Experiencing and Perceiving Visual Surfaces

Ken Nakayama and Shinsuke Shimojo

A theoretical framework is proposed to understand binocular visual surface perception based on the idea of a mobile observer sampling images from random vantage points in space. Application of the generic sampling principle indicates that the visual system acts as if it were viewing surface layouts from generic not accidental vantage points. Through the observer's experience of optical sampling, which can be characterized geometrically, the visual system makes associative connections between images and surfaces, passively internalizing the conditional probabilities of image sampling from surfaces. This in turn enables the visual system to determine which surface a given image most strongly indicates. Thus, visual surface perception can be considered as inverse ecological optics based on learning through ecological optics. As such, it is formally equivalent to a degenerate form of Bayesian inference where prior probabilities are neglected.

When we see objects in the world, what we actually "see" is much more than the retinal image. Our perception is three-dimensional (3-D). Moreover, it reflects constant properties of the objects and the environment, regardless of changes in the retinal image with varying viewing conditions. How does the visual system make this possible?

Two different approaches have been evident in the study of visual perception. One approach, most successful in recent times, is based on the idea that perception emerges automatically by some combination of neuronal receptive fields. In the study of depth perception, this general line of thinking has been supported by psychophysical and physiological evidence. The "purely cyclopean" perception in the Julesz random dot stereogram (*1*) shows that depth can emerge without the mediation of any higher order form recognition. This suggested that relatively local disparity-specific processes could account for the perception of a floating figure in an otherwise camouflaged display. Corresponding electrophysiological experiments with single cell recordings demonstrated that the depth of such stimuli could be coded by neurons in the visual cortex, receiving input from the two eyes (*2*). In contrast to this more modern approach, there exists an older tradition, which asserts that perception is inferential, that it can cleverly determine the nature of the world with limited image data. Starting with Helmholtz's unconscious inference (*3*) and with more recent formulations such as Gregory's "perceptual hypotheses," this approach stresses the importance of problem solving in the process of seeing (*4*). So far, however, "perceptual inference" theories have not been successfully linked to physiological findings, and they are not easily distinguished from other theories of mental processes, including those that attempt to account for thinking and reasoning.

In this article, we argue strongly for the importance of inference but provide the beginnings of what we think is a low-level mechanistic explanation of how such inferences could be learned. We argue that the observer's experience of optical sampling during locomotion provides a key to understand what will be perceived later on.

Our domain is stereoscopic vision, commonly thought to be dictated by early, prewired, local mechanisms. Instead, we consider stereopsis to be an example of surface representation, not obviously linked to currently understood properties of visual neurons (*5*) or to higher stages of object recognition.

Julesz's random dot stereogram defined much of the subsequent work in the field of binocular stereopsis. Ever since, most visual scientists assumed, either explicitly or implicitly, that stereopsis depends most critically on the solution to the "matching" problem. This is indeed an important and difficult problem because the rich local texture of random dot stereograms requires that the visual system must find the correct binocular match of individual points in the face of numerous possible "false matches" (*1, 6, 7*).

Random dot stereograms, however, are not entirely representative of the local details of everyday scenes. Such dense textures occur only occasionally in natural images, and some scenes contain large regions that are effectively untextured. Whereas human perceptual systems correctly interpret such scenes, current models, which were originally designed to handle densely textured stereograms, do not (*8*). Thus, we think it important to examine how the visual system handles image regions where texture is largely absent.

Failure of Depth Interpolation in Untextured Stereograms

Note the stereogram in Fig. 1A and consider the binocular disparity information available (*9*). Because this cross has no interior

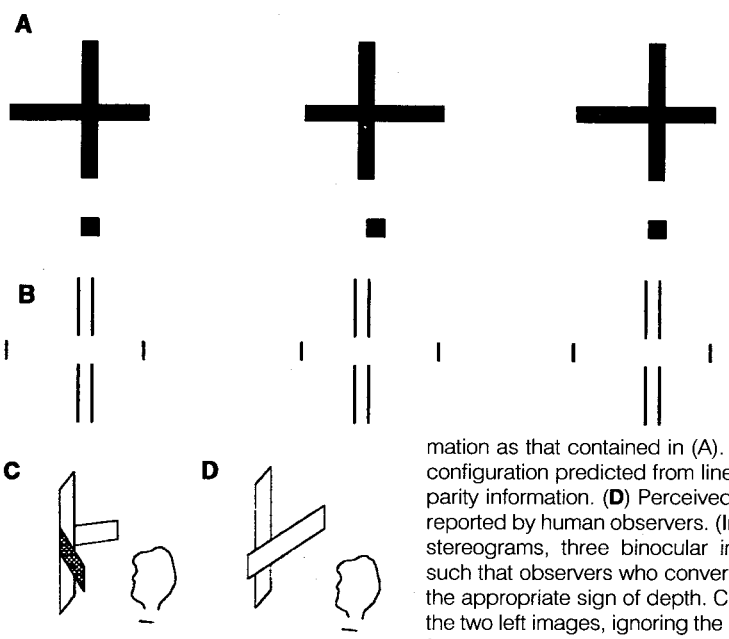


Fig. 1. Case 1: (**A**) Cross stereogram. Because the outer edges of the horizontal limb of the cross have crossed binocular disparity, these edges should be seen in front. Depths of the untextured interior regions of the cross are not specified by binocular disparity. (**B**) Reduced line stereogram having the same disparity information as that contained in (A). (**C**) Perceived surface configuration predicted from linear interpolation of disparity information. (**D**) Perceived surface configuration reported by human observers. (In these and in all other stereograms, three binocular images are presented such that observers who converge or diverge can see the appropriate sign of depth. Convergers should fuse the two left images, ignoring the right; divergers should fuse the two right images, ignoring the left).

