure 1 Slant and tilt of local surface patches on a sphere. The 3-D orientation of any local surface patch can be specified by two parameters: slant and tilt. Slant refers to the angle between the line of sight to the surface patch and its surface normal (the direction perpendicular to the surface), as indicated by the degree of elongation of the ellipses in this figure. Tilt refers to the direction of the surface normal projected onto the frontal plane.

This analysis of depth perception seems to lead to the conclusion that veridical depth perception is impossible. But how does this seemingly inescapable conclusion square with the fact that people consistently achieve accurate depth perception every minute of every day? The answer is that only infallible depth perception under all possible

circumstances is logically impossible. Since human depth perception is hardly infallible under all possible circumstances, there is no logical contradiction. As we will see, there are many conditions under which people are fooled into seeing depth inaccurately. You are already familiar with examples from everyday life, although you probably don't think about them as fooling you: Flat photographs portray depth relations quite convincingly, motion pictures do so even more compellingly, and so-called 3-D movies and virtual reality displays create depth perception of uncanny realism. All four are examples of your visual system being fooled because the resulting perception of depth is illusory; the depth that you so readily perceive arises from looking at images that are completely and utterly flat.

Equally important, however, is the fact that different depth cues generally converge on the same depth interpretation. It is this convergence of multiple sources of information that allows depth perception to be as accurate as it is. Only in the vision scientist's laboratory or under other conditions designed specifically to deceive the visual system do we regularly fail to apprehend the actual distance to environmental surfaces.

1 Marr's 2.5-D Sketch

How might visual information about the layout of surfaces in depth be represented? The most influential proposal to date has been David Marr's conception of the 2.5-D sketch (see Figure 2). As the name implies, the 2.5-D sketch is somewhere between the 2-D properties of an image-based representation and the 3-D properties of a true object-based representation. It summarizes the many converging outputs of different processes that recover information about the depth and orientation of local surface patches in the environment into a convenient representation of orientation at a distance.

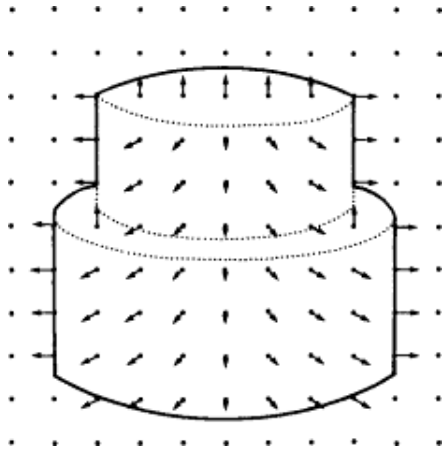

Figure 2 Marr's 2.5-D sketch. At each position in the visual field, a vector is shown (called the surface normal) that is perpendicular to the local patch of surface at that point and looks like a needle sticking out of the surface. The 2.5-D sketch also includes information about the distance along the line of sight.

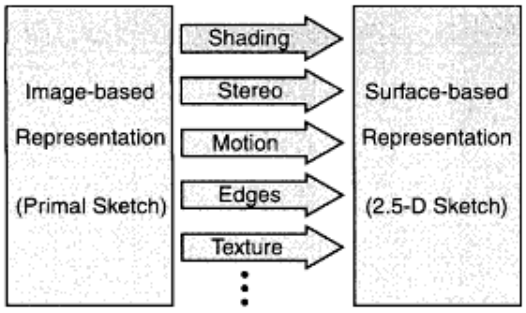

Figure 3 Flowchart of depth-processing modules. A 2.5-D surface-based representation is thought to be computed from a 2-D image-based representation by a set of parallel and quasi-independent processes that extract information about surface orientation and depth from a variety of sources, such as shading, stereo, and motion.

The view of surface recovery and depth processing that underlies the 2.5-D sketch is that there are many independent (or nearly independent) processing modules computing depth information from separate sources, as indicated in Figure 3. Each module processes a different kind of information, which then provides different constraints on the final common depth interpretation in the 2.5-D sketch. In the computer vision community, these processes are frequently referred to as the "shape-from-X modules," where X is the source of depth information, as in "shape-from-shading" or "shape-from-motion." Although this terminology is catchy, it is seriously misleading because shape is not explicitly represented in the output of such modules at all. What is represented is either depth or surface orientation, so it seems more appropriate to call them "depth-from-X" or "orientation-from-X" modules. The representation of shape is a great deal more complicated.

There are many different sources of depth information, as we shall see now.

2 Ocular Information

Ocular information about the distance to a fixated surface arises from factors that depend on the state of the eyes themselves and their various components. Of particular importance for depth perception are the focus of the lens (accommodation) and the angle of between the two eyes' lines of sight (convergence).

2.1 Accommodation

Accommodation is the process through which the muscles in the eye control the optical focus of the lens by temporarily changing its shape. It is a monocular depth cue because it is available from a single eye, even though it is also present when both eyes are used. The lens of the human eye becomes thin to focus light from faraway objects on the retina and thick to focus light from nearby ones (see Figure 4). If the visual system has information about the tension of the muscles that control the lens's shape, then it has information about the distance to the focused object.

Although accommodation is generally considered to be a weak source of depth information, experimental results indicate that people use it at close distances. Beyond 6-8 feet, however, accommodation provides little or no depth information. At this

distance the muscles that control the shape of the lens are already in their most relaxed state, so the lens cannot get any thinner.

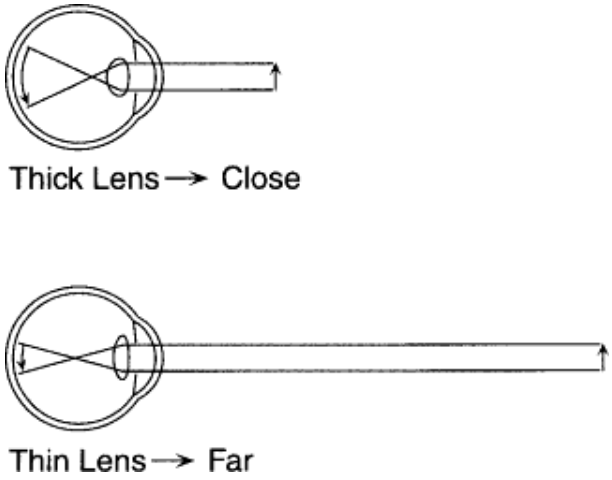

Figure 4 Depth information from lens accommodation. The lens of a human eye changes shape to focus the light from objects at different distances: thin for objects far away and thick for ones nearby.

Accommodation provides information about absolute depth. That is, it can specify the actual distance to the fixated object, provided the visual system is properly calibrated. Most optical depth cues merely provide information about relative distance, indicating which of two things is closer or the ratio of the distances of two objects. Such relative depth information is quite important, of course, but absolute depth information is required for people to estimate actual distances to environmental objects. If we couldn't determine actual distances, especially at close ranges, we would constantly stumble into things that were closer than we thought and reach for things that were out of range.

2.2 Convergence

The other ocular source of information about depth comes from eye convergence: the extent to which the two eyes are turned inward (toward each other) to fixate an object. The eyes **fixate** a given point in external space when both of them are aimed directly at the point so that light coming from it falls on the centers of both foveae simultaneously. Since each fovea has only one center, only one point can be precisely fixated at any moment (see Figure 5). The crucial fact about convergence that provides information about fixation depth is that the angle formed by the two lines of sight varies systematically with the distance between the observer and the fixated point. Fixating a close object results in a large convergence angle, and fixating a far object results in a small one, as illustrated in Figure 5. Because convergence depends on the observer using both eyes, it is a binocular source of depth information, unlike accommodation. Like accommodation, however, convergence provides information about the absolute distance to the fixated object.

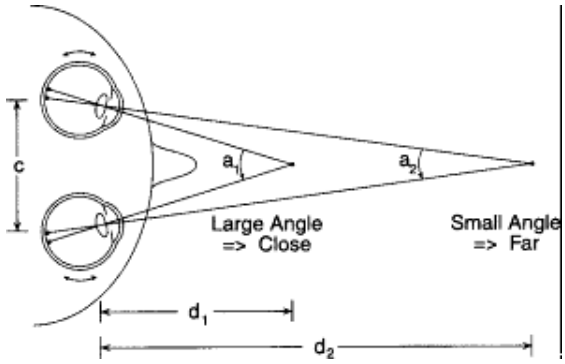

Figure 5 Depth information from eye convergence. The angle of convergence between the two eyes (a) varies with the distance to the object they fixate: smaller angles for objects far away (a_2) and larger angles for objects nearby (a_1).

The geometry of binocular fixation and convergence is illustrated in Figure 5. The equation relating distance to the angle of convergence is given in Figure 6, based on the trigonometry of right triangles. This equation shows that if the distance between the two eyes is known, the angle of eye convergence could be used by the human visual system to determine distance to the fixated point. But is it? Controlled experiments have shown that it is, but only up to a distance of a few meters. The reason for this limitation becomes apparent when the angle of convergence is plotted as a function of distance to the fixated object, as shown in Figure 6. At close distances the convergence angle changes rapidly for points that are only slightly different in depth, but at distances beyond about 6-8 feet, convergence changes very little as it approaches an asymptote (or limiting value) of zero degrees when the eyes are directed straight ahead to converge on a point at infinite distance.

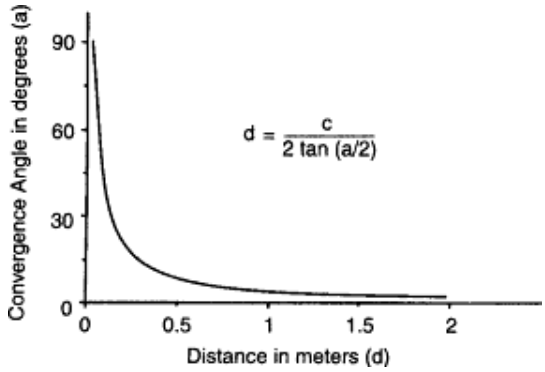

Figure 6 Convergence as a function of distance. The angle of convergence changes rapidly with distances up to a meter or two but very little after that.

3 Stereoscopic Information

Perhaps the most compelling experience of depth comes from **stereopsis**: the process of perceiving the relative distance to objects based on their lateral displacement in the two retinal images. Stereopsis is possible because we have two laterally separated eyes whose visual fields overlap in the central region of vision. Because the positions

of the eyes differ by a few inches, the two retinal images of most objects in the overlapping portion are slightly different. That is, the same point in the environment projects to locations on the left and right retinae that are displaced in a way that depends on how much closer or farther the point is from the fixation point. This relative lateral displacement is called **binocular disparity**.

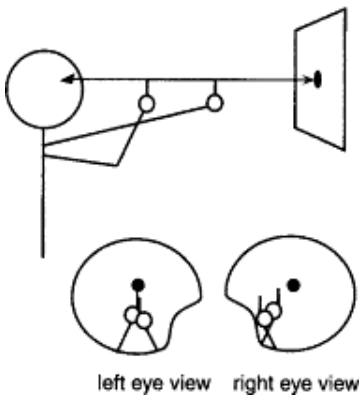

Figure 7 Demonstration of binocular disparity. The fact that the two eyes register somewhat different views of the world can be demonstrated by performing the finger experiments described in the text.

