Toward a Neural Understanding of Visual Surface Representation

K. NAKAYAMA* AND S. SHIMOJO†
The Smith-Kettlewell Eye Research Institute, San Francisco, California 94115

Independent Visual Modules?

Owing to the exploitation of available techniques in modern neuroanatomy and neurophysiology, an increasingly detailed picture of the primate cortical visual system is emerging. First, a topographic map of the retina appears to be duplicated many times in the cerebral cortex (Hubel and Wiesel 1965; Zeki 1978), and more than a dozen separate cortical maps of the retina have been described (Maunsell and Newsome 1987). Second, there seems to be a set of parallel anatomical pathways mediating visual function. This segregation begins as early as the retina and appears to continue through many cortical areas. Segregation is manifested particularly in the responses of individual cells. Cells differ in their selectivity to various characteristics of the visual image: binocular disparity, line orientation, motion direction, size, color, etc. (for review, see Livingstone and Hubel 1987).

This diversity in the coding properties of neurons and their anatomical distribution has given credence to the idea of relatively independent modules emerging early for the processing of different attributes of visual images. Such notions underlie various models, including those encoding motion (Adelson and Bergen 1985), binocular disparity (Sperling 1970; Julesz 1971; Nelson 1975; Marr and Poggio 1976), and color (Land 1977). What characterizes each of these models is a distinct module, composed of simple within-module elements. Thus, color and contour are not influenced by depth, depth is not influenced by color, etc. The most explicit support for this notion came from Julesz's demonstration of "purely cyclopean" perception (Julesz 1971). The idea of modularity also gains credibility from studies examining the selective damage to visual structures in the monkey. Restricted lesions can selectively knock out the perception of motion (Newsome and Pare 1988), light spots as opposed to dark (Schiller et al. 1986), and color (Schiller et al. 1990).

The assumption of independence between modules in fact has had a certain heuristic value. Yet, the success of such an approach should not blind us to the rather primitive nature of our current understanding of vision and the need to ask new questions. If we do have such modules, how are their activities coordinated so that we see as we do? Ordinarily, we see textured and colored surfaces bounded by contours. We do not see color and motion separately disembodied from objects and surfaces. The problem of how specific visual attributes or features are combined has been identified as the "feature-binding problem," and a number of hypotheses have been advanced to address this issue, often positing specific neural or cognitive processes to perform such tasks (Treisman and Gelade 1980; Crick 1984).

Our immediate approach to this seemingly difficult problem is to sidestep it, at least temporarily. Rather than attempting to build up perception from the outputs of these hypothetically independent neural modules, we think it perhaps more fruitful to look out into the world, to examine how the optical properties of the real world constrain the structure of images formed on our retina. Then we ask how such images might be best interpreted for the brain to "see" as it does. This approach rests heavily on an appreciation of ecological optics (Gibson 1950) and is broadly related to computational vision (Marr 1982). It also relies on phenomenological observations, often used by Gestalt psychologists (Koffka 1935).

Natural Constraints and Surface Representation

In contrast to the approach that seeks to build up perception from sets of atomic features, based on physiological and psychophysical experiments, we suggest that it is as important to evaluate how the visual system might go about encoding real-world scenes by looking at the natural-optical constraints of this world. Furthermore, to understand the interactive nature of early vision, as opposed to its modularity, we think it of great importance to understand the encoding of surfaces.

A look at surfaces in the real world immediately reveals one of the most important facts about vision: Closer surfaces occlude more distant surfaces. The amount of occlusion varies greatly, depending on seemingly arbitrary factors—the relative positions of the distant surface, the closer surface, and 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.

Present addresses: *Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138; †Department of Psychology, College of Arts and Sciences, University of Tokyo, Komaba, Meguro-ku, Tokyo 153, Japan.

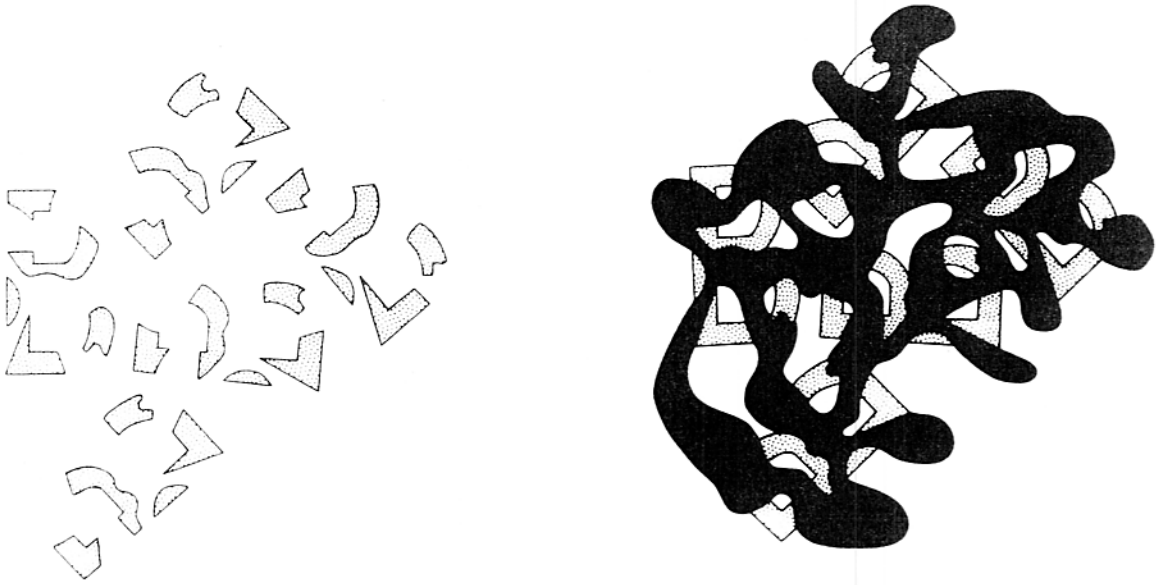


Figure 1. On the left are a series of letter fragments, made up of letter Bs that have been occluded by an invisible occluder. On the right are the exact same set of remaining fragments of the letter Bs, but here the black occluder is visible. Note that the letters can be recognized on the right but not on the left. (Reprinted, with permission, from Bregman 1981.)

We begin by thinking about the problem of object identification under occlusion. If we consider the image of several objects that have been occluded, some important things become apparent. Not only can an occluder conceal information from the viewer about the objects themselves, rendering them less visible for purposes of recognition, but it can also have two more effects which could be as damaging or more so. First, it could separate different parts of the same object from each other. Second, it could introduce spurious bounding contours around the object fragments themselves.

Perhaps these two points can be best appreciated by looking at the left portion of Figure 1 (from Bregman 1981). Here, we see a set of image fragments, remaining portions of letter Bs, scattered about. These Bs have been covered by an "invisible" occluder. Recognition is virtually impossible. Now examine the right portion of Figure 1, where the occluder is visible. Here, recognition becomes possible, almost easy. In each case, exactly the same fragments of the letters are visible, yet the difference is striking.

What accounts for this difference? To return to our points made earlier, it appears that without the visibility of the occluder, the separate fragments of each letter are not appropriately grouped. The invisible occluder divides the letters, and it is not possible to know which of the separate fragments go with which. With the visibility of the occluder, however, they are more correctly grouped.

Border Belongingness and Surface Continuation

To begin to understand what is going on, we would like to introduce some simple but useful concepts, similar in part to ideas outlined over 50 years ago by Koffka (1935). First is the notion of border "belongingness" or "ownership." If there is a common border between two image regions, we argue that it is important for the visual system to make an early distinction as to which image region "owns" the common border. As an example of border ownership, consider the various borders in Figure 1. On the left, the contours surrounding the stippled letter fragments clearly belong to the fragments themselves. Each fragment appears as complete within itself. They do not group with other fragments. As a consequence, it is very difficult to make out the letters. It is very different, however, for the case shown on the right. Here, the contours that are common to the black region and the stippled region are seen as belonging to the black region and not to the stippled region. As such, the stippled region is phenomenally bounded only where it meets the white background, not where it meets the black region. At this boundary the stippled region appears to continue "amodally" (Michotte 1954; Kanizsa 1979) behind the black region. Now each fragment is no longer complete within itself because it is not entirely enclosed with its own bounding contour. We argue that this lack of "boundedness" enables it to link up with other similarly unbound image regions, thus forming larger units which are otherwise fragmented in the image.