K. Nakayama is in the Vision Sciences Laboratory, Department of Psychology, Harvard University, Cambridge, MA 02138. S. Shimojo is in the Department of Psychology, University of Tokyo, Komaba, Meguro-ku, Tokyo 153, Japan.

texture, only the bounding contours are available to convey binocular disparity information. Moreover, because binocular disparity is not available from the horizontally oriented contours of the figures, vertically oriented contours provide the only source of horizontal disparity information. We have emphasized this point by constructing a partial stereogram having exactly the same disparity information as the cross. It contains only vertical lines (Fig. 1B).

Given this paucity of local disparity information in the whole figure, one might ask how depth gets assigned to the interior portions of the figure where disparity is not explicitly defined. Classical stereopsis makes no specific prediction as to the depth of these untextured regions. Yet, the least arbitrary assumption would be that the perceived depth of given positions in an untextured region is a simple linear interpolation between points having a defined disparity (*10*).

The ends of the horizontal limb of the cross have crossed disparity. That should indicate that these contours are nearer to the observer than to the contours defining the vertical limb (*11*). Assuming depth interpolation for the stereogram in Fig. 1A, we might expect a simple continuity of depth from the center of the cross (seen in back) to the ends of the horizontal limbs (seen in front). The observer should see a vertical bar in back flanked by horizontal "wings" that are slanted toward the observer (Fig. 1C).

We have shown this stereogram (Fig. 1A) to several hundred observers, and only a tiny minority observe what we have just outlined. Instead, the most frequently seen configuration is that of a horizontal bar in front of a vertical bar (Fig. 1D). In keeping with the perception of a straight horizontal bar in front, observers also see a subjective occluding contour, which is not present in the image itself but which perceptually segregates and completes the bar in front (*9, 12*).

In our second stereoscopic demonstration (Fig. 2), we again show that the visual system violates the expectation of simple interpolation by allowing a break in the perceived surface pattern. This demonstration also illustrates a qualitatively different phenomenon, the perception of transparency, which is accompanied by color spreading into otherwise uncolored regions (*9, 13*). In this example (Fig. 2A), the viewer observes a set of four bipartite bars, divided into red and white regions against a black background. The ends of the bars are in the zero disparity plane, and the dividing line between the regions has crossed disparity. Simple interpolation theory would predict that this edge would be seen in front and that the other two edges would be seen in back, forming a folded surface (Fig. 2B).

What is seen, however, is qualitatively different. Instead of seeing a set of folded surfaces (Fig. 2B), each visible face of which recedes back from the viewer, one usually sees this configuration as two disconnected surfaces, one transparent in front, the other opaque in back, each of which does not recede but is frontoparallel to the observer (Fig. 2C). This transparent surface appears to "complete" in front and merge as a single surface, which is in front of all four bars. Furthermore, it is "contained" by subjective contours, which bound the color that spreads into the black region (Fig. 2C).

Image Ambiguity in Stereograms

Depth interpolation failed to account for what is seen. Instead, it appears as if the visual system reached a conclusion with only the scantiest evidence. Is what is perceived consistent with the binocular image data?

In case 1 (Fig. 1), the perception of surface breakage, although not predicted from depth interpolation, is nonetheless consistent with the disparities presented to the observer. A real-world bar configuration as in Fig. 1D, as well as that in Fig. 1C, could have given rise to the disparities seen in the stereogram. The same is true for case 2. A transparent red surface lying in front of white bars (Fig. 2C), as well as folded surfaces (Fig. 2B), could have given rise to the disparities seen in Fig. 2A. Thus for both stereograms presented, the 3-D interpretation is ambiguous. The observer is presented with image data that can be interpreted in more than one way.

The issues raised here are not entirely new. Both traditional and modern students of visual perception have noted the ambiguity of the visual image. For example, it has been pointed out that, when one is presented with an image in the form of a triangle, there is an infinite number of triangles in space that could have given rise to this image (*14*). More recently, in studies of computational vision, it has been noted that vision is ill-posed in that the