3.1 Binocular Disparity

You can experience the lateral disparity of corresponding points rather dramatically, however, in the following demonstration. First, hold up your left index finger at full arm's length and your right index finger at half arm's length. Then close your right eye, and align your two fingers using your left eye so that they both coincide with some distant point in the environment, as illustrated in Figure 7. Now, keeping your fingers in the same place and continuing to focus on the distant object, close your left eye and open your right. What happens to the images of your two fingers? They are no longer aligned either with the distant fixated point or with each other but are displaced markedly to the left. These differences occur because the few inches of separation between your eyes provide two slightly different views of the world.

The lateral displacement of your fingers relative to the distant object in this situation is a clear demonstration that binocular disparity exists. When such disparity is registered by your visual system in two simultaneously present retinal images, it is interpreted as your two fingers and the distant object being at different depths.

The information that binocular disparity provides is actually much more precise than we have yet suggested. The direction of disparity provides information about which points are closer and which are farther than the fixated point. The magnitude of this disparity provides information about how much closer or farther they are.

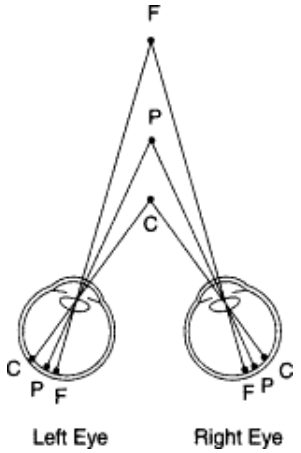

Figure 8 Crossed versus uncrossed binocular disparity. When a point P is fixated, closer points (such as C) are displaced outwardly in crossed disparity, whereas farther points (such as F) are displaced inwardly in uncrossed disparity.

For example, consider the diagram of two eyes fixated on a point, P, shown in Figure 8. By definition, point P falls on the foveae of both eyes and thus stimulates corresponding points. Now consider the projections of a closer point C while the eyes are still fixated on point P. As indicated in Figure 8, they do not fall on corresponding retinal points, since the one on the right retina is to the right of the fovea and the one on the left retina is to the left of the fovea. This outward direction is called crossed disparity (which you can remember because it begins with "c" just like "close"). Crossed disparity for the two images of a point (C) indicates that it is closer than the fixated point (P). How much closer it is depends on how far apart the disparate points are in the crossed direction. Now consider the point F which is farther than the fixated point P. This time, the one on the right image is to the left of the fovea and the one on the left image is to the right of the fovea. This inward (or nasal) direction is called uncrossed disparity. It indicates that the point that gave rise to it is farther away than the fixated point.

Stereograms. Perhaps the most powerful demonstration that binocular disparity can produce the experience of surfaces at different depths comes from stereograms: pairs of images that differ in the relative lateral displacement of elements such that, when viewed stereoscopically, they produce compelling illusions of depth from a completely flat page. Stereograms were invented by Charles Wheatstone when he analyzed the geometry of binocular disparity in 1838. Wheatstone realized that if the left and right eyes could be presented with images that differed only by the appropriate lateral displacement of otherwise identical objects, they should be experienced as located at different depths. Simple optical devices for viewing stereograms became fashionable in the late nineteenth century, and more sophisticated versions continue to be popular as children's toys. Stereograms can also be perceived in depth without any such apparatus, however, as we will now demonstrate.

The key feature of a stereogram is that corresponding objects in the left and right images are laterally displaced, producing binocular disparity. The direction of disparity (crossed or uncrossed) and the amount of disparity determine the depth that is perceived. Figure 9A shows one such stereo pair. Notice that, relative to the square

borders, the circles are displaced somewhat toward the outside, and the squares even more so. When these two images are stereoscopically fused by crossing the eyes, as described below, this lateral disparity produces a percept in depth like the display shown in Figure 9B. Figure 9C shows a pair with the reverse disparity, which produces the same percept, except reversed in depth, as depicted in Figure 9D. This result is precisely what we would predict given our previous discussion of crossed versus uncrossed disparity.

To experience these stereograms in depth, you must get your two eyes to register different images that your brain can fuse into a single image. There are two ways to do this: the crossed convergence method and the uncrossed convergence method. The images in Figure 9A and 9C have been constructed to produce the percepts illustrated below them using the crossed convergence method, which we will describe first. Viewing them using the uncrossed convergence method will reverse perceived depth, so that Figure 9A will look like Figure 9D, and Figure 9C will look like Figure 9B.

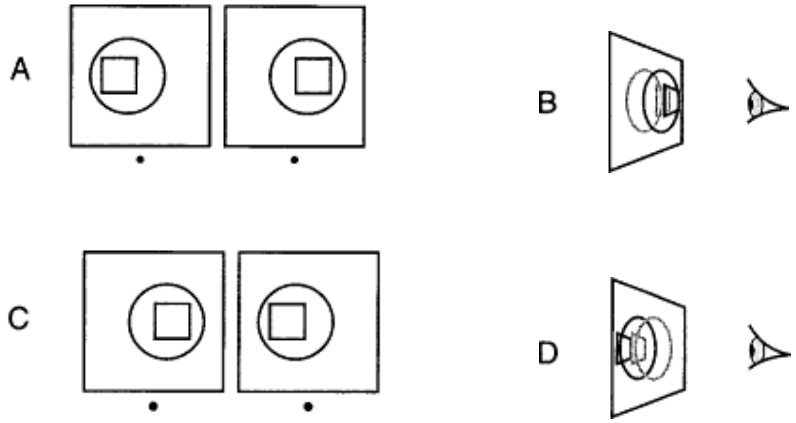

Figure 9 Binocular disparity and stereograms. If the two images in part A are stereoscopically fused with crossed convergence (see text for instructions), the circle and square will appear above the page as indicated in part B. If the two images in part C are cross-fused, the circle and square will appear behind the page, as indicated in part D.

To achieve stereoscopic fusion of a stereo pair by the crossed convergence method, you need to cross your eyes by just the right amount to bring the left image in your right eye into registration with the right image in your left eye. Because there are two physical images registered in each eye, you will generally experience four images once you misalign them by crossing your eyes. The goal in crossed convergence fusion is to adjust this misalignment so that there are exactly three images, the middle one of which is the alignment of two images that together produce the perception of depth.

Misaligning your eyes in just the right way is not easy to do without practice, but here is a useful trick. Focus your gaze on a pencil point positioned on the page midway between the two black dots below the two images. Then move the pencil toward your eyes, focusing continually on its tip. As you do so, you will experience diplopia (double images) for each half of the stereogram, producing a total of four images and

four dots. At some distance above the page, the borders of two of the four images will overlap exactly, and this is the point at which you should stop moving the pencil, whose point should now be aligned with the middle dot of three dots below the stereo images. (If your eyes are not horizontally aligned with the images on the page, they will not line up exactly, so you may need to tilt either your head or the book slightly as a final adjustment.) Once the images are exactly aligned, without moving your eyes, focus your attention on the middle image and look for the square and circle to float magically above the background border. This may take a few moments, but once you have achieved it, you will experience a vivid sense of depth.

To fuse stereograms using the uncrossed convergence method, begin by bringing the book very close to your face, with your nose touching the page. Relax your eyes—that is, do not try to focus on the page by crossing your eyes—and slowly move the book away from your face. Again you will see four images and four dots most of the time, but at some distance from the page, you will see exactly three images and three dots. Stop at this point and attend to the middle image, which will eventually appear in depth—if you have stereoscopic vision. Notice that the depth you perceive this way is the reverse of what you saw when you used the crossed convergence method.

In both methods, your eyes are misconverged for the distance to the page. In the crossed method, your eyes are overconverged, and in the uncrossed method, they are underconverged. These deviations from normal viewing convergence are what allow such stereograms to produce the illusion of depth.

3.2 The Correspondence Problem

Thus far we have talked about stereoscopic vision as though the only problem it poses for the visual system is to measure the direction and amount of disparity between corresponding image features in the two retinal images. But we have not yet come to grips with the far more difficult problem of determining which features in one retinal image correspond to which features in the other. This is called the correspondence problem for obvious reasons. We avoided it in our previous discussion because we followed light into the eyes from the environment, starting with a single environmental point and following its projections to distinct points in the left and right retinal images. In doing so, we assumed the proper correspondence of images in the two eyes. But the visual system faces the much more taxing inverse problem: It starts with two images and has to discover which features in the left image correspond to which features in the right one. How our visual systems determine this correspondence requires an explanation.

For many years, theorists assumed that this problem was solved by some sort of shape analysis that occurred before stereopsis. Shape was assumed to be analyzed first, separately for the left and right images, so that the results could be used to solve the correspondence problem. The rationale was that although it would be difficult to determine which point of light among thousands in one retinal image went with which point among thousands in the other, it would be easy to determine that, say, the tip of a German shepherd's nose in one retinal image went with the tip of a German shepherd's nose in the other. Ambiguity in the correspondence problem would therefore be enormously reduced if shape analysis came first. But does it? The

alternative possibility is that stereopsis might actually come first and occur without the benefit of monocular shape information.

Random Dot Stereograms. Bela Julesz, then working at Bell Telephone Laboratories, realized that he could test the shape-first theory of the correspondence problem by constructing what he called random dot stereograms. A random dot stereogram is a pair of images consisting of thousands of randomly placed dots whose lateral displacements produce convincing perception of depth when viewed stereoscopically so that one image stimulates one eye and the other image stimulates the other eye. Figure 10 shows an example of such a stereo pair that encodes a square floating above the page.

When each image of a random-dot stereogram is viewed by itself, the dots look random in the sense that no global shape information is present in either image alone. The shape-first theory of stereoscopic correspondence therefore predicts that it should be impossible to perceive depth by fusing random-dot images stereoscopically because it assumes that the correspondence must be based on recognized monocular shape information.

To test this prediction yourself, use the crossed convergence method described above to fuse them. It may take a while for the square to emerge in depth because random dot stereograms are quite a bit harder to see than standard stereograms, such as those in Figure 9. But once you succeed, a randomly speckled square will stand out clearly against a speckled background. (If you fuse Figure 10 using the uncrossed convergence method, the square will be perceived behind the page through a square hole in the background.) Since there are no monocular shapes to be matched in the two retinal images, the appropriate conclusion is that the shape-first theory is *incorrect*. The stereoscopic system seems to be able to solve the correspondence problem without monocular shape information because Julesz's random dot stereograms contain little, if any, such information. Another stereogram is shown in Figure 11 for your enjoyment. Be forewarned that it is a good bit more difficult to perceive than the simple square, but the results are well worth the extra effort. It depicts a spiral surface coming out of the page.