Why does the boundary between the black and stippled region "belong" to the black region? Note that there is a T-junction where these two image regions meet. Ever since Helmholtz (1910), such junctions have been hypothesized to provide evidence for depth and occlusion, with the line forming the top of the T being the occluding contour (see Guzman 1984). Thus, the T-junction provides evidence that the black edge is an occluding contour. As such, it belongs to the black region and not to the stippled region.

We postulate that if a given image region is bounded by a contour that does not belong to it, the surface corresponding to this region is not bounded there. As such, it will be seen as continuing behind the bounding contour that belongs to its neighbor.

To summarize, contour ownership or belongingness is critical for depth and to surface continuation. Under most circumstances, if a surface is seen as nearer, the common boundary belongs to this nearer surface. This optical constraint of occlusion, in turn, ensures that the neighboring distal region is seen as continuing behind.

Something very similar can be seen in Figure 2, which should be first viewed monocularly, not as a stereogram. Here, it should be clear that one sees a large letter C, occluded by a gray patch in all three panels. Again, as a consequence of the T-junction, the gray patch is seen as in front, and it thus owns the common border. This means that the two letter fragments are effectively "unbounded" at this point and they are seen to continue behind the patch. This enables us to see the C as a single object.

What would happen, however, if we were to reverse the depth seen in this figure and to make the gray patch appear in back rather than in front? If our hypothesis regarding depth and contour belongingness is correct, this reversal should have a major effect. Under these conditions, the contour would no longer belong to the gray patch because it is no longer in front. The ownership would be transferred to the fragments of the letter. As a consequence of this ownership transfer, the fragments would be effectively bounded on all sides. As such they could not be linked to form a single letter.

Our strategy has been to use stereopsis because it can often overcome monocular depth cues, such as T-junctions. To see this, the reader should fuse the images shown in Figure 2 as stereograms. Supporting our hypothesis, it should be clear that there is a dramatic difference when the gray patch is seen as behind. The figure is seen as two image fragments, two U-shaped objects, separated. It can hardly be seen as a single object. If one views the stereogram in the reversed configuration, however, so that the gray patch pops out of the page, we then see the fragments linked and we are again aware of the unified C as expected.

The data presented so far are purely phenomenological. We argue that such observations provide important and valid evidence for an understanding of perception. Yet, we have also linked such observations to more objective measures of pattern recognition (Nakayama et al. 1989). In brief, we presented a set of human faces to naive experimental subjects, asking them to remember these faces for a recognition test to take place immediately thereafter. During the testing phase, the image of the face was incomplete, replaced in part by a set of horizontal "noise" strips. An example of such a stimulus can be seen in the stereograms in Figure 3. Two conditions with identical image information were presented in the stereograms. In one case, the strips of faces were seen in back, in the other case they were seen as in front, occasioned by a simple right-eye left-eye reversal of images. All subjects reported that the faces were more recognizable when seen in the back plane. The separate pieces of the face appeared to continue behind the occluders and recognition appeared much more certain in comparison to the case where the strips containing the face fragments were in front. These impressions were clearly confirmed in the performance accuracy in the recognition task. Recognition accuracy was significantly better if the horizontal pieces of the face were seen in back rather than in front. Note that the two depth conditions were identical in terms of pictorial information for face recognition, different only in the sign of disparity. This ensures that relative depth of surfaces is critical for

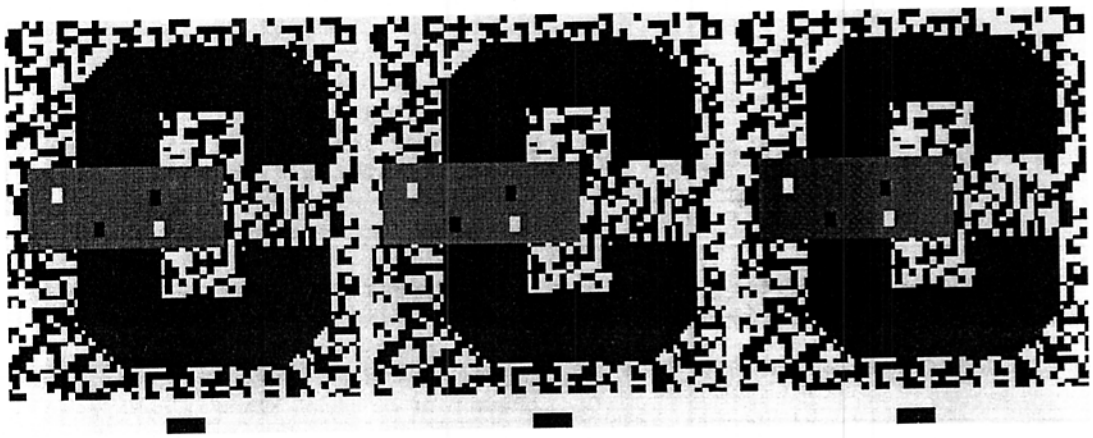


Figure 2. Note that the letter C is easily visible when not viewed as a stereogram. It is as if the hidden portions of the C continued behind the gray patch. If viewed as a stereogram such that the gray patch recedes in depth, the perception of the C disappears. One only sees two U- shaped objects, unconnected. If one views the reverse stereogram such that the gray patch pops out in depth, the C will be visible again. (Adapted from Nakayama et al. 1989.)

Note: To view this 3-section stereogram (as well as others in this paper) in its normal configuration, persons who cross their eyes should fuse the left and center images. Those who diverge to fuse should view the center and right images. To view the stereograms in the reversed configuration, do the opposite. Thus, cross-fusers should fuse the center and right images.

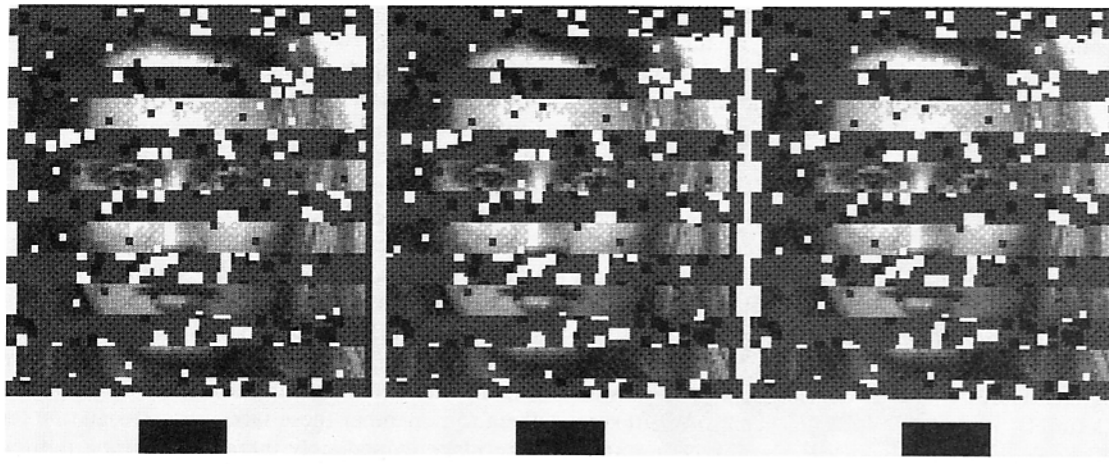


Figure 3. If one fuses these stereograms such that the faces are seen in the back plane, they are much easier to recognize than if they appear on the front disparity plane. Compare normal and reversed disparity case. (Adapted from Nakayama et al. 1989.)

image segmentation and grouping, which is a necessary stage of processing for object recognition.

Figure 4 summarizes our current understanding of how borders and surfaces are encoded. If the image patch labeled M is coded in front, contours C and C′ are owned by and belong to region M. As a consequence, surfaces corresponding to regions U and L are not bounded at C and C′. As such, they can have the opportunity to continue behind M. Contours A and A′ as well as B and B′ are then perceptually linked. Conversely, however, if region M is coded in back, then C and C′ belong to regions U and L, respectively, and A and A′ as well as B and B′ can no longer be linked.