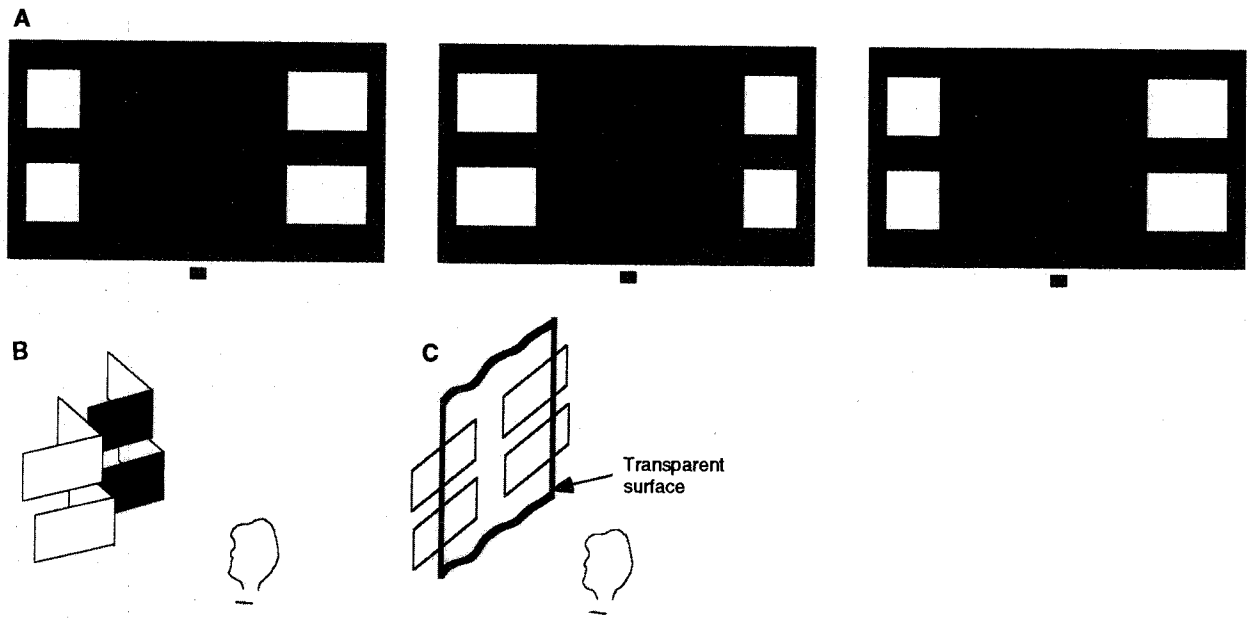


Fig. 2. Case 2: (**A**) Bipartite stereogram where the center line in each of the four bars has crossed disparity and should be perceived in front of the ends of the individual bars. (**B**) Perceived surface in depth predicted by linear interpolation of disparity information: four folded sheets. (**C**) Perceived surface configuration reported by human observers: a single frontoparallel transparent surface in front of four bars in back. If this stereogram, as well as that shown in Fig. 6A, is viewed in the reverse configuration (with right and left eye views exchanged), the red region will look opaque.

information available in the image by itself is insufficient to recover the structure in the real world (*15*).

Therefore, simply checking to see whether the image array is consistent with a given perceptual interpretation is not sufficient to permit one to decide which interpretation is true. Nor can one apply a simple rule such as depth interpolation among disparities to reach the perceived solution.

Ecological Optics and the Importance of Viewing Position

Rather than starting with an image and thinking about automatically reconstructing the surface by interpolation, we advocate a conceptual shift, arguing that the problem can be best understood from the perspective of ecological optics. As discussed by Gibson (*16*), we must remember that, when one is viewing surfaces in the world, the vantage point of the viewer is rarely stationary. The observer locomotes, and new sets of image samples continuously arise. Thus, there is of necessity a one-to-many mapping from the physical layout of surfaces to the image. For example, in Fig. 3A we indicate three different sets of surfaces, S_1, S_2, and S_3, each of which potentially gives rise to more than one image as a consequence of the viewer taking differing vantage points.

If this mapping were literally as described above, the task of the visual system would be relatively easy: given a visual image, say, I_1, it would simply designate the corresponding surface S_1. What makes this task more difficult is the fact that one image arising from one surface can also arise from a variety of other surfaces (Fig. 3B). So the mapping is not only one-to-many, it is also many-to-one.

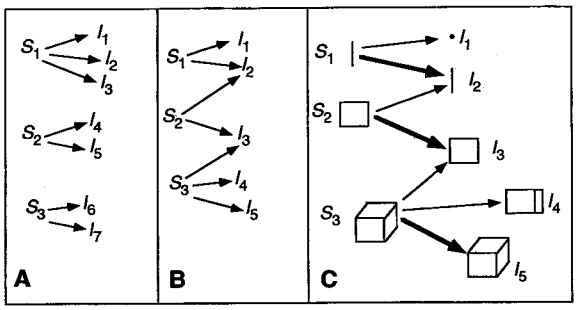


Fig. 3. Mapping of surfaces to images as defined by ecological optics. Classes of surface structure in the real world (S_1, S_2, and S_3) can give rise to sampled images (I_1, I_2, . . .) as an observer assumes differing vantage points. (**A**) Image sampling without stimulus ambiguity where each image class is tied to only one surface configuration. (**B**) Image sampling with stimulus ambiguity where some image classes (I_2 and I_3) can arise from more than one surface. (**C**) Specific example of line, square, and cube, where thick and thin arrows indicate generic and accidental samplings, respectively.

To illustrate, consider a set of various real-world objects: a line, a square, and a cube, labeled S_1, S_2, and S_3, respectively, in Fig. 3C. Depending on the location of the observer's vantage point, a line can give rise to an image of a line or a point; similarly a square can give rise to an image of a quadrilateral or a line; and finally a cube can give rise to images of polygonal figures having either one, two, or three faces. Thus, changes in the visual image occasioned by differing viewer positions can be summarized in terms of particular topological classes of images as initially proposed by Koenderink and van Doorn (*17*).

Now, think of the various viewing positions that could have given rise to each of these images. If the observer were to assume a random position in space around the cube, image I_5 (three faces) would be much more probable than image I_4 (two faces), which in turn would be much more probable than image I_3 (one face). Furthermore, as viewer distance is increased, this inequality would be accentuated, with the likelihood of three faces tending toward unity in the limit and the likelihoods of two faces and one face tending toward zero. Thus those viewing positions in space where I_5 (three faces) is encountered are called "generic" vantage points, with I_5 designated as a generic image (*18*). I_4 and I_3 are correspondingly called "accidental" images, and the viewing positions in space where they are encountered are called accidental vantage points. When one is confronted with image I_2, it would make more sense if one sees a line rather than a square. Similarly, when confronted with image I_3, one would see a square instead of a cube. It is only when confronted with I_5 that one would see a cube.