It is important not to overstate the conclusion reached from the perception of random dot stereograms. The fact that people can perceive stereoscopic depth in random dot stereograms does not prove that there is no shape analysis prior to stereopsis. It shows only that stereoscopic depth can be perceived without monocular shape information. There may well be some primitive shape or contour analysis before stereopsis that aids in solving the correspondence problem when monocular shape information is present. The orientation of local lines and edges is one example of further information that would be useful in solving the correspondence problem, for only lines or edges of similar orientation would be potential matches. Such orientation information is available relatively early in the visual system in the output of simple cells in area V1 (Hubel & Wiesel, 1962). In fact, the difficulty of achieving depth perception in random dot stereograms compared to stereo pairs of normal photographs suggests that monocular shape information is indeed useful. Even so, random dot stereograms show that in the absence of such information, binocular depth perception is possible.

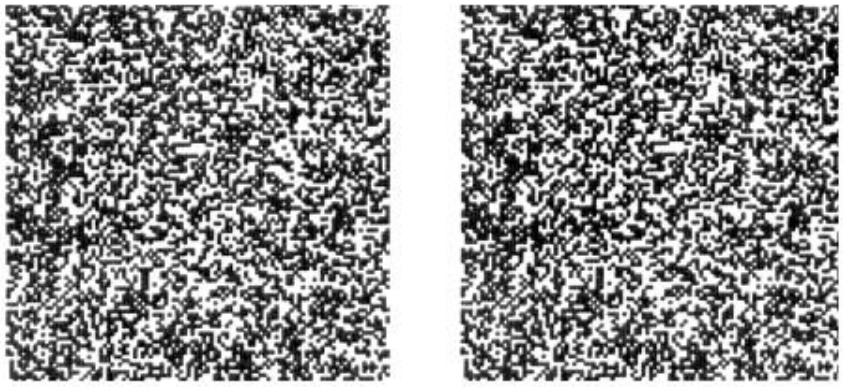

Figure 10 A random dot stereogram. These two images are derived from a single array of randomly placed squares by laterally. displacing a region of them as described in the text. When they are viewed with crossed disparity (by crossing the eyes) so that the right eye's view of the left image is combined with the left eye's view of the right image, a square will be perceived to float above the page.

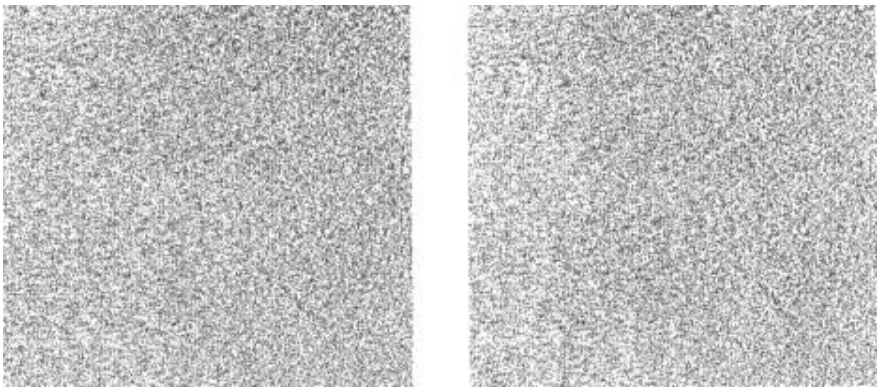

Figure 11 A random dot stereogram of a spiral surface. If these two images are fused with crossed convergence (see text for instructions), they can be perceived as a spiral ramp coming out of the page toward your face. This perception arises from the small lateral displacements of thousands of tiny dots.

Autostereograms. Another kind of stereogram has become popular in the past few years that does not require special glasses or viewing apparatus. They were initially called autostereograms, but they are now more widely known as magiceye stereograms. Figure 12 shows an example. When viewed normally, it looks like what it is: a flat picture with a repetitious horizontal structure. When viewed somewhat differently, however, it creates a compelling illusion of stereoscopic depth. To experience this illusion, hold the book in the frontal plane and cross your eyes somewhat so that the two black circles at the top appear as three dots in a line. When you have accomplished this, the elements in the autostereogram will start to fuse, and you will begin to get the impression of objects at different depths. Eventually, you will see a checkerboard of random dot squares floating above an underlying plane of random dot background texture.

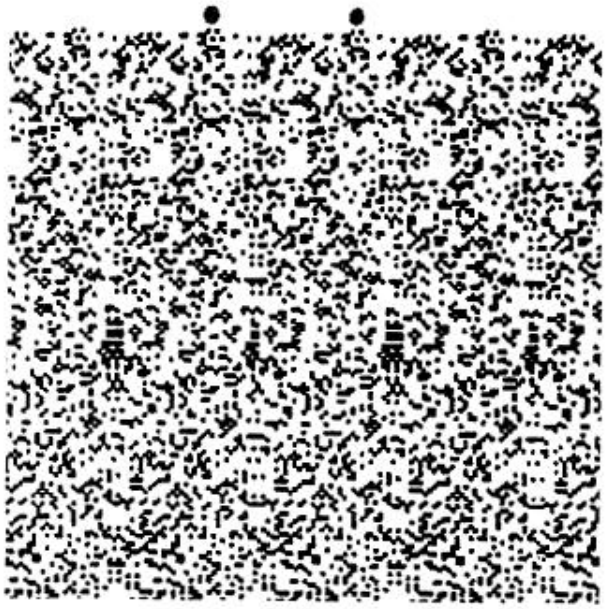

Figure 12 An autostereogram. When this image is viewed as directed in the text, a square can be seen floating in front of the rest of the figure, without the aid of special glasses.

To understand how such patterns manage to create the perception of depth, consider the much simpler autostereogram shown in Figure 13. View it using the crossed convergence method, and when you see two adjacent circles with white dots inside them, the different rows of shapes will appear to be located at different distances away from you: squares closest, diamonds farthest, and circles in the middle. If you view this auto-stereogram with uncrossed convergence, the depth of the shapes will reverse.

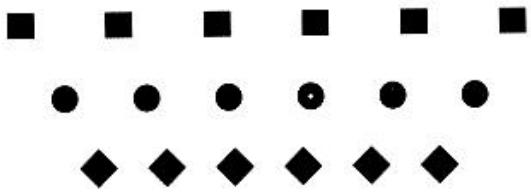

Figure 13 A simplified autostereogram. To experience a depth illusion with this figure, follow the instructions in the text to achieve crossed convergence. The rows of figures should begin to appear in different depth planes. The squares will appear closer than the circles, and the diamonds will appear farther away than the circles.

The illusion of depth is created when the two eyes fixate on two different objects and fuse them as though they were the same object. (You can verify that this must be the case because there is actually only one circle with a white dot in it, yet you see two such circles whenever you perceive illusory depth.) By crossing your eyes, you can induce the visual system to fixate different objects with your two eyes. The objects in the same row are identical in shape to enable this fusion error to occur. When this happens, the two fixated objects appear to be a single object in the depth plane on

which the eyes are converged. This plane lies in front of the actual depth plane, as illustrated in Figure 13a. Once this happens, the other objects in that row can also be fused. If their spacing is even, wrongly matched pairs will be perceived to be located in the same plane as the illusion of the fused fixation objects.

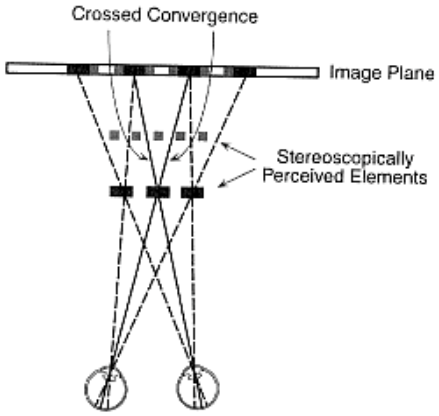

Figure 13a. Autostereograms explained. Illusions of depth occur in viewing autostereograms when the eyes are misconverged such that different elements in the repetitive pattern are fused.

Figure 13 is a very simple autostereogram compared to the complex, computer-generated example in Figure 12 and the much more complex ones available in recent commercial books (e.g., Magic Eye: A New Way of Looking at the World, by N. E. Thing Enterprises, 1993), but the basic principles are the same.

3.3 Computational Theories

We will now return to our earlier question: How does the visual system solve the correspondence problem in random dot stereograms when there is no global shape information? To appreciate the difficulty of this problem, consider precisely what must be done. Most (but not all) dots in the left image have some corresponding dot in the right image that came from the same environmental location. The visual system must somehow figure out which pairs go together. To simplify the problem somewhat, we will consider only the pairs of points that lie along the orientation of binocular displacement, which is horizontal if the head is upright. If there are 100 points along a given horizontal line in one image, then the number of logically possible pairings is 100! (one hundred factorial). This is a very large number indeed.

A number of different computational approaches to the correspondence problem have been explored. Some match individual points (pixels), others match lines and edges, and still others match local regions in one form or another. We will now consider an approach to solve the correspondence problem for the simple case of horizontal binocular displacement. A more detailed account on the general case of arbitrary binocular displacement, as well as the problem of depth recovery, will be given in the Chapter on Stereo later.

The First Marr-Poggio Algorithm to Solve Correspondence. One interesting and well-known algorithm was devised by David Marr and Tomaso Poggio of M.I.T. in

1977. It is neither the latest nor the best theory of human stereopsis, but it is an interesting—and historically important—example of how a dynamic neural network can be constructed to solve a computationally difficult visual task. It is also a good example of how heuristic assumptions or constraints can be implemented in such networks. We will therefore examine it in some detail.

The basic idea behind the Marr-Poggio (1977) algorithm is to solve the correspondence problem by matching individual pixels in the left and right images. The starting point for understanding how this is done is the concept of an inverse projection from the two retinal images back into the world, as illustrated in Figure 14. This diagram depicts a top view of two black-and-white striped surfaces in the environment, a small one situated in front of a larger one, as shown at the top. Light reflected from these surfaces registers on the left and right retinal images—which have been flattened here for simplicity—as shown in the middle of the diagram. To form an inverse projection, each pixel of these images is then projected back into the "mirror-image environment" shown below. The shaded cells in this matrix represent positions at which there are color matches between pixels in the two images. The points match (and are shaded) if both pixels are black or both are white; they do not match (and are unshaded) if one is black and one is white. If you trace the projections from a shaded cell back to their pixels in the left and right retinal images, you will find that both have the same color. Among these numerous matches are the correct ones that correspond to the visible portions of the actual surfaces in the real world. These correct matches are shaded more darkly in Figure 14 to differentiate them from the false matches that also arise.