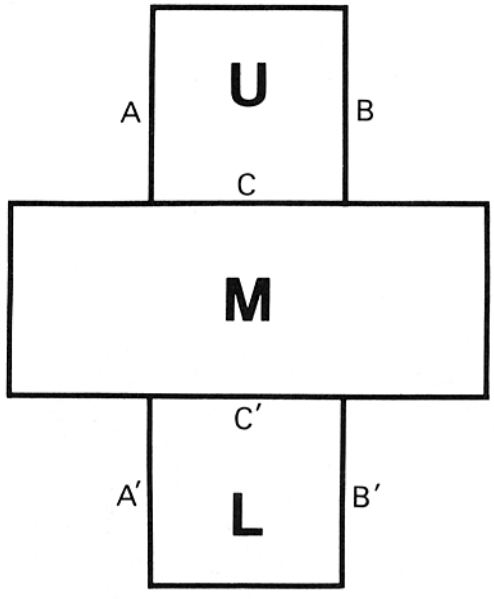


Figure 4. Schematic diagram illustrating hypothesized relationship between depth, contour ownership, and completion of image fragments behind occluders. If M in this diagram were to be encoded in front, M would own contours C and C′. As such, regions U and L would appear to continue behind M because they are effectively "unbounded" at C and C′. Then contours A and A′ as well as contours B and B′ would tend to be linked. If, however, M were to be encoded as behind, its common border with regions U and L would be owned and belong to U and L, respectively. As such, U and L would be effectively bounded and no perceptual linking between them would occur. (Reprinted, with permission, from Nakayama et al. 1989.)

Can we come up with a plausible neural mechanism to understand such large differences in perception? We can only speculate at this point. Yet, it is curious to note that frequently encountered cell types in striate and extrastriate cortex could combine their outputs to implement some of the seemingly "intelligent" distinctions mentioned above. For example, one of the most important decisions to be made in this context is whether a given line terminator in an image represents a real line termination in the three-dimensional world or whether it is more likely to continue behind another surface (see Shimojo et al. 1989).

Cells originally labeled "hypercomplex" by Hubel and Wiesel (1965) and now more recently designated as "end stopped" might indicate that a line ended in an image. Such cells respond less vigorously for long lines and fire more vigorously if the line stops in the image plane. However, such cells alone could not signal whether a real line might continue behind an occluder or whether it actually stops in the real world. Combining the output of these cells with cells sensitive to depth, however, could resolve the issue for the visual system. Consider a disparity-tuned, end-stopped cell with a receptive field coincident with contour B′ in Figure 4. It would fire vigorously because the line stops abruptly in the image. Suppose, however, that such a cell were to feed its output to higher order cells which also received input from cells that encoded the depth of region M. If M were coded as closer, then such higher order cells could indicate that the line continued behind. Conversely, if M were coded as further, such cells would provide information as to whether the line really did end in the three-dimensional world.

Up to now, we have only considered the possibility of surfaces that are opaque, showing how surfaces and contours might be seen as continuing behind occluders, depending on local depth signals. In addition, we suggested how plausible signals from different early mechanisms might interact to encode surface termination versus surface continuation.

Transparency and Color Spreading

Interaction between different putative early local mechanisms becomes even more apparent in our studies of surface formation. In these studies, we show that the perception of contour, depth, and color are closely coupled and are critically related to whether a given configuration is perceptually encoded as transparent or opaque.

Recently, in considering factors involved in space perception, there has been an implicit dichotomy between physiological versus cognitive cues to depth. These attitudes may have stemmed in part from the existence of neuronal units selective to binocular disparity (Barlow et al. 1967) and the lack of comparable evidence for pictorial cues. Such a bias toward stereopsis might lead one to assume that binocular disparity signals alone are sufficient to specify depth and that so-called cognitive occlusion cues were perhaps acquired later, built on early experience with stereopsis or perhaps motion parallax. If disparity did indeed play such a dominant role, it would seem to follow that if one reversed the disparity in a simple stereoscopic scene, the depth of elements in the scene should undergo a corresponding reversal. This is indeed the case for a large set of contrived stereograms. However, a closer examination of many types of other simple configurations indicates that local disparity alone is far from sufficient to specify the depth values of points in many scenes. As such, the reversal of right and left eye views can have effects very different from an expected exchange of depth values.

Examine the cross, shown in Figure 5. Although appearing as a cross if viewed monocularly, it should be clear that there is a difference in relative disparity between the vertical and horizontal limbs of the cross if fused stereoscopically. In such an untextured figure, classical stereopsis makes no obvious prediction regarding the depth value of the untextured region at the very center of the cross. Will it take a value intermediate to the front and back region of the stereogram—i.e., some form of averaging? If so, might not one assume the average would be weighted more by the depth defined by the vertical limb of the cross? We mention this possible expectation, because disparity signals along the vertical line defining the vertical limb are certainly closer to the center than the short vertical lines far out on the horizontal limb. As such, they might be expected to contribute more, that is, if we assume simple local propagation of depth signals.

If the reader fuses the stereograms seen in Figure 5, it should be obvious that perception does not correspond to any of the expectations outlined above, and yet the answer is clear and unambiguous. The middle of the cross is seen in front even when the disparity of the closer vertical limb indicates that this region is in back.

More telling are the specific details of the perceived surface layout. If the ends of the horizontal portion of the cross are nearer, one sees a horizontal bar in front of a vertical bar. In the reversed case, the vertical bar is in front. Yet something new is also seen. Although each monocular cross is homogeneous, this homogeneity is now broken by an obvious contour when viewed stereoscopically. An illusory occluding contour "com-

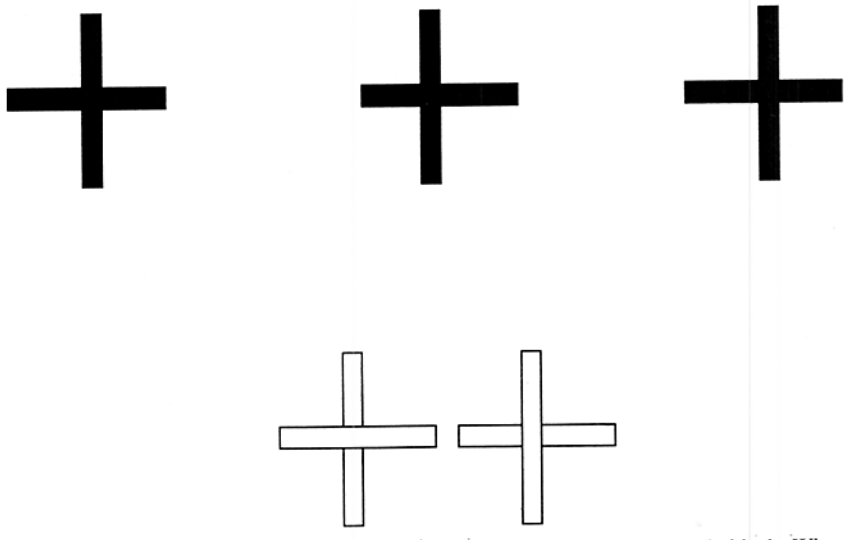


Figure 5. Top is a cross when viewed monocularly. Areas within the cross are homogeneously black. When viewed in its normal stereo configuration, however, one sees a black horizontal bar in front of a vertical bar. In the reversed stereo configuration (right-eye left-eye reversal), one sees a vertical bar in front of a horizontal bar. Bottom two figures portray what is seen in the fused stereograms, showing the subjective occluding contours.

pletes" the vertical or horizontal bar in front. A schematic description of these two perceptual outcomes can be seen in the bottom of Figure 5.

Although the perception of this stereogram reminds us that disparity is certainly important and can propagate when necessary, more importantly, it also tells us that local disparity alone is not sufficient to explain what is going on in this scene. No model based on local disparity signals can account for these results. It appears that when the visual system is confronted with this simple stereogram, it acts by "constructing" a set of depth values that correspond more to possible configurations in the world rather than performing an algebraic summation of local disparity signals. Thus, instead of providing a smoothed output profile of depth, as current neural models might suggest, the visual system comes up with a concise and seemingly intelligent answer: in this case, one surface occluding another.

This "problem solving" or "inferential" aspect of perception is perhaps even more dramatically supported in the next demonstration shown at the top of Figure 6. Here, the exact same cross is colored red and embedded in a white cross which itself has no "within" differences in disparity. As such, one might expect to see another version of what we have just described: a horizontal bar in front of a vertical bar, lying within a white cross.

The perception, however, is strikingly different. No longer does one see two bars, the horizontal in front of the vertical. Instead one sees a red transparent surface, sitting in front of a white cross. Furthermore, even though there is no red coloring in the black regions around the center of the disk, one sees a distinct reddening of this area, enclosed by a circular "subjective" contour bounding this disk. All of these descriptions are most succinctly summarized by saying that one sees a red filter in front of a white cross.

Reversing the disparity (by looking at the other pair of the three half-images) also leads to something unexpected from our previous discussion. Instead of seeing a vertical bar in front of a horizontal bar, one sees a flat circular red disk, lying behind, seen through a cruciform aperture. Note that the spreading of red color into the black region is essentially absent and that the surface color is matte rather than transparent. It is as if one sees a Japanese flag behind a window.