With these ideas in mind, we return to our untextured stereograms, starting with the cross in Fig. 1. The surface arrangement that would result from linear depth interpolation is the vertical bar flanked by the horizontal wings. Let us apply ecological optics to understand how these real-world surfaces might give rise to images sampled. If an observer were to view the configuration with the horizontal wings, he could encounter the binocular image in question (I_2) but only from a restricted set of vantage points (Fig. 4A). The observer is required to be at the same vertical level of the surfaces, neither above nor below, otherwise the horizontal wings would no longer appear collinear (*19*). As a telling comparison, consider the horizontal bar in front of the vertical bar (Fig. 4B). Here all images (I_4, I_5, and I_6) arising from this pair of surfaces are qualitatively the same as I_2. Instead of arising from just one particular

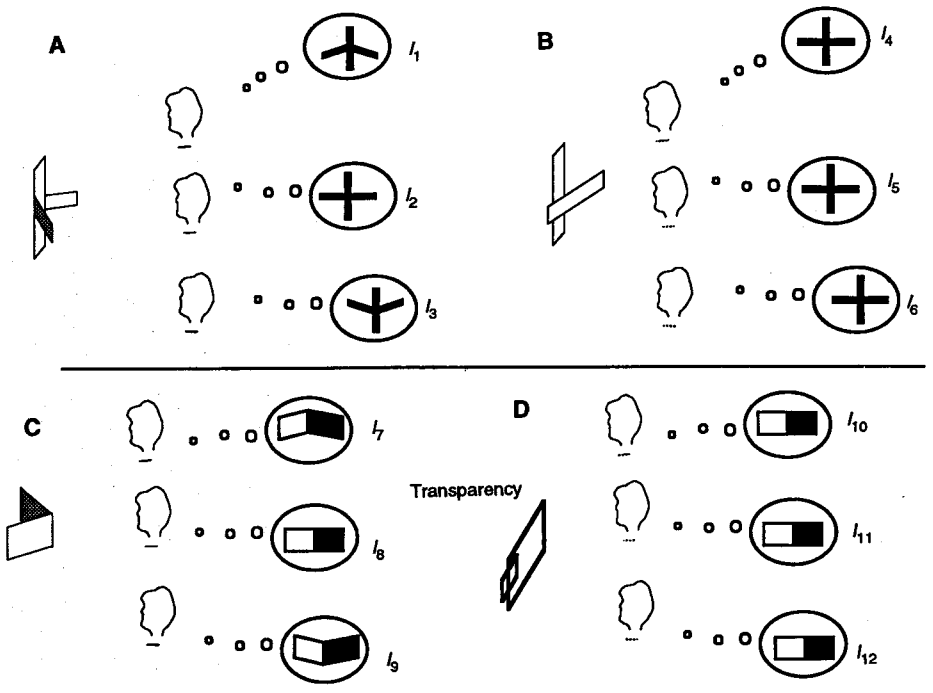


Fig. 4. Changes of image as the location of the observer's vantage point changes. Real-world surface structure is illustrated on the left, while changes of views with differing vertical elevations are shown on the right of each figure. (**A**) For the cross with wings bent. (**B**) For the horizontal bar in front of the vertical bar without a bend. (**C**) For the prism-like folded surfaces. (**D**) For a transparent surface in front.

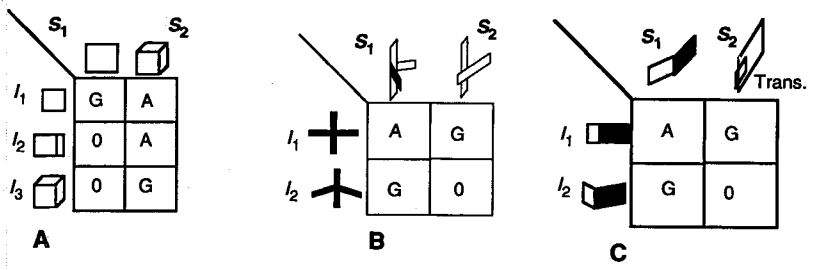


Fig. 5. Image sampling matrices for (**A**) cube versus square, (**B**) cross versus bent wings, and (**C**) fold versus transparency (Trans.). Potential surface sets in the real world are denoted at column headings (S_1 and S_2), and possible images are denoted by row headings (I_1, I_2, . . .). The likelihood that a given surface will give rise to a given image is categorized as either generic and very likely (G), accidental or rare (A), or impossible (0).

viewing position, the image in question can arise from a range of elevations. The image sampling process is different in each case, even though the same image can arise from two different surface layouts. The sampling is accidental for the situation depicted in Fig. 4A (bar with wings), whereas it is generic for the case shown in Fig. 4B (crossed bars).

The same analysis applies to the stereograms shown in Fig. 2. The surface arrangement that is predicted from linear depth interpolation is depicted as Fig. 4C (folded surface). The image I_8 could arise from a folded surface but only from a restricted set of vantage points. The observer is required to be just at the same vertical level of the surfaces, neither above nor below, otherwise the horizontal boundaries would no longer remain collinear. In contrast, in the case of the transparent surface in front of a white bar on a black background (Fig. 4D), all images sampled (I_{10}, I_{11}, and I_{12}) are qualitatively the same. Thus in this case, changes in viewing position have little effect, and a sampled image, which is categorically identical to I_8, can arise frequently. Again, there is a large difference in image sampling between the two surface layouts: In the first case the image I_8 is accidental; in the second the same image is generic.

The Principle of Generic Image Sampling

One of the major themes of this paper is that the observer's experience of optical sampling during locomotion provides a key to understanding what will be perceived later on. Here we develop this idea by applying the principle of generic image sampling. The principle can explain most of our findings and others to be described: *When faced with more than one surface interpretation of an image, the visual system assumes it is viewing the scene from a generic, not an accidental, vantage point.*

This principle is not entirely new. It has been one of the core assumptions of machine vision algorithms (*20*) as well as a theory of human object recognition (*21*). However, it can also be used to explain psychophysical phenomena that are generally thought to be part of early visual processing, namely, the encoding of depth in simple stereograms.