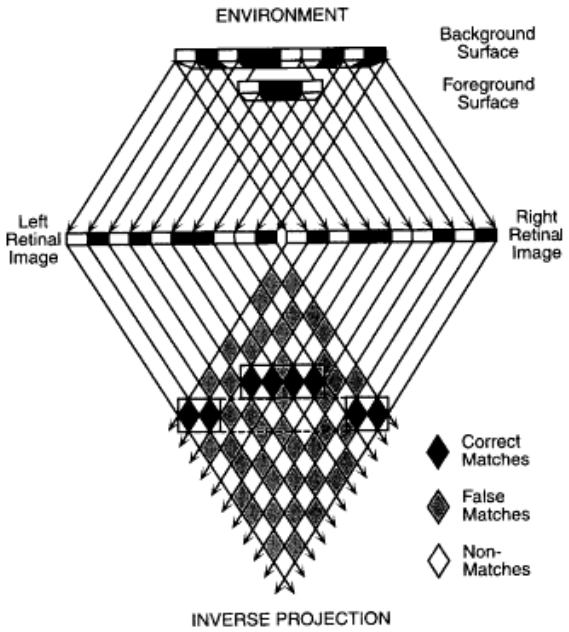

Figure 14 Inverse projection of two surfaces in depth. The upper half of the figure shows a top view of how two surfaces project to (flattened) left and right retinal images. The lower half shows how these images can be used to form an inverse projection in which there are many same-color matches, some of which correspond to the correct depth solution (dark elements) and many of which are false matches (gray elements).

The fact that there are false matches as well as true ones reflects the fact that this is an underconstrained inverse problem, one that has many possible solutions. The problem for the visual system is how to determine which matches are correct and which are false.

Marr and Poggio (1977) proposed that this could be accomplished by the dynamic neural network diagrammed in Figure 15. It shows the left and right images from Figure 14 activating internal nodes in a neural network that represent the set of all possible correspondences. That is, each node represents a potential match between two pixels: the one from the left image that projects to it and the one from the right image that projects to it. Only the intersections that come from pixels of the same color are possible matches because pixels that are projected from the same point in the environment must have the same color (both white or both black).

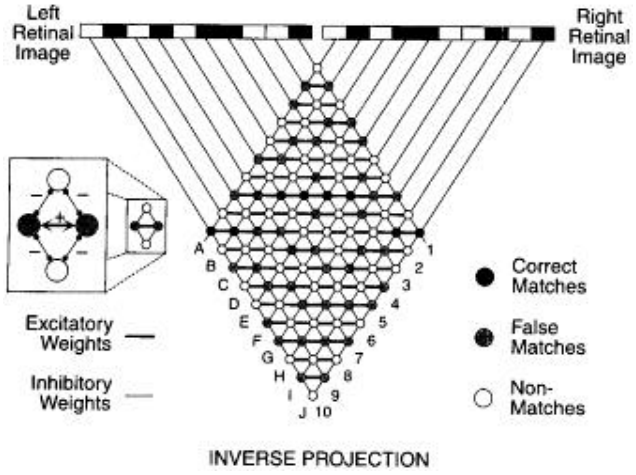

Figure 15 The first Marr-Poggio stereo algorithm. A neural network performs an inverse projection of the two retinal images, activating nodes corresponding to positions where the left and right images are the same color. Inhibitory connections (see inset) implement the surface opacity constraint, and excitatory connections implement the surface continuity constraint. (See text for further details.)

Matching by color does not produce a unique solution to the correspondence problem, however, because there are still many color matches for each point in the left and right images. This is why random dot stereograms pose such a difficult problem for the visual system. The question is: What further heuristic constraints might be brought to bear on this formulation of the problem that would produce a unique, correct solution most of the time? Marr and Poggio (1977) employed two such further constraints:

1. Surface opacity: The opacity constraint states that because most surfaces in the world are opaque, only the nearest one can be seen. Thus, if correspondence A10 is correct in Figure 15 (that is, if pixel A in the right image actually corresponds to pixel 10 in the left image), then correspondences B10, C10, D10, and so forth cannot be correct, and neither can A9, A8, A7, and so forth.

2. Surface continuity: The continuity constraint states that because surfaces in the world tend to be locally continuous in depth (except at occluding edges), the correct solution will tend to be one in which matches are "close together" in depth, as they would be on locally continuous surfaces.

Note that both constraints are heuristic assumptions that are usually true but not always. If they are true, the solution the algorithm finds will generally be correct. If they are not—that is, if one or more of the surfaces are transparent and/or if the surfaces are not locally continuous—the solutions that it finds will tend to be incorrect.

Marr and Poggio implemented these two constraints in the connections between nodes of the neural network in Figure 15. The model works by first activating all of the nodes in the "intersection network" that represent like-colored pixels in the left and right images. These are just the shaded nodes in Figure 15, indicating that they have been activated in the initial phase of the algorithm. This set of possible correspondences is then subjected to the opacity and continuity constraints by the nature of the connections between nodes in the network as illustrated in the enlarged four-node inset at the left of the figure. Opacity is implemented by having mutual inhibition among all nodes along the same line of sight in the network. This part of the architecture, denoted by the diagonal connections in Figure 15, is called a **winner-take-all network** because it allows only one node in each diagonal line to remain active after activation has settled into a stable state.

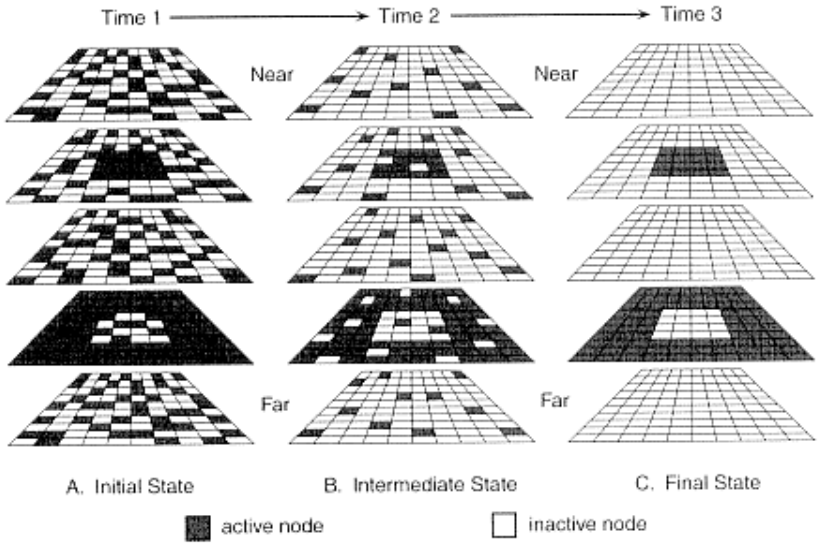

Figure 16 Dynamic behavior of the Marr-Poggio stereo network. Active nodes corresponding to different depth planes are represented by black squares. (A) Initially, all positions where there are same-color matches are activated, including both true and false correspondences. (B) As excitatory and inhibitory interactions occur, false matches diminish, but so do some true ones. (C) Further interactions leave only the true matches.

The continuity constraint is implemented in the network by having mutual excitation between pixels in the same or nearby depth planes. These interactions between nodes are indicated in Figure 15 by the thicker horizontal connections. They have the effect that possible correspondences in the same environmental depth plane tend to "help each other out" by activating each other through mutual facilitation. This will cause solutions to tend toward coplanarity as the continuity constraint implies.

The joint effect of these two constraints is to progressively reduce the complete set of possible matches to the single set most compatible with (1) the disparity information available in the sensory images, (2) the opacity constraint, and (3) the continuity constraint. The network churns away, sending activation and inhibition back and forth through the excitatory and inhibitory connections, until it eventually settles into a stable state. This final state of the network usually—but not always—corresponds to the actual state of affairs in the environment. Figure 16 shows how activation in the intersection network changes over time for the simple case of a square figure in front of a background plane, as in the stereogram in Figure 10. The initial pattern of activation is completely determined by the set of pixel pairs whose colors match. As the nodes begin to interact via the excitatory and inhibitory connections, the opacity and continuity constraints begin to take effect. Eventually, the initial correspondences are reduced to just the set representing a square in front of a background.

Marr and Poggio's first algorithm is an interesting example of how a process of unconscious inference can be implemented in a neural network without invoking deductions based on either numerical calculations or the sequential application of symbolic logical rules. Instead, the assumptions are built into the connection strengths within the network, and they work simply by having a favorable effect on its behavior with respect to achieving veridical perception.

Edge-Based Algorithms

A few years after publishing their first stereo algorithm, Marr and Poggio (1979) suggested a second one that differs from their first in a number of important respects:

1. Edge-based matching. The second Marr-Poggio algorithm finds stereoscopic correspondences by matching edges in the right and left images rather than individual pixels. This is more efficient because it allows the process to take information into account that simply is not available in matching individual pixels, such as the orientation of the edges and their polarity (light to dark versus dark to light). Edges that do not match in orientation and polarity can be eliminated from consideration, thus adding further constraints on the solution.

2. Multiple scales. The second algorithm exploits the multiple size (or scale, or spatial frequency) channels in the visual system by first looking for corresponding edges at a large spatial scale (that is, at low spatial frequencies) and only later at smaller scales (that is, at high spatial frequencies). In terms of solving random dot stereograms, this means that the early, large-scale process works not on individual dots, as their first algorithm did, but on larger regions of the image. Only after possible edge matches are found at this global level is a more detailed matching attempted at the finer-grained level of individual dot contours.

3. Single-pass operation. The edge-based algorithm is a noniterative process, meaning that it does not require many cycles of interaction to converge, as did the first Marr-Poggio algorithm. Instead, it simply finds the best edge-based correspondence in a single pass through a multistage operation. Because iteration is a time-consuming process, computer implementations of the second algorithm are much faster than those of the first.

Many of the benefits of the second Marr-Poggio algorithm derive from the fact that the matching operation is performed on the output of edge detectors rather than individual pixels. This is more plausible biologically because binocular processing begins in area V1 of the cortex, after the outputs of individual receptors have been recombined into the more complex elongated receptive fields of the cortex. If these cells are indeed doing edge detection, as Marr (1982) claimed, then solving the correspondence problem by matching edges is a sensible approach. It also has distinct computational advantages because edge detectors carry information about more complex features of the image (that is, oriented edges of a given polarity) than do individual receptors, and this complexity enables many potential matches to be rejected. The second Marr-Poggio algorithm also agrees more closely with the results of psychophysical experiments using human subjects (Marr, 1982). It is thus a better model of human stereo vision than their first algorithm.

4 Dynamic Information (Motion)

Dynamic Visual information refers to changes in visual structure that occur over time due to object or self motion. When an observer moves with respect to the environment, the direction and rate at which different objects are retinally displaced depends not only on the observer's motion, but on how far away the objects are and on where the observer is fixated.

4.1 Motion Parallax

One way in which depth can be recovered from motion information is due to what is known as motion parallax: the differential motion of pairs of points due to their different depths relative to the fixation point. You can demonstrate the existence and nature of motion parallax by carrying out some finger experiments very much like the ones we used for binocular disparity. Here's how:

1. Hold your two index fingers in front of your face, one at arm's length and the other halfway to your nose.

2. Close your right eye and align your two fingers with some distant object, focusing on the distant object.

3. Keeping your fingers as still as possible, slowly move your head to the right. Notice that both of your fingers move leftward relative to the distant object, but the closer finger moves farther and faster.