Another configuration that yields a very similar set of results can be seen by inspecting the stereogram seen at the bottom of Figure 6. In the normal configuration, one sees a transparent green square in front. Note the spread of color into the black background. In the reversed configuration, one only sees pieces of a green square viewed through four circular apertures. Here the green region is no longer transparent and it does not spread into the black background (for details, see Nakayama et al. 1990).

These very striking changes in perception with small changes in the display are puzzling if we think only about local "within-module" signal processing. How can such a small change in color and configuration lead to such a very different global perception? In particular, how can just the addition of the outer limbs of the cross, combined with changing the color of the inner limbs, lead to such a dramatic change in appearance? A small change in the configuration leads to the creation and/or destruction of contours, depth, and color. Each of these seemingly primitive features appears as subordinate to the global interpretation of the scene as a set of surfaces.

Discontinuity Edges, Transparency, and Contrast Conditions

To extend this observation even further, we choose what is perhaps the simplest one-dimensional configuration (Fig. 7). The pattern consists of sets of horizontal bars, which if they were not divided into white and gray regions, would have no depth when viewed stereoscopically because the ends of these bars all have the same disparity. What makes these bars interesting is that the vertical dividing border between the gray and white regions has crossed disparity. As a consequence, it should be seen as in front. At issue is the perceived depth of the intermediate regions. Again, these portions of the image are untextured, and classic stereopsis has no explicit prediction as to their perceived depth. If one takes the now popular view, however, that the visual system is looking to maximize smoothness in terms of interpreting the image (Hildreth 1984; Poggio et al. 1985), one might expect some form of interpolation: either a set of tilted planes if continuity alone is maintained, or a bowed plane (similar to a spline fit) if both continuity and smoothness are preserved. These expected interpretations (labeled 1 and 2) are shown in Figure 8A.

Our perception is quite different from either of these expectations. Instead of the kink or bulge in a continuous surface as might be predicted from an interpolation of disparity-based depth signals, two distinct planes in depth emerge. We see a transparent surface in front of white regions in back. In the top stereogram of Figure 7, we see a transparent filter covering the central region as schematized in the bottom of Figure 8A (labeled "perceived." We also see the spreading of color into the black regions as well as vertical subjective contours, appearing to contain the spreading color. If we examine the lower set of stereograms shown in Figure 7, something very different happens. Here we have exactly the same stereograms as shown in the top of Figure 7, except that the luminances have been exchanged. Regions that were gray are now white and vice versa.

Such a reversal of contrast has a dramatic perceptual effect. No longer is the center region transparent, but transparency has shifted to the outer gray regions. Also striking is the concomitant change of depth (see schematic drawing in Fig. 8B). In this demonstration, with exactly the same disparity, this central region is no longer seen in front, but in back; the outer regions are now in front. Thus, by the simple manipulation of luminance, we obtain a radically different perception of

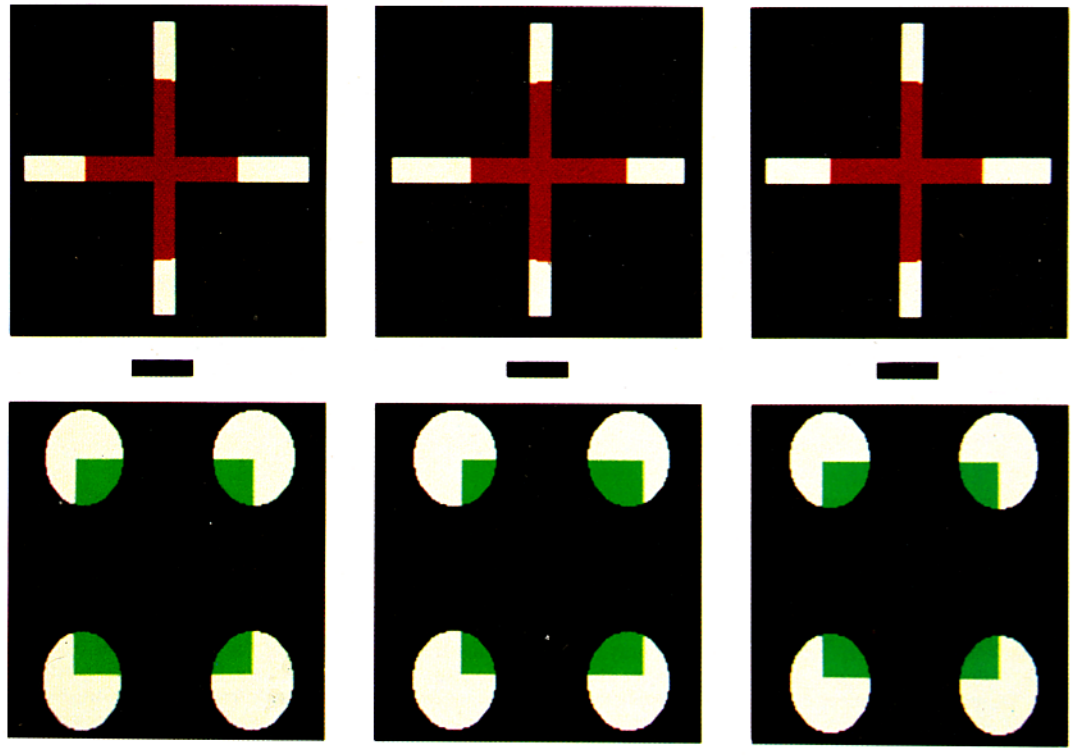


Figure 6. Top row shows the stereoscopic modification of a configuration introduced by Redies and Spillmann (1981). Inner red cross has the same disparity as the cross shown in Fig. 5, yet the perception is very different. Instead of seeing a horizontal bar in front of a vertical bar, one sees a transparent colored disk, covering the central cross. Thus, color spreads out into the dark regions nearby, confined by a circular "illusory" contour forming the outer perimeter of the transparent disk. When the stereograms are viewed in their reversed configuration, one perceives a more distant disk, seen through a cruciform aperture. Color is matte and no longer spreads. Bottom row shows the stereoscopic modification of the Varin (1971) configuration where transparency, opacity, and subjective contours are also determined by the sign of disparity. (Adapted from Nakayama et al. 1990.)

depth. These results support Metelli's (1974) observation that the region of intermediate luminance is seen as the transparent surface. They go further, however, in indicating that hand in hand with this switch in the region seen as transparent, there is a corresponding switch in depth. A region previously appearing opaque and in back now looks transparent and in front. Thus, a small change in luminance can have a profound influence on perceived depth. It is as if the visual system with exactly the same disparity signals makes a decision from various sorts of information (including luminance) to come up with a coherent interpretation of the scene.

This brief series of results indicates that we cannot consider depth, color, and contour as elemental primitives, acting within confined regions of the visual field according to within-module local rules. There is far too much interaction between these seemingly distinct primitives such that small changes in one can influence not only the surface interpretation, but also the same primitives themselves. Thus, changes in disparity can determine whether a surface is seen as transparent or opaque, and this in turn can determine whether subjective contours are visible and whether color is confined to a particular region or spreads elsewhere. In short, it seems that to understand surface perception we need to come to grips with terms like intelligence or inference.

It appears that our vision acts as if it were a detective, piecing together image data to come up with the most plausible interpretation of the scene. Such a view is not new. It was espoused forcefully by Helmholtz (1910) in his discussion of "unconscious inference" in perception, and this general view has had a number of contemporary adherents (Gregory 1970; Rock and Anson 1979; Hochberg 1981).

With the hope of delving more deeply into the specific nature of visual intelligence, let us again think about the striking changes in perception as we made the seemingly small modification from the stereogram in Figure 5 (the crossed bars) to the transparent surface seen in the top of Figure 6. How does such a small change in target configuration (with no changes in disparity) lead the visual system from seeing the inner cross as two crossed bars in one case (Fig. 5) to something that seems quite far-fetched, constructing in a seemingly unexpected way a transparent surface (as in Fig. 6)?