To develop this line of thinking in a more general framework, we summarize the relations between surfaces and images in the form of contingency matrices, which indicate the likelihood of obtaining images, given certain real-world surfaces. To illustrate, consider the two by three array in Fig. 5A. On top are possible sets of surfaces in the real world: a square and a cube. Along the side are the image classes that one might encounter (containing one, two, and three faces, respectively). In this array, we label the likelihood of images as either likely (G, generic), unlikely (A, accidental), or impossible (0).

So to enumerate the possible images that can be sampled from a square, we examine the cells in the first column corresponding to images I_1, I_2, and I_3. Only I_1 could have arisen from a generic vantage point. As such, it is labeled as G. The other two images (I_2 and I_3) could not have arisen, and these cells are thus labeled 0. The second column outlines the possible images that could arise from a cube. Images I_1 and I_2 can arise only from privileged viewpoints and are thus accidental and labeled A. I_3 can arise from many viewpoints and is thus generic and labeled G accordingly. Applying the principle of generic image sampling is now straightforward. When I_1 (the single face) is presented, it is clearly a generic view of S_1 (the square) and, as such, a square is perceived. S_2 (the cube) is not perceived because I_1 is an accidental view of it, not a generic one.

Having introduced these contingency matrices with familiar objects, we can now provide a framework to understand the two stereograms presented. At least two real-world surface configurations could give rise to the cross (Fig. 5B, I_1), either the bent wings or the horizontal bar in front of the vertical bar. When presented with the binocular image of the cross (I_1), the observer sees the horizontal bar S_2 and not the horizontal wings S_1 because I_1 is the generic view of the bar and not of the wings. It is only an accidental view of S_1. For the case of the folded surface versus transparency, the exposition is similar and is depicted in Fig. 5C. Thus, the visual system deals with stimulus ambiguity by picking the interpretation based on the generic sampling assumption.

To provide an even stronger case for the principle of generic image sampling, we add a third demonstration where the identical crossed bars of Fig. 1 can give rise to an entirely different global configuration, that of a transparent disk. All that is required is that the same cross be embedded in a new context. Consider the stereogram in Fig. 6. The inner red portion of this figure is geometrically identical to the stereogram shown in Fig. 1. Thus, the ends of the horizontal limbs of the cross have crossed disparity with respect to the vertical limbs and should be seen as closer. However, this red cross is now embedded in a larger white cross that has zero disparity.

What should we expect to see? We already know from Fig. 1 that the familiar bent bar configuration (Fig. 6B, S_1) is not seen. Instead, one sees the cross as a horizontal bar in front of a vertical one. As such, we might expect to see this same configuration embedded in the middle of a white outer cross (Fig. 6B, S_2).

The actual perception of this stereogram is totally different from either of these expectations (*22*). What is seen is a transparent red disk hovering in front of a white cross (Fig. 6B, S_3). So why do we see a transparent disk when earlier we just saw a horizontal bar in front of a vertical one in an essentially identical configuration?

Again, we appeal to the principle of generic image sampling, arguing that the visual system's preference is based on the likelihoods of images arising from different sets of surfaces (Fig. 6B). For the cases of I_2 and I_3, the task for the visual system is easy because only one surface interpretation is possible for each (see horizontal arrows). For the case of I_1 (Fig. 6B), many surface interpretations are possible, and here the principle of generic image sampling reveals its predictive power. I_1 is only an accidentally sampled image of S_1 and S_2, whereas it is a generically sampled image of S_3. Therefore, the transparent disk (S_3) has priority over S_1 and S_2 as a surface interpretation of I_1.

Although this third case provides powerful support for the generic sampling idea, it is even more revealing if we focus more closely on specific local aspects of the configurations, searching for local primitives upon which surface perception may depend. Transparency is seen in cases 2 and 3.

Can we identify a local feature common to both yet absent from the other case? Indeed, each has a specific type of stereoscopic T-junction. These are shown inside of the circles of Fig. 7, A and B, which reproduce relevant portions of case 2 and case 3. The stem of the T has a crossed disparity in relation to neighboring contours, and the transparent side of the stem is darker.

Monocularly viewed T-junctions have long been considered as evidence for occlusion, with the top of the T interpreted as occluding the stem (*3, 23*). No explicit formulation, however, has been made for the stereoscopic T-junctions, particularly where the stem is in front. In contrast to the monocular version, such a junction is incompatible with occlusion because the top of the T cannot act as an occluding contour if the stem is in front of it. Yet, the principle of generic image sampling can apply to this local configuration in the same way as we have applied it to whole figures. For example, consider two sets of stereoscopic junctions (Fig. 7, C and D), where one of the lines, L_3, has crossed disparity and is thus coded as in front. In Fig. 7D, two of the lines forming the intersection (L_1 and L_2) are collinear, whereas in Fig. 7C they are not. With ecological optics we can determine how likely it is that each image junction could arise, given an overlying transparent or a folded opaque surface. The T-junction in Fig. 7D could have arisen frequently from a transparent surface in front but not from a folded surface. As such, it is a generic image junction for a transparency and an accidental image junction for an opaque fold. On the other hand, the junction in Fig. 7C is a generic image for a fold. This analysis explains why the seemingly small step of embedding the smaller cross in a larger one leads to a dramatic change in surface perception. By adding an outer limb, which is collinear to the inner cross, we introduced a powerful local feature, a stereoscopic T-junction, which was essentially incompatible with opacity.