4. Now move your head to the left and notice that your fingers move rightward, the closer one again moving farther and faster.

The differential motion of your fingers in this demonstration illustrates the nature of motion parallax: The images of points at different distances from the observer move at different retinal velocities as the observer's stationpoint changes.

The close informational kinship between motion parallax and binocular disparity can be appreciated by comparing the corresponding finger experiments that you performed. In the case of binocular disparity, you kept your head still and compared the left retinal image with the right one, both of which are normally available at the same time. In the case of motion parallax, you moved your head over time and compared an earlier image with a later one. Thus, binocular disparity involves the difference between a pair of displaced simultaneous retinal images, and motion parallax involves the difference between a pair of displaced sequential retinal images.

Motion parallax is also like binocular disparity in that it provides only relative information about depth. That is, it does not specify the actual distance to an object, but only how much closer or farther it is than the fixated one. Unlike binocular disparity, however, motion parallax can provide effective depth information at great distances. If you were to drive past two hills, for example, one of which was 2 miles away and the other 4 miles away, you would be able to determine which one was closer from relative motion parallax. Binocular disparity would provide no information in such a case because the distances between the two eyes is too small in comparison with the distance of the mountains. With motion parallax, however, the separation that is achieved from successive views can be much greater—particularly in a speeding car—thus affording much greater parallax.

4.2 Optic Flow Caused by a Moving Observer

In naturally occurring perception, relative motion parallax of two isolated points is seldom, if ever, encountered. Observers are usually moving about and actively exploring cluttered environments, engaging in activities that cause complex patterns of optic flow.

Figure 17 shows some examples of optic flow patterns. The diagrams indicate how a sample of image points changes over time as the observer moves, as though tracing their paths in a time-lapse photograph. Part A indicates the pattern of optic flow when the observer moves smoothly in the direction indicated by the large arrow below the figure and fixates the point indicated in the center of the scene. The fixated point does not move on the retina, of course, because the observer tracks it with head and eye movements, so it stays on the fovea. The rest of the visual field moves at a speed and in a direction that depend on the depth relation between it and the fixated object.

Another common pattern of optic flow is optic expansion or looming, as illustrated in Figure 17B. It occurs when an observer moves directly toward a surface in the frontal plane, fixating on the point toward which he or she is heading. Ocular expansion results, for example, when you walk toward a wall while looking straight ahead.

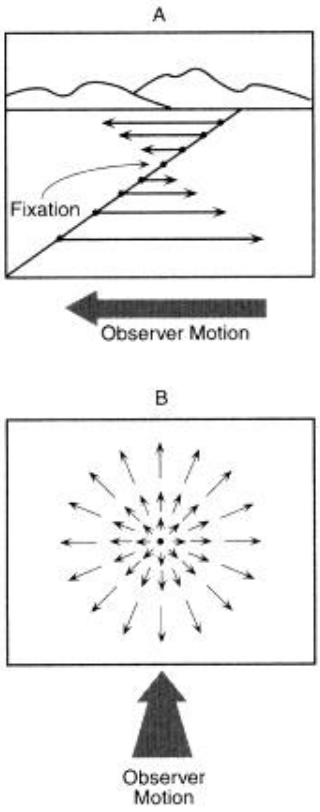

Figure 17 Motion gradients produced by a moving observer. (A) The optic flow created by an observer moving leftward (large arrow) while fixating the point in the middle of the line. (B) The optical expansion pattern that results from an observer moving toward a fixation point straight ahead, as in walking toward a wall.

These motion gradients are just a few special cases of very simple optic flow patterns that arise from very simple motions with respect to a single environmental surface. Optic flow patterns become exceedingly complex as the situation begins to approximate naturally occurring conditions. More complicated flow patterns result from changes in the direction of gaze (due to head and eye movements), up-and-down movements of walking, curved paths of motion, and complex environments consisting of many surfaces at different orientations, distances, and degrees of curvature. Realistic optic flow patterns produced in the course of normal activities in normal environments are so complex that they could not possibly be catalogued in terms of a few simple types. Their structure can be discerned only through sophisticated mathematical analysis. Such computational analysis shows that the relative depth of each pair of points on a curved, rigid, textured surface can be recovered from a moving monocular point of view and that the path of observer motion can be recovered from the same sensory data (Longuet-Higgens & Prazdny, 1980; Prazdny, 1980). We will discuss these theories in a later chapter.

5 Pictorial Information

Although stereopsis and motion produce particularly compelling experiences of depth, they are by no means the only sources of depth information. The most obvious

argument in favor of this conclusion is that if you close one eye and keep your head still, the world continues to look quite three-dimensional. Another is that photographs and certain kinds of realistic drawings and paintings can produce compelling impressions of depth. This is obvious from even a brief glance at Figure 18, which shows a simple photograph of a 3-D scene. The remaining sources of depth information are known collectively as pictorial information because they are all potentially available in static, monocularly viewed pictures. Pictorial information can be very powerful, for it often leads to good depth perception in 2-D pictures when both stereo and motion information indicate that they are quite flat. Indeed, it can even overcome stereo depth information.

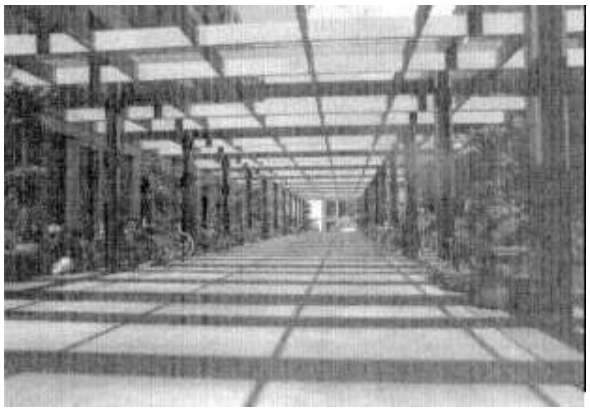

Figure 18 A demonstration of pictorial sources of depth information. This photograph contains a great deal of optical information about depth that arises from stationary, monocular structure in the image.

5.1 Familiar Size

Figure 19 Relative size as a cue to depth. The cheerleaders in this line are all about the same size, but their projected images become smaller as their distances from the camera increase. If one assumes that they are actually about the same size, their relative distances can be recovered from their image sizes.

The depth cue of relative size is independent of the identity of the objects involved. But many objects—perhaps even most—tend to have a characteristic size or range of sizes with which experienced perceivers are familiar. Adult men vary somewhat in height, but the vast majority are between 5 feet 6 inches and 6 feet 2 inches. Similarly, tables are about 2 feet 6 inches off the floor, cars are about 5 feet high; ceilings are about 8 feet above the floor, and so on. The importance of these facts is that if the size of an object is known to the perceiver, then the size-distance equation can be solved for its actual distance from the observer. The knowledge involved here is not conscious knowledge, nor is solving the equation deliberate symbolic manipulation. Rather, they are rapid, unconscious processes that occur automatically, without our even thinking about them.

5.2 Texture Gradients

Another important manifestation of the structure of perspective projection in depth perception is what is known as texture gradients: systematic changes in the size and shape of small texture elements that occur on many environmental surfaces. Examples of naturally occurring texture include the blades of grass in a lawn, the pebbles in a driveway, the strands of yarn in a carpet, the warp and woof of a woven fabric, the tiles in a bathroom floor, and so forth. An example is shown in Figure 20. In addition to providing information about depth, texture gradients also can inform observers about the orientation of a surface in depth and about its curvature. Quite complex surface shapes can be realistically depicted by textural variations, as the example in Figure 21 shows.

Stevens (1979) demonstrated that two aspects of textural variation—element size and shape—provide independent sources of information about surface orientation. The overall size of texture elements diminishes with distance because all dimensions decrease as the distance to the stationpoint increases. Element size can therefore be used to estimate the relative distance to different parts of the surface and thus to recover the orientation of the textured surface. But notice that this will be true only if the texture elements are actually about the same size.

This is another example of heuristic assumptions in depth perception, since the perceptual conclusion about the distance to texture elements based on their image-size will be accurate only if the objects forming the texture are similar in size. If they are not, then illusions of depth and surface orientation will result.

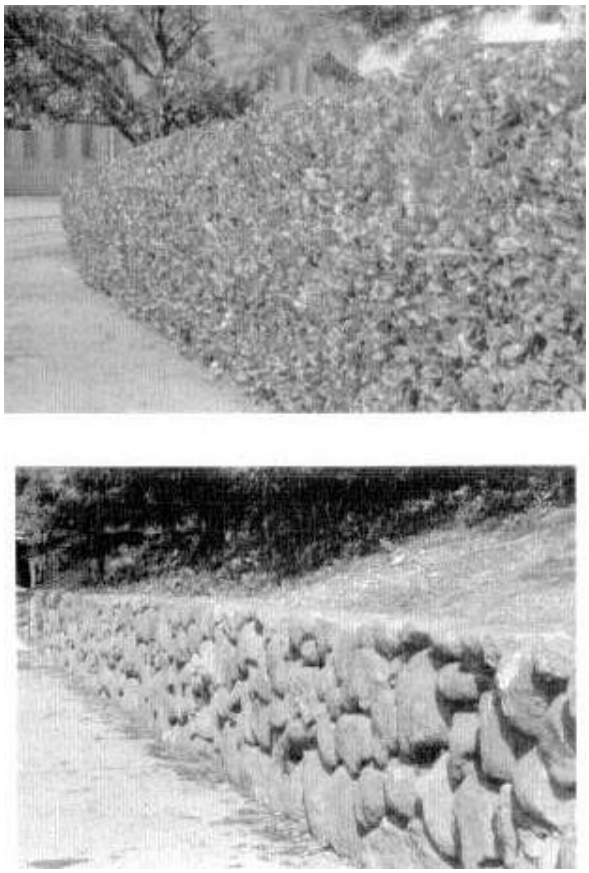

Figure 20 Natural texture gradients. Many natural surfaces are defined by texture elements of about the same size and shape so that surface depth and orientation can be recovered from systematic changes in their projected sizes and shapes.

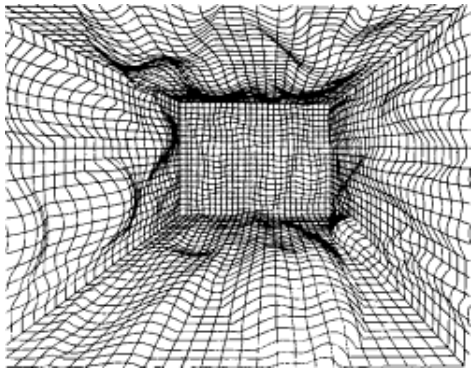

Figure 21 Artificial texture gradients. Artificial surfaces of arbitrarily complex shapes can be rendered by using identical texture elements in computer graphic displays.