Generic View and Conditional Probabilities

Thinking in terms of the detective analogy, it would seem that the visual system would need to ask a number of questions. First, could a particular interpretation

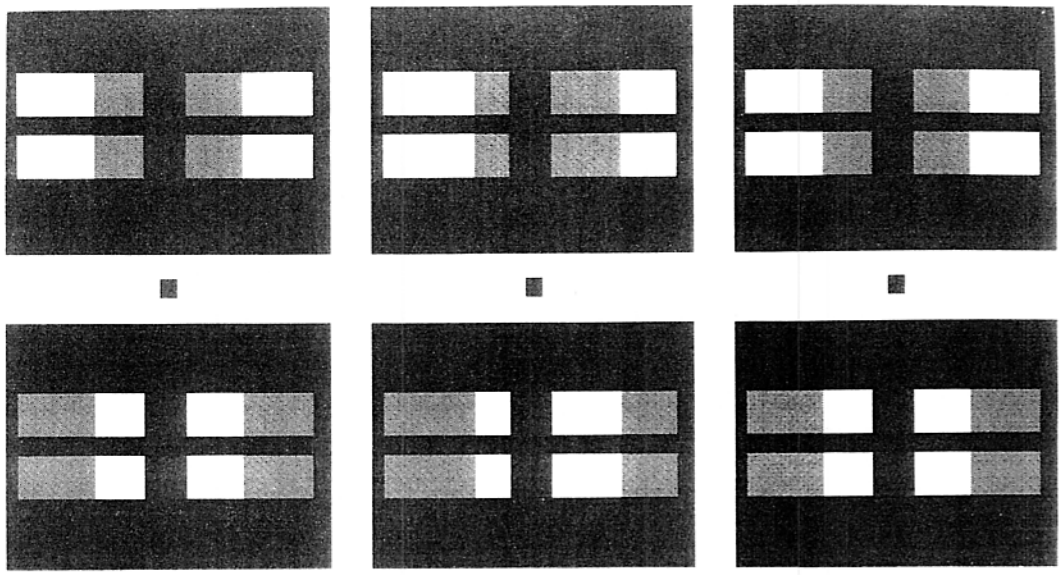


Figure 7. Top stereogram shows four horizontal bars whose right and left endings have the same disparity. These bars are each divided into two regions, the border of which has crossed disparity when viewed in the regular configuration. Thus, it should appear in front. What is perceived is a transparent gray figure covering the central region of the two bars; this colored region is bounded by "illusory" contours, oriented vertically. Bottom stereogram is the same except the gray and white have been exchanged. Despite the identical disparity, this simple switch in luminance leads to a very different perception. Now there is a closer surface on the outer edges of the figure; the closer surface in the inner region is gone. Note that this stereogram should be viewed only in its normal, not its reversed, configuration.

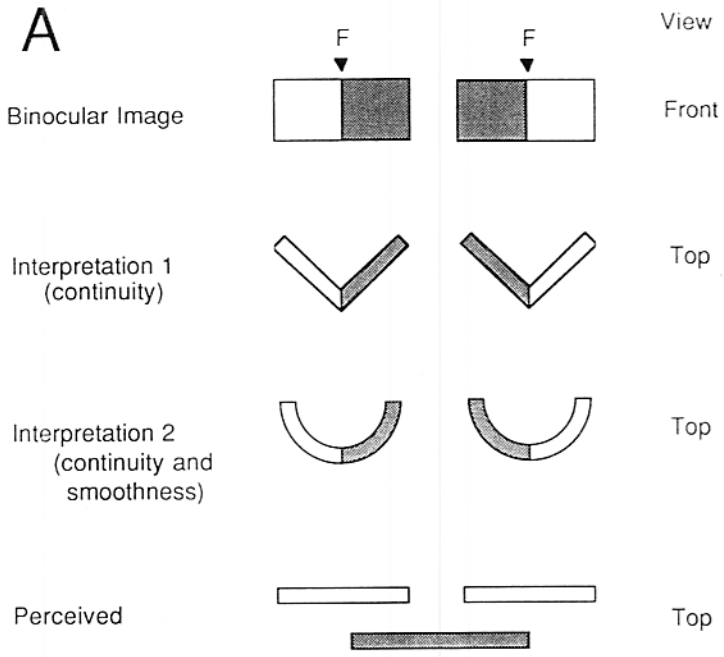
under consideration give rise to the facts at hand? Second, given the "reality" corresponding to a given interpretation, could it have plausibly given rise to these same facts?

The first question can be rephrased as follows. Does the perceived configuration correspond to the image data? The answer is yes for the perceived transparency configuration. One can (in retrospect) understand how the particular spatial pattern perceived (shown in the top of Fig. 6) could have given rise to the image data shown in the stereograms. The luminance and disparity relations conform to the underlying physics of such a scene. Yet, the answer to this question is not sufficient because there was at least one other interpretation; namely, the one we expected from an analysis of Figure 5 (two crossed red bars in front of a white cross). Why did the system choose the transparent disk rather than this second and seemingly reasonable interpretation?

The second question is perhaps more critical. Given a particular scene, how likely could it have given rise to the image data? Such conditional probabilities are dictated by the geometry and optics of scene viewing. In this context, we briefly introduce two interlinked concepts characterizing the totality of vantage points in space from which this scene could be viewed: the generic and the accidental views (see Richards et al. 1987; Koenderink 1990). By generic viewpoint, we mean those positions where, for arbitrarily small perturbations of this vantage point, there are no qualitative changes in the image. Conversely, by accidental viewpoint, we mean those very few vantage points in space where small perturbations will lead to qualitative differences in the image. Our hypothesis is that, given a choice, the visual system will show a strong preference for a scene interpretation based on the generic rather than the accidental view.

As an illustration, consider the two figures seen in Figure 9. We see a square on the left and a cube on the right. Why do we only see a square on the left and not a cube? In terms of fitting the image data, it is certainly possible for a cube to be turned so that one just sees its face. Yet, if we appeal to the notion of the generic view, stated above, it is clear that if the object were a cube, the image on the left could only have arisen from a very small number of accidental viewpoints out of an infinite number of generic viewpoints. On the other hand, if it were a square, it could be seen from many (generic) viewpoints. We argue that the visual system has implicit knowledge of the conditional probabilities associated with these various views and hence rejects any interpretation based on an accidental view.

Now we use the same reasoning to explain why the visual system chooses the transparency interpretation seen in Figure 6 (top). Here, there are at least two physical interpretations, but only the transparent disk interpretation survives the generic view test. If the surface configuration in the real world was indeed a set of red crossed bars in front of a white cross, the binocular image seen in Figure 6 could have arisen only from a privileged or accidental view, such that the observer's position led to an exact alignment of image contours. It is obviously accidental because a small perturbation of viewing position leads to a very different image (note the top stereogram in Fig. 10 and compare it with the



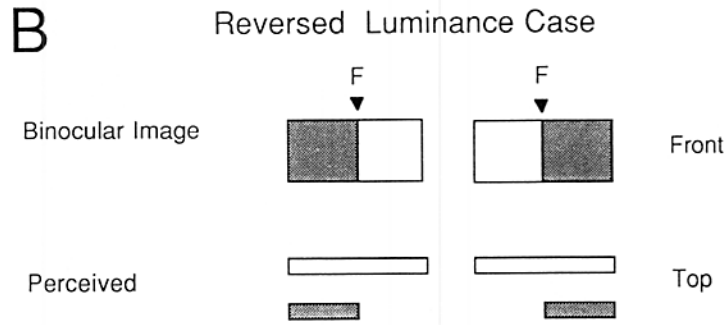


Figure 8. Schematic illustration of expected and actual perceived depth of single one-dimensional stereograms. In *A* we show the case for the top stereogram shown in Fig. 7, including the binocular image, the expected depths (from considerations of continuity and smoothness), and the central transparent surface perceived in front. F refers to the edge coded in front by binocular disparity. In *B* we show that the simple reversal of luminance leads to a dramatic reversal of depth. Now the transparent surface moves from the central region to the periphery.

top one in Fig. 6). If, on the other hand, the surface configuration was a transparent disk in front of a white cross, the image could have arisen from a generic view: Here for a similar change in vantage point, there is no qualitative change in image data. Note the invariant nature of the perception (aside from the color difference) as one compares the bottom stereogram in Figure 10 with that shown in the top of Figure 6.

Like a good detective, the visual system does not simply jump to conclusions but avoids a major pitfall. It does not base its "reasoning" on assumptions that are highly improbable. As such, it rejects any interpreta-

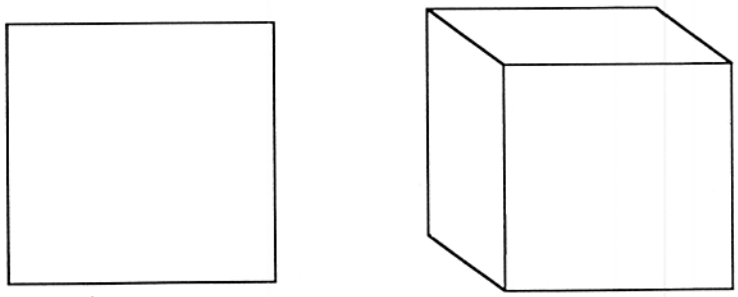


Figure 9. Accidental vs. generic view I. We hypothesize that we do not see a cube on the left because such a view can only arise from a small number of accidental vantage points (see text).