Role of Visual Experience

When the observer locomotes in the world, new images arise (*16*) with each class of image corresponding to a certain range of vantage points that the observer can assume in space (*17*). Thus, given a particular surface layout S_n, the probability that a class of image I_m will arise can be plausibly estimated from geometry, from considering the spatial range of vantage points under which a given image class is sampled, divided by the totality of possible vantage points. This is roughly equivalent to the quotient of two solid angles: the numerator being the solid angle under which a given image class is sampled, the denominator defining the solid angle from which the surface can be viewed. Thus, the viewer's experience can be expressed as a conditional probability $P(I_m|S_n)$, the probability of a given image I_m, given a real-world surface layout S_n (Fig. 8).

We may formalize our analysis in a more general contingency table (Fig. 8). For each column, identified by a surface S_n, the likelihoods or conditional probabilities of all possible image samplings ($I_1 \ldots I_m$) are listed, thus forming the exhaustive sample space of possible images. This analysis summarizes the totality of image sampling for the mobile observer. We argue that $p(I_m|S_n)$ is represented in the nervous system as an associative strength, between a surface and a visual image.

From this associative strength to perception is a short step, for the task facing the visual system is inverse ecological optics. Given an image I_m, it must decide which surface structure is the best candidate for what actually exists in the external world. Then, given an image I_m, the probability of S_n being the cause of I_m can be determined from a comparison of these learned conditional probabilities, $P(I_m|S_1)$, $P(I_m|S_2)$, $P(I_m|S_3)$, or associative strengths. In terms of our matrix in Fig. 8, this passive process acts as if it selects the cell having the highest probability in the same row, thereby finding the appropriate surface.

If we think of perceptual learning, the conditional probability terms assume special importance as they provide an opportunity for us to estimate visual experience simply from geometry. We suggest that these image sampling probabilities could be learned passively as the moving organism assumes essentially random positions with respect to real surface configurations. It follows, therefore, that the principle of generic image sampling emerges from this associative process.

The critical cues, such as collinearity, binocular disparity, and luminance contrast, are all local and primitive visual properties, the kind of selectivities commonly observed at early stages of the cortical visual processing. This implies that inverse ecological optics could be implemented in a strictly bottom-up, retinotopic representation, not requiring "higher order" inference. This view goes against the classical notion in psychology that perception is a form of problem solving or a hypothesis-testing process under strong top-down cognitive influence (*4*). Our views are perhaps closer to Helmholtz's unconscious inference (*3*), albeit at a retinotopic stage of representation, because we suggest that inferencelike processes can be constructed through associative learning in early vision without appealing to a homoculus or detectivelike processing.

Image Sampling Probability Versus Prior Probability

Inverse ecological optics based on image sampling probabilities are similar in part to Bayesian reasoning, as formulated in the Bayes theorem

$$P(S_n|I_m) = P(I_m|S_n)P(S_n)/ [P(I_m|S_1)P(S_1) + P(I_m|S_2)P(S_2) + \ldots P(I_m|S_n)P(S_n)] \quad (1)$$

where image sampling probabilities are denoted as conditional probabilities $P(I_m|S_n)$ and the posterior probability of surface S_n is denoted as $P(S_n|I_m)$. The major difference between the Bayes theorem and our formulation, Eq. 2 (Fig. 8, matrix), is our omission of prior probabilities or base rates $p(S_n)$

$$p(S_n|I_m) = p(I_m|S_1)/ [p(I_m|S_1) + p(I_m|S_2) + \ldots p(I_m|S_n)] \quad (2)$$

Conceivably, this term could be added to our framework, rendering it consistent with a full Bayesian approach. In terms of our matrix, each column could be multiplied by a factor proportional to the base rate of each surface structure to obtain the posterior surface probabilities described in Eq. 1. In terms of a neural network represented by such a matrix, base rates could be incorporated by increasing excitation along the corresponding vertical columns in proportion to prior probability. Yet, for many reasons, we do not think that this strategy is appropriate for the visual perception of surfaces.

First, in contrast to the learning of image sampling probabilities, the estimation of prior probability is problematic. Let us consider this in the context of transparent surfaces. Because transparent surfaces are generally infrequent, one would need a large period of time over which to obtain a sufficiently large set of samplings for a reliable measure of their prior probability. In addition, the frequency of transparent surfaces can change over varying environmental contexts, suggesting that a single general estimate of prior probability is likely to be meaningless. This necessitates that the estimate be obtained in varying environmental contexts and that these contexts be tagged for future use. Moreover, the fact that prior probabilities for transparency are likely to be small, combined with the fact that the numerator in Eq. 1 is a product, makes Bayesian decisions extremely unstable, particularly when estimates of prior probabilities are not immune to noise.

Even if such an estimate were made and could be assumed to be reliable and valid, it is not clear that such knowledge actually biases our perception appropriately. Many studies have shown that perception is largely impervious to prior knowledge, that