The projected shape of texture elements can also carry information about the orientation of the surface, as illustrated in Figure 21. Again, however, this information can be recovered from the image only if additional assumptions are made about the actual shapes of the texture elements in the environment. Stevens (1979) used the aspect ratio (the ratio of the longest to the shortest dimension) of texture elements to estimate the orientation of the elements themselves and the surface on which they lie. His analysis rested on the assumption that the dimensions of real-world texture elements are approximately uniform over different orientations. Witkin (1981) proposed another algorithm based on the assumption that the edges of texture elements tend to be distributed isotropically, meaning that the amount of contour at different orientations will be approximately the same, or at least equally distributed over orientations. This is a useful heuristic because when isotropic texture elements are viewed at a slant, their edges will not be isotropic in the image plane. Rather, they will be biased toward orientations that are perpendicular to the direction of tilt owing to foreshortening along the axis of tilt, as illustrated in Figure 22.

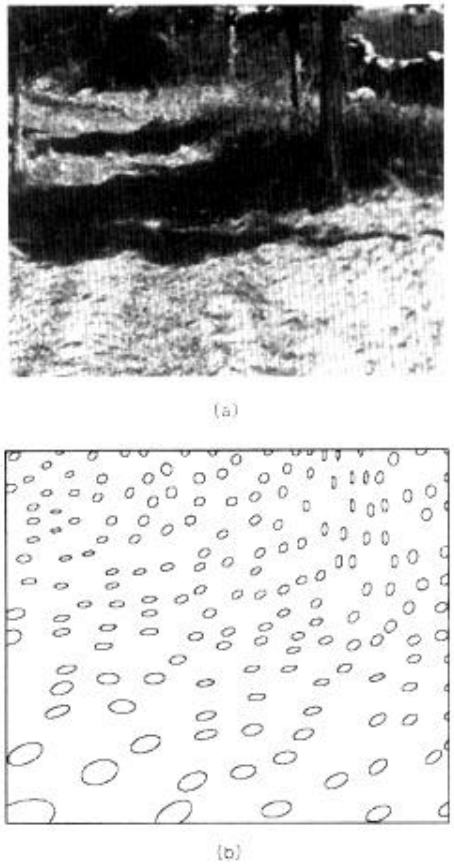

Figure 22 Estimating local surface orientation from texture. Panel A shows a natural scene that includes significant texture information, and panel B shows the output of Witkin's program in which the shape and size of the ellipses convey the estimated depth and orientation of the local regions of surface. (From Witkin, 1981.)

5.3 Edge Interpretation

One very important class of pictorial information about depth comes from the interpretation of edges or contours. In Figure 23, for example, people invariably perceive a square behind and partially occluded by a circle. All that is actually present, of course, is a 2-D configuration of regions bounded by edges, yet we perceive these edges as indicating a depth relation: The circle is in front of the square.

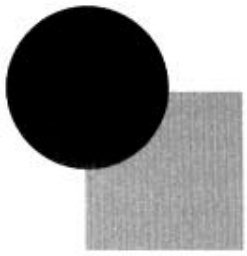

Figure 23 Partial occlusion as depth information. When one object partly occludes another, the occluding object is perceived as closer and the occluded object as farther.

A sophisticated theory of edge interpretation has evolved within computer vision in the **blocks world** environment. The goal of this theory is to determine the best interpretation for all edges in a line drawing of a blocks-world scene. Edges can arise in an image in several different ways, and the output of an edge interpretation program is a classification of the edges in terms of their environmental sources. Some arise from one surface occluding another, others arise from two surfaces meeting along a common border, still others arise from a shadow being cast over a single surface, and so on. Such distinctions are important for pictorial depth information because some edge interpretations imply specific depth relations. For example, if edge A partly occludes edge B, then A is necessarily closer to the observer than B is.

We will now examine this computational theory of edge interpretation to find out how it reveals information about the relative depth of surfaces from their bounding edges. The beginnings of a formal analysis

Four Types of Edges

We begin by assuming that earlier processes have produced an image in which the edges have been accurately identified. Thus, the input to this edge interpretation algorithm is an essentially perfect line drawing of all the edges in the scene. The task is to interpret them in terms of the environmental situation that produced them. To begin, we distinguish among four types of edge interpretations:

1. Orientation edges. Orientation edges refer to places in the environment in which there are discontinuities in surface orientation. These occur when two surfaces at different orientations meet along an edge in the 3-D world. They usually arise at internal edges within a single object (e.g., a cube) or where one object abuts another, such as a block sitting on a table. Examples of orientation edges in Figure 24 are labeled with O's.

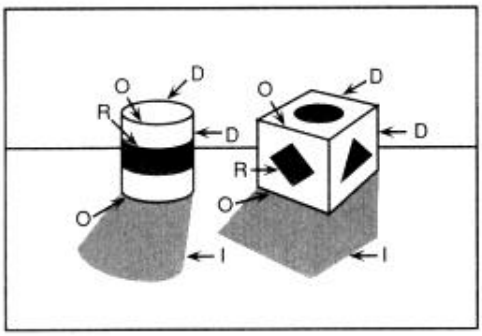

Figure 24 Four kinds of edges. This scene contains four different kinds of luminance edges: orientation edges (O) due to abrupt changes in surface orientation, depth edges (D) due to gaps between surfaces at different distances, reflectance edges (R) due to different surface pigments or materials, and illumination edges (I) due to shadows.

2. Depth edges. Depth edges refer to places where there is a spatial discontinuity in depth between surfaces, that is, places in the image where one surface occludes another that extends behind it, with space between the two surfaces. If they actually touch along the edge, then it is classified as an orientation edge. Examples of depth edges in Figure 24 are labeled with D's.

3. Illumination edges. Illumination edges are formed where there is a difference in the amount of light: falling on a homogeneous surface, such as the edge of a shadow, highlight, or spotlight. Examples of illumination edges in Figure 24 are labeled with I's.

4. Reflectance edges. Reflectance edges result where there is a change in the light-reflecting properties of the surface material. The most obvious examples are when designs are painted on an otherwise homogeneous surface. Examples of reflectance edges in Figure 24 are labeled with R's.

Now we will focus on orientation and depth edges because these edges provide the strongest constraints on depth interpretations of the scene. We therefore begin with a line drawing that contains only orientation and depth edges.

Edge Labels

Orientation and depth edges in objects with flat surfaces are mutually exclusive. If an edge in the image is caused by two differently oriented surfaces meeting, it is an orientation edge; if it is caused by one surface occluding another with space between, it is a depth edge. Each edge in the line drawing is therefore either an orientation edge or a depth edge and can be unambiguously labeled as one or the other. The goal of a theory of edge interpretation is to discover a process that labels every edge in the way that corresponds to people's perception of the same scene.

We need to further differentiate the labeling system so that there is a unique label for each qualitatively different type of orientation and depth edge. Two kinds of

orientation edges and two kinds of depth edges are required. The two types of orientation edges are called convex and concave, and they carry important information about the depth of the edge relative to the surfaces.

1. Convex orientation edges occur when two surfaces meet along an edge and enclose a filled volume corresponding to a dihedral (two-faced) angle of less than 180°. Convex edges indicate that the edge's angle points toward the observer, as do the external edges of a cube seen from the outside. Convex edges are illustrated in Figure 25 by the lines labeled with a "+."

2. Concave orientation edges occur when two surfaces meet along an edge and enclose a filled volume corresponding to a dihedral angle of more than 180°. Concave edges indicate that the edge's angle points away from the observer, as do the internal edges of a hollow cube seen from within. Concave edges are illustrated in Figure 25 by the lines labeled with a "-."

In the simple trihedral planar objects analyzed by Huffman and Clowes, each orientation edge is either convex or concave, never both.

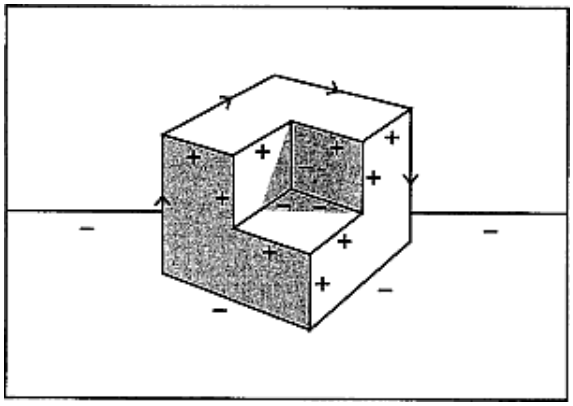

Figure 25 Convex versus concave orientation edges. Convex edges, labeled by a "+," arise when two surfaces meet at an interior angle of less than 180°. Concave edges, labeled by a "-," arise when two surfaces meet at an interior angle of more than 180°. Arrows indicate that the closer surface along a depth edge is on the right.

There are also two cases of depth edges that need to be distinguished: one in which the occluding surface is on one side of the edge and another in which it is on the other side. Depth edges are labeled by single arrowheads running along the edge, and the convention for its direction is a right-hand rule: The arrow is placed facing in the direction along which the closer, occluding surface is on the right side of the edge in the image (and the farther, occluded surface is on the left). In other words, if you imagine yourself moving forward along the edge in the direction of the arrow, the closer surface is always on your right. These two possible labels for each depth edge—an arrow in one or the other direction along the edge—are mutually exclusive, since the occluding edge can be on only one side. The correct labeling carries important depth information because it designates which surface is closer to the observer.

Thus far we have four possible labels for each edge in a line drawing containing only orientation and depth edges, all of which contain significant depth information. This means that if there are n edges in the drawing, there are 4^n logically possible labelings for it, corresponding to 4^n qualitatively different depth interpretations. This is an astronomically large number even for very simple scenes. For example, although there are only 20 edges in Figure 25, this simple line drawing allows for 1,048,576 logically possible edge labelings! In this case, however—and in most other cases—people generally perceive just one. How can the set of logically possible interpretations be reduced to a manageable number, much less just one?

Physical Constraints: Huffman and Clowes based their analyses on the crucial insight that not all logically possible labelings are physically possible. They examined local constraints at vertices of trihedral objects—objects whose corners are formed by the meeting of exactly three faces—and found that only a small fraction of logically possible labelings could be physically realized. Consider, for example, the set of possible "arrow" junctions, in which three edges meet at an angle in the image plane of less than 180°. Because each edge could be labeled in any of the four ways described above—as a concave or convex orientation edge or as a right-handed or left-handed occluding edge—there are 4^3 (= 64) logically possible labelings for an arrow junction. However, Huffman and Clowes proved that only three of them are physically possible. The same reduction of 64 to 3 possibilities is achieved for "Y" junctions. Huffman's catalog of all physically possible realizations of trihedral angles is given in Figure 26.