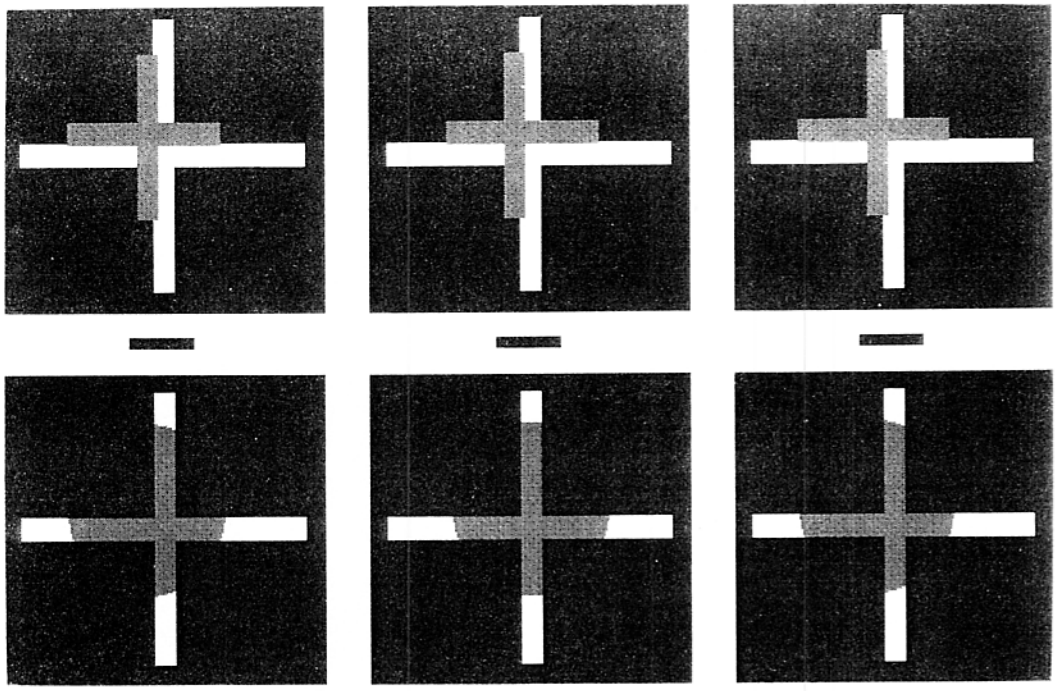


Figure 10. Accidental vs. generic view II. Binocular image of two possible surface configurations that could have given rise to the scene in Fig. 6 (top) under a perturbation of vantage point. Top stereogram shows the binocular image if the scene consisted of two crossed bars lying within a white cross. Bottom stereogram shows the binocular image of a transparent disk in front of a white cross. Note the large qualitative change in image for the first (crossed bars) but not the second (transparency) configuration.

tion based on the assumption of an accidental view.

Where in the brain might such "intelligence" reside? Is it in the visual system or is it higher up, part of some generalized cognitive function, as has been postulated by antiphysiological theorists (Gregory 1970; Rock and Anson 1979)? If we think of the word detective, we think of words like executive or agent "homunculus." If we look at the anatomy of the brain, it is not obvious from the structure of the system that there is an apparent "place" for this executive or agent, at least for vision. A more plausible alternative is that inference or intelligence must be distributed and that there needs to be no agent as such. Somehow, by the operation of many small distributed units, each showing some limited inferential capacities, the work of the apparent detective is done. Thus, there is no inherent reason why this process in vision cannot begin very early, possibly in areas now undergoing intensive neurophysiological investigation.

da Vinci Stereopsis and Possible Neural Mechanisms

To support this alternative way of thinking, we present our final set of demonstrations. Here, we make the case that at least one class of such inferences must begin very early in cortical information processing, possibly as early as primary visual cortex. To appreciate the logic of this argument, consider the case of surfaces viewed by the two eyes, not just one. Ever since Leonardo da Vinci, it has been known that a closer surface occludes a background surface differentially in the two eyes (see Wheatstone 1838). Thus, there will be background points to the left of an occluding surface that are visible only to the left eye, and correspondingly, there will be background points to the right of an occluding surface that are visible only to the right eye. Thus, points P_L and P_R shown in Figure 11A are visible only to the left or right eye, respectively. It should be clear that such unmatched binocular points can hardly be avoided if we view real-world scenes. An interesting question is how the visual system might interpret the presence of such left-eye-only and right-eye-only points in a simple stereogram where there are no disparity-based depth cues and where the occluder is not physically present in the display. Would the visual system make the correct "unconscious inference" that such an occluding surface did still exist?

To see whether this is indeed the case, we created a stereogram with four unpaired image points, where two image points were seen only by the left eye and where the other two image points were seen only by the right eye. The positioning of such points is illustrated in Figure 11B. The only plausible scene in real life that could have given rise to this configuration of binocular and monocular points would be a surface in front (partially delineated by the dashed lines in Fig. 11C).

What is remarkable is that when most viewers examined such a scene in a stereoscope, it was reported that such a surface was indeed perceived (Nakayama and Shimojo 1990). A faint but distinct triangular-shaped surface bounded by illusory contours (posi-

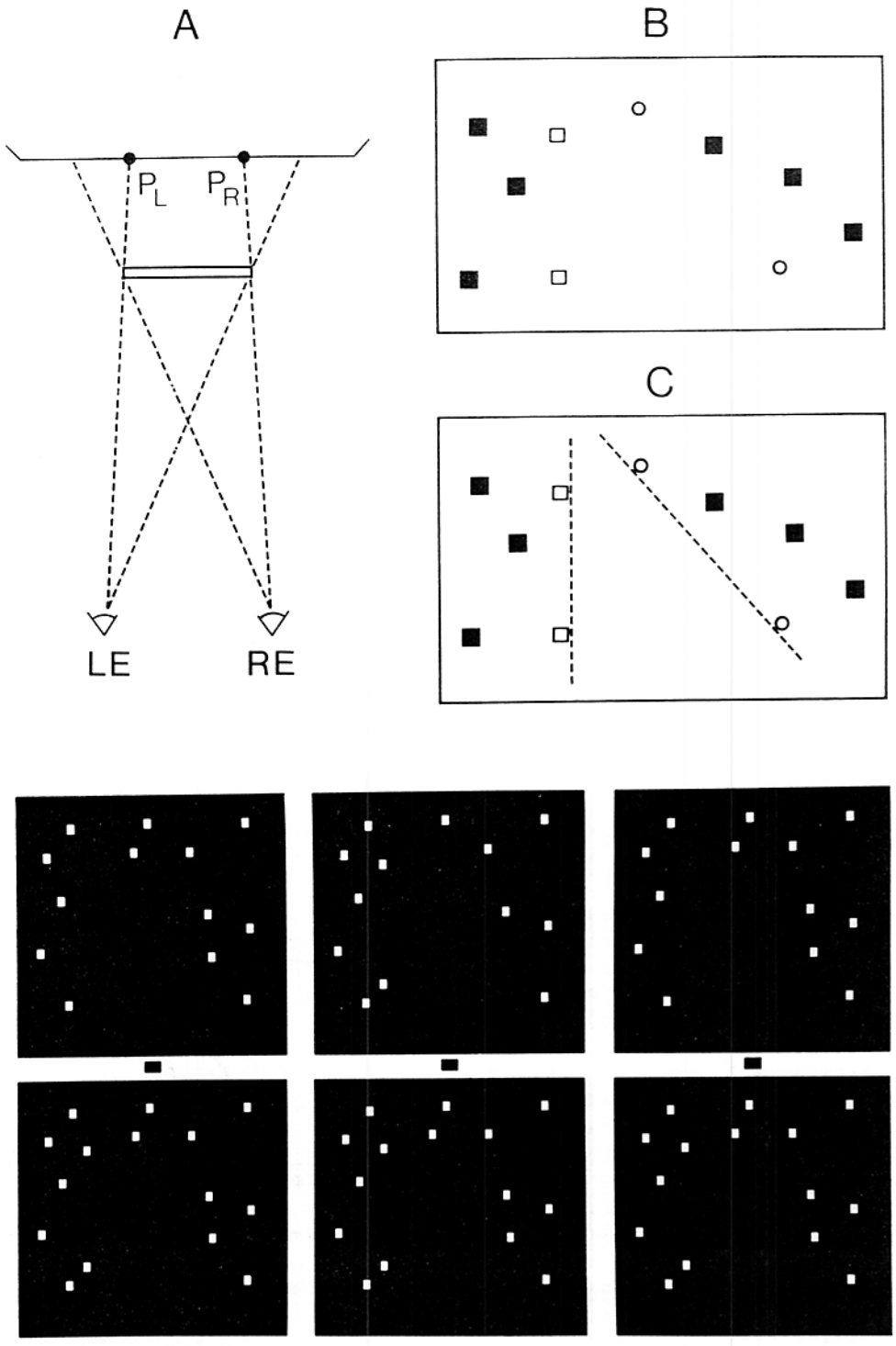


Figure 11. (*A*) Occlusion is different for the two viewing eyes. Because of the existence of an opaque surface, point P_L in the background is visible only to the left eye, and point P_R is visible only to the right eye. (*B*) Schematic of a stereogram where all filled circles represent points that are viewed binocularly and have zero disparity. Open squares and open circles identify points visible to the left eye only and right eye only, respectively. (*C*) Same as in *B*, except showing the boundaries of a surface that could have given rise to the monocular points. It also identifies where observers see subjective contours in the stereograms below. In the stereogram depicted in the next portion of the figure, the top portion has four unpaired points as illustrated in *B*. In the bottom portion, all points are paired.