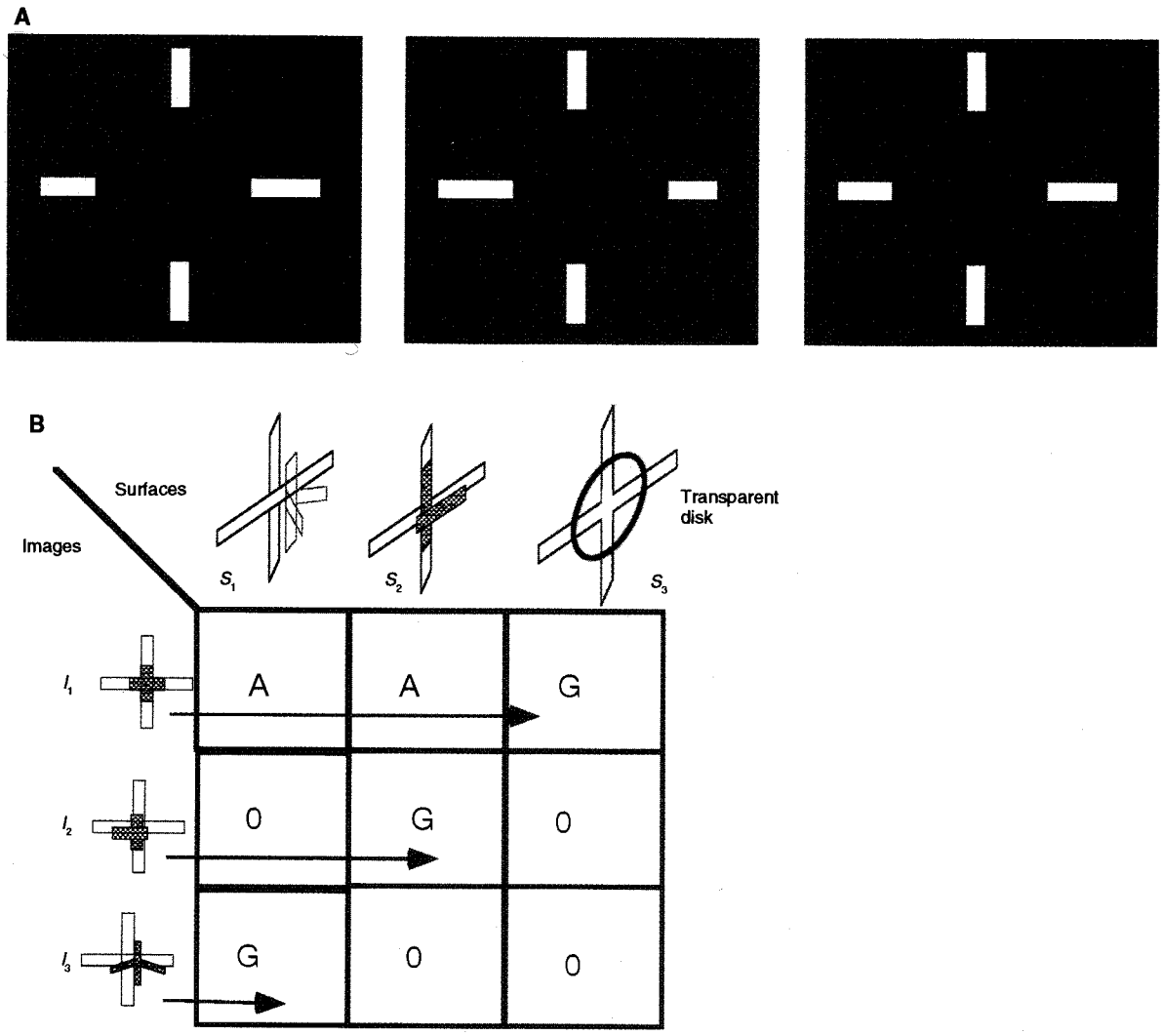


Fig. 6. Case 3: (**A**) Stereogram where the inner red cross is geometrically identical to that shown in Fig. 1A and is embedded in a larger white cross having zero disparity. (**B**) Image sampling matrix for this case. As in Fig. 5, surface classes are illustrated in columns, and sampled image classes are listed in rows. S_1, S_2, and S_3 refer to three possible surface configurations that could have given rise to this binocular image. I_1, I_2, and I_3 refer to images that could be possibly sampled.

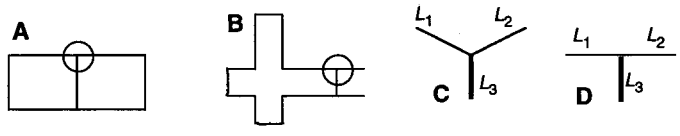


Fig. 7. (**A** and **B**) Local stereoscopic T-junctions embedded in the previous stereograms (Figs. 2 and 6, respectively). (**C**) Generic stereoscopic image sample from a folded surface. (**D**) Generic stereoscopic image sample from surface structure with transparency. Thick line, L_3, has crossed disparity and is coded as in front of L_1 and L_2.

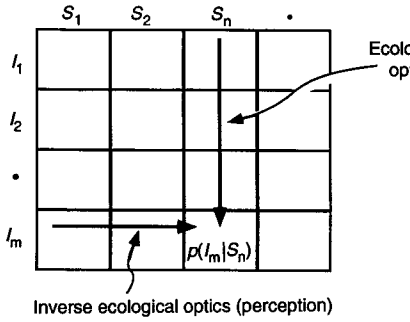


Fig. 8. Generalized form of the image sampli matrix.

seemingly compelling counterevidence at the cognitive level does not destroy strong perceptual illusions and other perceptual phenomena. Kanisza, for example, has shown many cases in which local perceptual rules essentially dominate our cognitive understanding of a scene (*24*). So in broad agreement with others (*16*, *25*), we suggest that the visual perception of surfaces is an autonomous process, minimally subject to object-specific knowledge about the world. Thus, our proposal is similar to a degenerate form of Bayesian inference where prior probabilities are unknown, set to equality, and ignored. This directly corresponds to inverse ecological optics as we have outlined it in the generalized matrix in Fig. 8 (*26*, *27*).

Need for Perceptual Categorization

For simplification, an important step has been missing in our discussion so far. We have suggested that sampled images can be associated with surfaces, not mentioning the representation of surfaces themselve By designating S_n as a real-world surface, v have glossed over the fact that it too mu have a neural representation. In particula we need to ask what kind of neural organ zation emerges for the perception of surfac so that they are seen as either connecte

folded, disconnected, transparent, or opaque, among others. For example, what gives transparency its characteristic appearance, identifiable even from opaque patches in a stereogram (Figs. 2 and 6)?