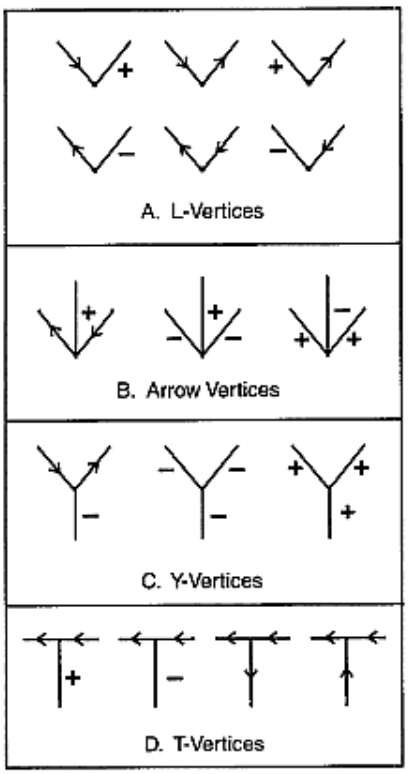

Figure 26 The catalog of vertex types for trihedral angles. All physically possible interpretations of vertices are shown for "L," "arrow," "Y," and "T" vertices.

Huffman and Clowes also pointed out that there are further constraints on edge interpretation that operate at a more global level. They result from the fact that within the constraints of the objects in blocks world—polyhedra with planar surfaces—each edge has a constant interpretation along its entire length. Convex edges cannot become concave and right-handed occluding edges cannot become left-handed unless there are vertices between the edges that allow the change in interpretation. By making sure that the interpretations assigned to the same edge at adjacent vertices are consistent in this way, the number of logically possible labelings can be further reduced. Figure 27 gives a simple example of how this global consistency constraint works for a tetrahedron:

1. Consider all physically possible labelings of the arrow vertex A. From the catalog shown in Figure 26, we know that there are just the three shown.

2. Next, consider how these possibilities interact with all physically possible labelings of the arrow vertex B. Of the nine logical possibilities, four are eliminated by physical constraints because they would require two different labels for the central edge.

3. Finally, consider whether the resulting five labelings of vertices A and B provide physically possible labelings for the "L" junctions at vertices C and D. This constraint eliminates two more, leaving only three physically possible interpretations.

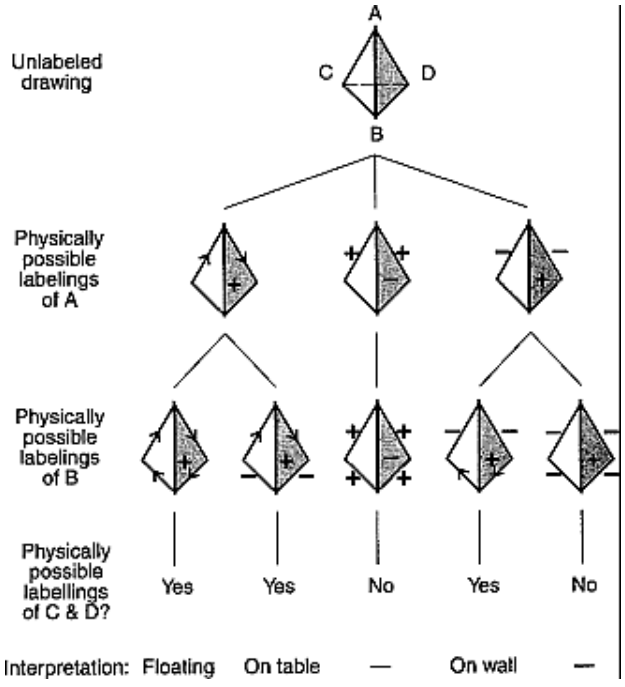

Figure 27 Interpreting the edges of a tetrahedron. A simple tetrahedron (top) can be interpreted by first labeling vertex A in all physically possible ways, then labeling vertex B in all physically possible ways, and finally eliminating impossible labelings of vertices C and D. Only three physically possible interpretations remain.

What are these three interpretations? The perceptually preferred one has concave orientation edges at the bottom, as would be found if the tetrahedron were sitting with its lower surface (BCD) on a table. The alternative with concave orientation edges at the top corresponds to the percept of the tetrahedron attached to a wall by its back surface (ACD). And the one with occluding depth edges all along the perimeter corresponds to a tetrahedron floating unsupported in the air.

Notice how drastically the set of possible labelings was reduced by imposing physical constraints. We began with a five-line drawing that could have 4^5 (= 1024) logically possible labelings. After analyzing purely physical constraints, only three remained! Notice that this edge analysis does not provide any way to decide among these three alternatives, and so further constraints must be introduced to settle on a single solution.

Extensions and Generalizations

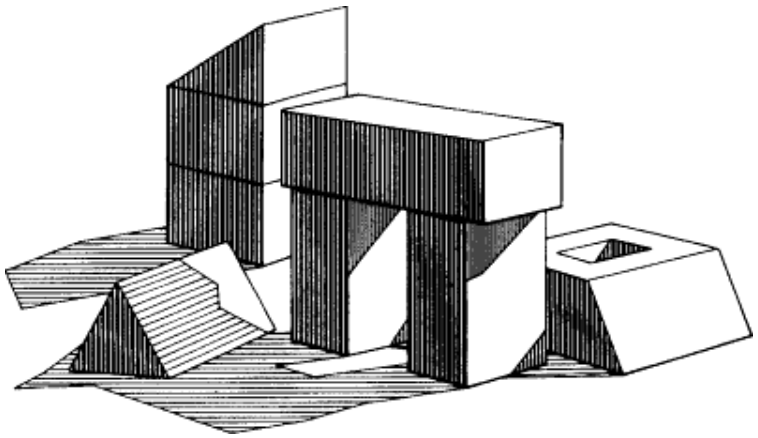

Figure 28 A blocks world scene with shadows. Waltz's algorithm for edge interpretation produces just one interpretation for this scene, which is the one people always perceive. Adding shadows makes the interpretation process more accurate because it provides further constraints.

The Huffman-Clowes analysis of physical constraints was a major conceptual breakthrough in the theory of edge interpretation. David Waltz (1975), then a graduate student at M.I.T., extended the Huffman-Clowes analysis to include 11 types of edges, including shadows and "cracks" (lines that result from the exact alignment of coplanar edges: essentially, orientation edges at 180° angles). This expansion of edge types proliferated the catalog of physically possible vertices to thousands of entries, but as you might guess by extrapolating the conclusions of the Huffman-Clowes analysis, it turned out to reduce even further the number of possible interpretations that could be assigned to a given shaded line drawing. For example, the complex drawing shown in Figure 28 yields only one physically possible labeling, which is the one people always perceive. As it turns out, the strongest additional constraints in Waltz's analysis come from corners that cast a shadow.

The success of Waltz's algorithm for assigning edge interpretations does not even approach human levels of competence, however, because its application is limited to planar polyhedra. It does not work for curved surfaces or for objects that contain thin

sheets (such as folded paper) rather than volumes. Neither of these pose serious problems for human observers, who can interpret line drawings of scenes involving quite complex objects of either type. On the other hand, within its own domain Waltz's program is uncanny in its ability to arrive at physically possible interpretations, some of which people seldom, if ever, perceive without explicit guidance.

Figure 29 A line drawing of a potted plant. Although human perceivers have no difficulty perceiving the shape of the leaves of this plant from the drawing, blocks-world edge algorithms cannot interpret it correctly.

5.4 Shading Information

Yet another useful and effective source of information about the shape of surfaces curved in depth comes from shading: variations in the amount of light reflected from the surface as a result of variations in the orientation of the surface relative to a light source.

Perceiving Surface Orientation from Shading

The ability of human observers to recover surface orientation and depth from shaded objects and pictures has been studied experimentally by Koenderink, Van Doorn, and Kappers (1992, 1996). They showed observers pictures of the human torso shown in Figure 30 and had them indicate the perceived orientation of a fairly dense sampling of positions on the surface. They made their measurements by giving the subjects control over a small computer-generated display, called a gauge figure, consisting of an oval surrounding a short line segment, as indicated for several positions in Figure 30. These gauge figures were used to probe surface orientation because they can easily and accurately be perceived as circles oriented in depth, at a particular slant and tilt, with short lines sticking perpendicularly out of their centers. (Gauge figures were used in Figure 1 to demonstrate the difference between slant and tilt, for example.) Subjects were instructed to adjust each gauge figure so that it appeared to be a circle lying flat on the surface of the object with the line sticking perpendicularly out of the surface, as illustrated by ovals A and B in Figure 30. For contrast, ovals C and D show examples in which they do not appear to be circles lying flat on the surface.

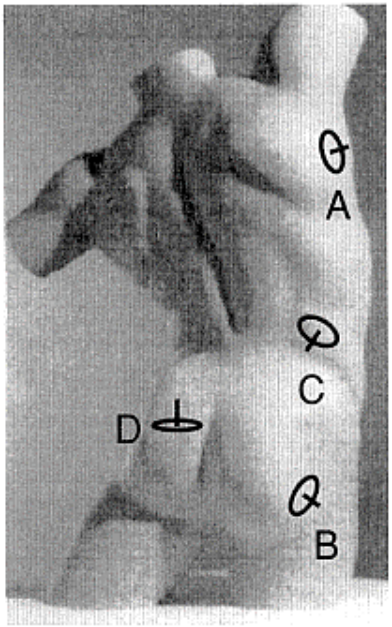

Figure 30 Studying the perception of surface orientation from shading. Subjects saw this picture of a male torso and adjusted the shape and orientation of oval test figures so that they looked like circles lying flat on the surface (A and B) rather than askew (C and D).

Figure 31 shows the results from one subject averaged over several sessions. Notice that this representation of the data approximates the appearance of the original torso in the texture gradient of many circles on its surface. Such findings allowed the experimenters to reconstruct people's perception of surface depth and orientation using minimal assumptions. They found that different subjects were remarkably similar in their qualitative perception of these surfaces but that they differed quantitatively in the amount of depth they perceived. Another important conclusion they reached was that observers were not using strictly local information in making their responses, but were integrating over a substantial region of the object's surface. These conclusions were not dependent on the surface depicting a familiar object such as a torso because they obtained similar re-suits using unfamiliar abstract sculptures. Exactly what perceptual process might have produced these global effects is not yet clear.

Like many other aspects of depth and surface perception, the visual analysis of shading often rests on heuristic assumptions. Perhaps the most striking is that our brains implicitly assume that illumination comes from above. Figure 32 shows an example of a surface with two rows of indentations. The top ones typically appear to be convex bumps, bulging outward toward the viewer, and the bottom ones appear to be concave dents, curving inward away from the viewer. In fact, this perception is veridical only if the illumination comes from above in the depicted scene. You can demonstrate this simply by turning the book upside down. This reverses the assumed direction of illumination (relative to the rows) and thus reverses the perceived

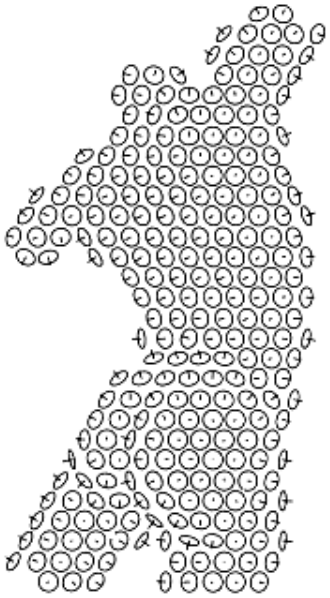

Figure 31 Local surface orientations reported by one subject. The average ovals produced by one subject for every point tested on the surface of the male torso shown in Figure 30.

convexity/concavity of the dimples. The ones at the top of the page (now the lower ones) appear concave, and the others (now the upper ones) appear convex. The assumption of illumination from above makes a great deal of sense because our visual environment almost always is illuminated from above. It is therefore another example of using plausible hidden assumptions to determine how we solve an underconstrained inverse problem.