tioned as the dashed lines in Fig. 11C) was seen to lie in front of the background dots. Thus, an illusory occluding contour was seen to the immediate right of the left-eye-only dots and to the immediate left of right-eye-only dots. Similar to the case of other surfaces bounded by illusory contours (see Petry and Meyer 1987), the surface was also seen as slightly darker than the dark background. The case just described is shown in the upper stereogram presented in Figure 11. A control case where all points are interocularly matched is seen in the lower stereogram. When presented in a stereoscope, it should be clear for most viewers (with normal stereoscopic vision) that an approximately triangular surface can be seen in the upper but not in the lower stereogram.

We think that this latest demonstration again makes the point. Perception is remarkably adaptive and intelligent. What makes this latest demonstration of added interest, however, is its very specific reliance on eye-of-origin information. In this demonstration as in others (see Shimojo et al. 1988; Nakayama and Shimojo 1990; Shimojo and Nakayama 1990), it is of importance for the visual system to know which eye was presented with the unpaired stimulus. The position of the illusory occluding contour had a very specific position in relation to whether the monocular dot came through the right or left eye. If it was presented to the left eye, the illusory occluding contour appeared to its immediate right. If it was presented to the right eye, the contour appeared to its immediate left. This is not surprising, given the nature of the real-world geometry described in Figure 11A. Putting this observation together with our current understanding of the properties of cells in striate and extrastriate cortex, however, leads us to propose an unexpectedly strong hypothesis; the inferential process responsible for the perception of the surface must begin very early in the cortical visual pathway, perhaps as early as striate cortex and at least as early as its immediate cortical projection zones.

The argument runs as follows. Eye-of-origin information is preserved in the ocular dominance structure of V1. Here, many cells respond preferentially to right-eye or left-eye stimulation. Beyond V1, however, this explicit coding of eye preference is essentially gone. Instead of finding cells that respond more to one eye than to the other, extrastriate cortex is characteristic in having cells that are equal in responsiveness to stimulation of either eye (Burkhalter and Van Essen 1986; Maunsell and Van Essen 1983; Hubel and Livingstone 1987). Thus, it would seem that the neural mechanism responsible for the illusory contours and the associated surface must have direct access to the signals from these monocular striate cells. At the very latest, therefore, cortical areas with such inputs from striate cortex comprise the possible candidates.

Cortical Organization and Surface Representation

Having provided some support for the view that the encoding of occlusion and surface formation can begin fairly early in cortical visual processing, we are in a position to return to the original problem presented in the introduction, namely, the feature-binding problem. How do separate features in an image get associated into unitary perceptual wholes?

Our approach to this problem has been to concentrate on the perception of surfaces, because it is here that one obtains clear evidence for differential linkage and visibility of features depending on global scene encoding. Depending on the final interpretation of the image data, the encoding of seemingly primitive processes such as contour, depth, and color can be radically altered. The perception of surfaces is influenced by primitives, and moreover, the final outcome can alter these same primitives. Obviously, the strong notion of modularity is inconsistent with this view. By the same token, this might be taken as evidence against any model based on a sequence or hierarchy of information processing. Yet, information obtained over the past 30 years indicates that cortex is highly organized and hierarchical, at least for vision (van Essen 1985).

To reconcile our perceptual results with the clear sequence and hierarchy seen in cortex, we briefly outline a speculative framework to consider the cortical analysis of surfaces. Figure 12 shows two levels of cortical organization, a hypothetical area V_m projecting to V_n (where index $m < n$). Thus, V_n represents an as yet unspecified extrastriate area (V2, V3, V4, etc.). In any given cortical area of interest, anatomical and physiological data support the view of vertical columns, patchy regions where function is more or less the same within a given vertical extent and differs more along a horizontal direction. Thus, one has thick, thin, and interstripes of V2. For the case of V2, color, contour, depth, and motion, for example, all have slightly differ-

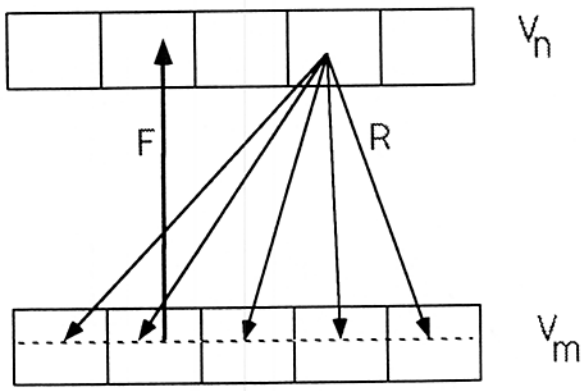


Figure 12. Hypothesized hierarchial relationship between two cortical visual areas, a lower level V_m projecting to a higher V_n showing its possible relation to visual surface formation. Primitive features segregated into different vertical columns of V_m project to V_n, where surface formation is hypothesized to occur. Forward connections (labeled F) between the regions are relatively local and tend to preserve segregation of pathways and topography. Reciprocal projections (labeled R), however, are not local but project widely and diffusely to a given lamina back in V_m (see text). Such backward connections enable different pathways to interact over wider retinotopic regions.

ent distributions with respect to these columnar subdivisions (Hubel and Livingstone 1987). As such, one might still consider these subregions as functional modules. In addition, these separate columns project forward to higher cortical regions in a reasonably local and discrete manner, roughly preserving segregation and hence modularity at each stage.

Of interest, however, is the fact that the backward projection (from V_n to V_m) appears to have a very different projection rule (Krubitzer and Kaas 1989; Zeki 1990). Instead of a given patch, say in V_n, projecting just to the patch from which it received its projection, the backward projection tends to be a laminar one (confined to particular horizontal layers), and it is much more diffuse, projecting across many columns in V_n. As such, a back projection has a much greater opportunity to operate across modules, as well as across larger retinotopic areas. This means that higher-order processing, say surface formation, could be influenced by and yet still influence the earlier processing of features.

Neurophysiological Implications

A number of hypotheses advanced in this paper may be directly testable in single unit experiments. We mention a few. First, regarding the possibility that eye-of-origin information participates in surface and depth perception, one could imagine a certain class of illusory contour cells (identified by von der Heydt et al. 1984; Peterhans and von der Heydt 1989; von der Heydt and Peterhans 1989) in which the responses of such cells would be stimulated by unpaired monocular regions, much in the way that such regions evoke illusory contours seen in Figure 11. Thus, one might predict that certain contour-specific cells would be more vigorously stimulated by unpaired rather than paired dots. In addition, one might also predict that the position of dots which would best stimulate such cells would correspond to the fact that contours appear to the right of left-eye-only and to the left of right-eye-only points.

Second, we think that our hypothesis regarding the close interdependence of contour, color, and depth might be explored by seeing whether color-specific cells might alter their firing rates in a predicted direction from a subtle manipulation of disparity. Thus, by simply reversing the disparity of figures like those seen in Figure 6, one might see changes in the response of color-specific cells whose receptive fields are in the "filled-in" region surrounded by physically colored regions. In addition, the removal of the outer limbs of the white cross from Figure 6 might be expected to stimulate illusory contour cells having horizontally oriented receptive fields, in accordance with the horizontal illusory contour seen in Figure 5.