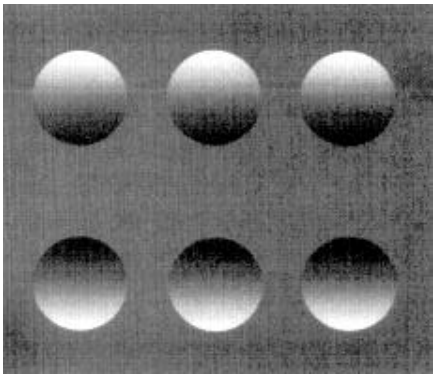

Figure 32 Direction of illumination and perceived convexity. The top row looks like convex bumps and the lower row like concave dents because the visual system assumes that illumination comes from above. If you turn the book upside down, the perceived convexity of these elements reverses.

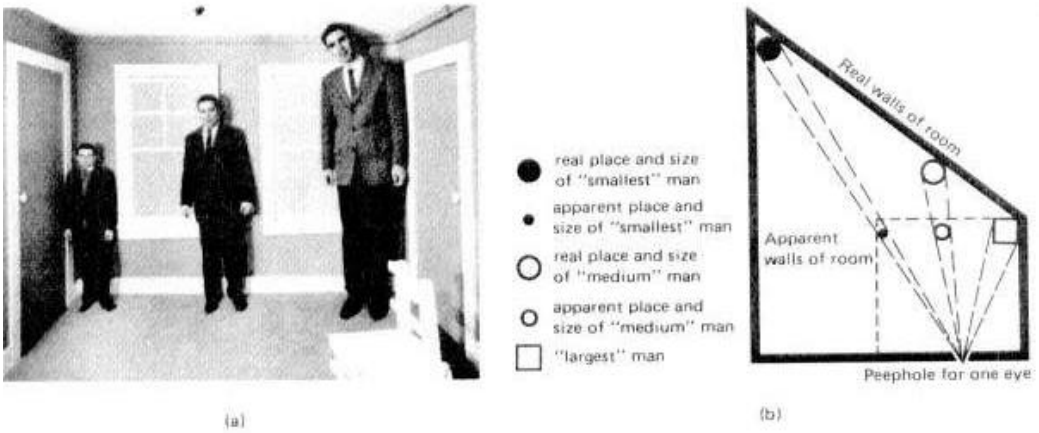

Figure 33 Conflicting information in an Ames room. Perspective information and familiar size information conflict when people are seen in an Ames room, a distorted room that appears to be rectangular from a single viewpoint. Perspective information dominates, so the people are perceived as greatly different in size.

We have now examined a large number of widely different information sources for perceiving depth in a scene. Since all these sources bear on the same perceptual interpretation of surfaces oriented in depth, they must somehow be put together into a coherent consistent representation. How does the visual system accomplish this integration?

Under normal viewing conditions, integrating different sources of depth information is largely unproblematic because they are very highly correlated. They therefore converge naturally on a single, coherent, and accurate representation of the distance and orientation of surfaces relative to the viewer. In the laboratory, however, different factors can be manipulated independently so that cues come into conflict. What happens in such cases, and what do the results imply about the rules for integrating different information?

Perhaps the simplest possibility is that one information source will dominate some other conflicting source with the result that the latter is completely ignored. This form of integration implies a hierarchy of depth sources such that those higher in the ordering dominate those lower down, although it is not clear what might happen in this hierarchy if several sources "gang up" on one outlier.

A well-known example of what appears to be dominance between depth cues is the Ames room, which pits perspective information against familiar size of objects. The Ames room is a greatly distorted room that looks normal from one particular viewpoint. Figure 33 shows an Ames room together with its floor plan. Even though it is not rectangular, it looks rectangular from the designated viewpoint. When objects known to be approximately equal in size are placed along the back wall of the room, such as the three people in Figure 33, observers at the special viewpoint invariably report two illusions: (1) the people are seen as equally distant and (2) they are seen as

differing greatly in size. The perspective information about depth in the 2-D image of the Ames room at the special viewpoint leads to the perception of a normal rectangular room, with corners at equal distances. If this were actually true, the people would have to be enormously different in actual size to account for their differences in retinal size, and this is what is perceived. Familiar size information suggests that the men are about the same size, but this possibility is overwhelmed by the evidence from perspective, which appears to completely dominate perception in this case.

5.6 Development of Depth Perception

As adults we perceive the distance to surfaces and their orientations in space with no apparent effort. Is this because we were born with the ability, because it matured autonomously after birth, or because we learned it and have become so well practiced that it has become automatic? Experiments with young children have begun to provide answers to this intriguing question.

Some of the earliest and most dramatic studies demonstrating that infants have functional depth perception early in life employed an apparatus called a visual cliff. The visual cliff consists of a glass-topped table with a central board across the middle, as illustrated in Figure 34. Textured surfaces can be placed below the glass on either side of the central board at varying distances beneath it, and the lighting can be arranged so that the glass is essentially invisible. In the classic visual cliff experiment, Walk and Gibson placed a textured checkerboard surface directly below the glass on the "shallow" side and an identical texture on the floor, 40 inches below the glass on the "deep" side.

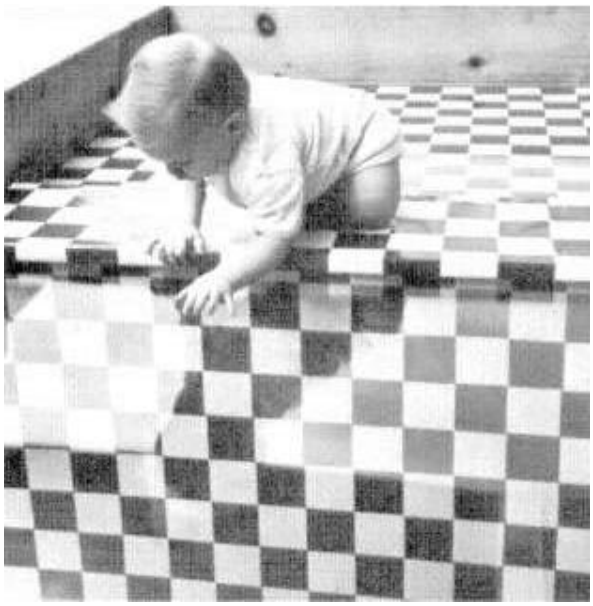

Figure 34 An infant on the visual cliff. Infants are placed on the central board over a thick sheet: of glass that covers a textured surface on each side, one of which is shallow, lying just beneath the glass and the other of which is deep. When beckoned by their mothers, most infants will cross the shallow side, but few will cross the deep side.

The experimenters then asked the baby's mother to put her child on the central board and to try to entice the baby to crawl across either the shallow side or the deep side. If the baby is able to perceive the distance to the two textured surfaces, it should be willing to crawl to its mother over the shallow side but not over the visual cliff. Of the 36 children tested from ages 6 to 14 months, 27 could be persuaded by their mothers to leave the central board. Of these, all 27 crawled across the shallow side at least once, whereas only three attempted to negotiate the visual cliff. This result shows that by the time babies can crawl at age 6 to 12 months, they have functional depth perception.

To find out whether babies who are too young to crawl perceive depth, the heart rates of infants as young as two months old on a visual cliff apparatus were recorded. It was found that when two- to five-month-old infants were placed on the shallow side, there was a small but not statistically reliable change in heart rate. When they were placed over the deep side, however, their heart rates slowed significantly. The fact that heart rate decreased rather than increased suggests that the infants were not afraid on the deep side—for that would have increased their heart rates—but were studiously attending to the depth information. Consistent with this interpretation, they cried and fussed less on the deep side than on the shallow side. It therefore appears that children as young as two months old are already able to perceive depth but have not yet learned to be afraid in the cliff situation.

Stereoscopic and Dynamic Information

Once infants are able to converge their eyes, they can develop stereoscopic vision. Several different procedures indicate that depth perception based on binocular disparity develops at about 3.5 months of age, slightly after direct methods indicate that the ability to converge properly develops.

As noted earlier, there are a variety of dynamic sources of information about depth: motion parallax, motion gradients, looming, etc. Perhaps the best candidate for inborn dynamic depth information is looming (the motion gradient of an approaching surface) because it applies to objects coming toward a stationary observer. Newborns cannot really move themselves enough to produce extensive self-generated optic flow, but biologically significant objects—such as Mom and their own limbs— do move toward and away from babies as soon as they are born. Human babies one to two months old respond to looming (that is, visually expanding) objects with appropriate defensive reactions. They push back with their heads and lift their arms in front of their faces when contours expand in their field of view. These researchers interpreted such findings as indicating that depth information from a looming object is present very early in life, perhaps even at birth.

Clowes, M. B. (1971). On seeing things. Artificial Intelligence, 2, 79-116.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture of the cat's visual cortex. Journal of Physiology (London), 160,106-154.

Huffman, D. A. (1971). Impossible objects as nonsense sentences. In M. Meltzer & D. Michie (Eds.), Machine intelligence (Vol. 6). Edinburgh, Scotland: Edinburgh University Press.

Koenderink, J. J., van Doorn, A. J., & Kappers, A. M. L. (1992). Surface perception in pictures. Perception & Psychophysics, 52(5), 487-496.

Koenderink, J. J., van Doorn, A. J., & Kappers, A. M. L. (1996). Pictorial surface attitude and local depth comparisons. Perception & Psychophysics, 58(2), 163-173.

Longuet-Higgins, H. C., & Prazdny, K. F. (1980). The interpretation of a moving retinal image. Proceedings of the Royal Society, London, B, 208, 385-397.

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

Marr, D., & Poggio, T. (1977). Cooperative computation of stereo disparity. Science, 194, 283-287.

Marr, D., & Poggio, T. (1979). A computational theory of human stereo vision. Proceedings of the Royal Society of London, B, 204, 301-328.

Prazdny, K. (1980). Egomotion and relative depth from optical flow. Biological Cybernetics, 36, 87-102.

Stevens, K. A. (1979). Surface perception from local analysis of texture and contour. Unpublished Ph.D. Dissertation, MIT, Cambridge, MA.

Waltz, D. (1975). Understanding line drawings of scenes with shadows. In P. H. Winston (Ed.), The psychology of computer vision (pp. 19-92). New York: McGraw-Hill.

Witkin, A. P. (1981). Recovering surface shape and orientation from texture. Artificial Intelligence, 17, 17-45.