CONCLUDING SUMMARY

In the course of attempting to understand the visual encoding of surfaces, we have made a number of points. First, we introduced the concept of contour belongingness or ownership and outlined its relationship to the linkage of image fragments. We demonstrated that if an image region is bounded by a contour that it does not own, linkage to other image regions becomes a possibility. Furthermore, we hypothesized that cell types encountered in visual cortex could begin to make the distinction between contours that actually ended in the three-dimensional world from those that continued behind occluding surfaces.

Second, we showed that the local pooling of disparity signals is insufficient to understand the perception of some very simple stereoscopic scenes. One must view the visual system as constructing a consistent interpretation of the scene rather than pooling local disparity signals.

Third, we argued from our studies of transparency that the visual system effectively avoids the assumption of the accidental view in interpreting scenes. These studies also show a strong two-way relationship between underlying features and surface formation. Thus, the triggering of transparency can lead to the destruction and creation of illusory contours, the spreading of color, and a radically different perception of depth.

Fourth, we showed that binocularly unpaired points can lead to the perception of a surface bounded by illusory contours. Because such eye-of-origin information is available only at the earliest stages of cortical processing, we suggested that the inferential process for surface formation must begin very early, as early as area V1 or its immediate cortical projection zones.

Fifth, we hypothesized a relationship between feature and surface representation that is derived from our perceptual observations and the known anatomy of extrastriate cortical areas.

ACKNOWLEDGMENTS

The current research is partially supported by Air Force Office of Scientific Research grant 83-0320. S.S. was supported by a fellowship from the Japanese Society for the Promotion of Science for Japanese Junior Scientists and by a Rachel C. Atkinson Fellowship.

REFERENCES

Adelson, E.H. and J.R. Bergen. 1985. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am.* **2:** 284.

Barlow, H.B., C. Blakemore, and J.D. Pettigrew. 1967. The neural mechanism of binocular depth discrimination. *J. Physiol.* **193:** 327.

Bregman, A.L. 1981. Asking the "what for" question in auditory perception. In *Perceptual organization* (ed. M. Kubovy and J.R. Pomerantz), p. 99. Lawrence Earlbaum, Hillsdale, New Jersey.

Burkhalter, A. and D.C. van Essen. 1986. Processing of color, form and disparity information in visual areas VP and V2 of ventral extrastriate cortex in the macaque monkey. *J. Neurosci.* **6:** 2327.

Crick, F. 1984. Function of the thalamic reticular complex: The searchlight hypothesis. *Proc. Natl. Acad. Sci.* **81:** 4586.

Gibson, J.J. 1950. *The perception of the visual world*. Houghton Mifflin, Boston.
Gregory, R.L. 1970. *The intelligent eye*. McGraw-Hill, New York.
Guzman, A. 1984. Decomposition of a visual scene into three dimensional borders: Fall Joint Conference 33. In *Information technology series*, volume VI. *Artificial intelligence* (ed. O. Fischein), p. 310. Reston, Virginia.
Helmholtz, H. 1910. *Treatise on physiological optics*. (Translated from the third German edition; reprinted in 1962) (ed. J.P.C. Southall), vol. 3. Dover, New York.
Hildreth, E. 1984. *The measurement of visual motion*. ACM Distinguished Dissertation Series. MIT Press, Cambridge.
Hochberg, J. 1981. Levels of perceptual organization. In *Perceptual organization* (ed. M. Kubovy and J.R. Pomerantz), p. 255. Lawrence Erlbaum, Hillsdale, New Jersey.
Hubel, D.H. and M.S. Livingstone. 1987. Segregation of form, color, and stereopsis in primate area 18. *J. Neurosci.* **7:** 3378.
Hubel, D.H. and T.N. Wiesel. 1965. Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat. *J. Neurophysiol.* **28:** 229.
Julesz, B. 1971. *Foundations of cyclopean perception*. University of Chicago Press, Illinois.
Kanizsa, G. 1979. *Organization in vision*. Praeger Publishers, New York.
Koenderink, J.J. 1990. *Solid shape*. MIT Press, Cambridge.
Koffka, K. 1935. *Principles of Gestalt psychology*, p. 107. Harcourt, Brace, and World, Cleveland.
Krubitzer, L.A. and J.H. Kaas. 1989. Cortical integration of parallel pathways in the visual system of primates. *Brain Res.* **478:** 161.
Land, E.H. 1977. The retinex theory of color vision. *Sci. Am.* **237:** 108.
Livingstone, M.S. and D.H. Hubel. 1987. Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *J. Neurosci.* **7:** 3416.
Marr, D. 1982. *Vision*. Freeman, San Francisco.
Marr, D. and T. Poggio. 1976. Cooperative computation of stereo disparity. *Science* **194:** 283.
Maunsell, J.H.R. and W.T. Newsome. 1987. Visual processing in monkey extrastriate cortex. *Annu. Rev. Neurosci.* **10:** 363.
Maunsell, J.H.R. and D.C. Van Essen. 1983. Functional properties of neurons in the middle temporal visual area (MT) of the macaque monkey. I. Selectivity for stimulus direction, speed and orientation. *J. Neurophysiol.* **49:** 1127.
Metelli, F. 1974. The perception of transparency. *Sci. Am.* **230:** 90.
Michotte, A. 1954. *La perception de la causalite*. Publications Universitaires de Louvain, France.
Nakayama, K. and S. Shimojo. 1990. daVinci Stereopsis: Depth and subjective occluding contours from unpaired image points. *Vision Res.* **30:** (in press).
Nakayama, K., S. Shimojo, and V.S. Ramachandran. 1990. Transparency: Relation to depth, subjective contours, luminance, and neon color spreading. *Perception* (in press).
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.
Nelson, J.I. 1975. Globality and stereoscopic fusion in binocular vision. *J. Theor. Biol.* **49:** 1.
Newsome, W.T. and E.B. Pare. 1988. A selective impairment of motion perception following lesions in the middle temporal visual area (MT). *J. Neurosci.* **8:** 2201.
Peterhans, E. and R. von der Heydt. 1989. Mechanisms of contour perception in monkey visual cortex. II. Contours bridging gaps. *J. Neurosci.* **9:** 1749.
Petry, S. and G.E. Meyer. 1987. *The perception of illusory contours*. Springer Verlag, New York.
Poggio, T., V. Torre, and C. Koch. 1985. Computational vision and regularization theory. *Nature* **317:** 314.
Redies, C. and L. Spillmann. 1981. The neon-color effect in the Ehrenstein illusion. *Perception* **10:** 667.
Richards, W., J.J. Koenderink, and D.D. Hoffman. 1987. Inferring three-dimensional shapes from two-dimensional silhouettes. *J. Opt. Soc. Am. A* **4:** 1168.
Rock, I. and R. Anson. 1979. Illusory contours as the solution to a problem. *Perception* **8:** 665.
Schiller, P.H., N.K. Logothetis, and E.R. Charles. 1990. Functions of the colour-opponent and broad-band channels of the visual system. *Nature* **343:** 68.
Schiller, P.H., J.H. Sandell, and J.H. Maunsell. 1986. Functions of the ON and OFF channels of the visual system. *Nature* **322:** 824.
Shimojo, S. and K. Nakayama. 1990. Real world occlusion constraints and binocular rivalry interaction. *Vision Res.* **30:** 69.
Shimojo, S., G.H. Silverman, and K. Nakayama. 1988. An occlusion-related mechanism of depth perception based on motion and interocular sequence. *Nature* **333:** 265.
———. 1989. Occlusion and the solution to the aperature problem for motion. *Vision Res.* **29:** 619.
Sperling, G. 1970. Binocular vision: A physical and neural theory. *Am. J. Psychol.* **83:** 461.
Treisman, A. and G. Gelade. 1980. A feature integration theory of attention. *Cognit. Psychol.* **12:** 97.
Van Essen, D.C. 1985. Functional organization of primate visual cortex. *Cereb. Cortex* **3:** 259.
Varin, D. 1971. Fenomini di contrasto e diffusione chromatica nell' organizzazione spaziale del campo percettivo. *Riv. Psicol.* **65:** 101.
von der Heydt, R. and E. Peterhans. 1989. Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *J. Neurosci.* **9:** 1731.
von der Heydt, R., E. Peterhans, and G. Baumgartner. 1984. Illusory contours and cortical neuron responses. *Science* **224:** 1260.
Wheatstone, C. 1838. On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philos. Trans. R. Soc. Lond. B* **128:** 371.
Zeki, S. 1978. Functional specialization in the visual cortex of the rhesus monkey. *Nature* **274:** 423.
———. 1990. The motion pathways of the visual cortex. In *Vision: Coding and efficiency* (ed. C. Blakemore), Cambridge University Press, United Kingdom. (In press.)