Because we have no specific data to address this issue directly, we can only speculate. Consider the various distinctive properties of images that are related to surface transparency. This would include, say, stereoscopic T-junction, contrast relations that satisfy Metelli's rule (*28*), simultaneous depth coding from a front and a back plane (*22*), and semispecular reflection at the surface. Given the associative power of theoretical neural networks (*29*), we hypothesize that, if these properties occur simultaneously when the observer locomotes in front of a transparent surface, an associative linkage is formed across these features. Then later, when an image contains a subset of these co-occurring features, the visual system can recall the whole pattern of features. This is presumably why we see transparency in our stereograms even though no transparency exists in the literal sense. Most important for our present discussion, it provides a plausible cluster of neural connections to represent a surface that can then be associated with specific image classes sampled (*30, 31*).

REFERENCES AND NOTES

1. B. Julesz, *Bell Syst. Tech. J.* **39**, 1125 (1960).
2. H. B. Barlow, C. Blakemore, J. D. Pettigrew, *J. Physiol. (London)* **193**, 327 (1967); G. F. Poggio and B. Fischer, *J. Neurophysiol.* **40**, 1392 (1977).
3. H. Helmholtz, *Handbuch der Physiologischen Optik* (Verlag, Hamburg, 1910) [J. P. C. Southall, Ed. *Helmholtz's Treatise on Physiological Optics* (Dover, New York, 1962)].
4. J. Hochberg, in *Perceptual Organization*, M. Kubovy and J. R. Pomerantz, Eds. (Erlbaum, Hillsdale, NJ, 1981), p. 255; R. L. Gregory, *The Intelligent Eye* (McGraw-Hill, New York, 1970); I. Rock, *The Logic of Perception* (MIT Press, Cambridge, MA, 1983).
5. Related demonstrations indicating the importance of surface representation have been summarized by V. S. Ramachandran, [*Percept. Psychophys.* **39**, 3361 (1986)]; see also V. S. Ramachandran and P. Cavanagh [*Nature* **317**, 527 (1985)].
6. B. Julesz, *Foundations of Cyclopean Perception* (Univ. of Chicago Press, Chicago, 1971).
7. G. F. Poggio and T. Poggio, *Annu. Rev. Neurosci.* **7**, 379 (1984).
8. D. G. Jones and J. Malik, *J. Invest. Ophthalmol. Vision Sci.* **31** (suppl.) 529 (1990).
9. K. Nakayama and S. Shimojo, *Cold Spring Harbor Symp. Quant. Biol.* **40**, 911 (1990).
10. In sparsely textured random dot stereograms, perceived depth of the surface region between dots appears as smoothly interpolated between individual dots (*6*).
11. The technical terms "crossed" and "uncrossed" disparity refer to those disparities that would lead to perceived near and far distances, respectively.
12. S. Shimojo and K. Nakayama, *Perception* **19**, 285 (1990).
13. H. F. J. van Tuijl, *Acta Psychol.* **39**, 441 (1975); C. Redies and L. Spillmann, *Perception* **10**, 667 (1981).
14. W. H. Ittelson, *Visual Space Perception* (Springer, New York, 1960).
15. T. Poggio *et. al.*, *Nature* **317**, 314 (1985).
16. J. J. Gibson, *Perception of the Visual World* (Houghton Mifflin, Boston, 1950); *The Senses Considered as Perceptual Systems* (Houghton Mifflin, Boston, 1966).
17. J. J. Koenderink and A. J. van Doorn, *Biol. Cybern.* **24**, 51 (1976).
18. Although the terms "generic" and "accidental" have been borrowed from the mathematics literature, the terms' exact meanings here and in computer vision have strayed from the precise mathematical definition originally assumed. In our usage, we consider an image I_1 more generic than the other image I_2 when the volume or area in space in which the vantage point can randomly move without causing qualitative change in the image is larger (*17*).
19. The observer should also be relatively far from the display, such that it is effectively a plane parallel projection, otherwise the horizontal limbs will not be collinear because of perspective. Furthermore, the head must not be tilted.
20. J. Malik, *Int. J. Comput. Vision* **1**, 73 (1987).
21. I. Biederman, *Comput. Vision Graphics Image Process.* **32**, 29 (1985). See also W. Richards, J. J. Koenderink, D. Hoffman, *J. Opt. Soc. Am. A* **4**, 1168 (1987).
22. K. Nakayama, S. Shimojo, V. S. Ramachandran, *Perception* **19**, 497 (1990).
23. A. Guzman, in *Automatic Interpretation and Classification of Images* A. Grasselli, Ed. (Academic Press, New York, 1969), pp. 243–276; D. A. Huffman, in *Machine Intelligence*, B. Metzler and D. Michie, Eds. (Edinburgh Univ. Press, Edinburgh, Scotland, ed. 6, 1971), pp. 295–323.
24. G. Kanisza, *Organization in Vision: Essays on Gestalt Perception* (Praeger, New York, 1979).
25. D. Marr, *Vision* (Freeman, San Francisco, 1980).
26. The neglect of prior probabilities may be related to recent findings in animal conditioning [R. Rescorla and P. C. Holland, *Annu. Rev. Psychol.* **33**, 265 (1982)]. These studies show that it is not simple contiguity between events that forms associations, but rather that the conditioned stimulus (CS) and unconditioned stimulus (US) must be correlated such that the CS provides unique predictive information about the US. Thus, sheer frequency of pairing alone is not a sufficient precondition for learning. This suggests that, if the frequency of transparent surfaces were extremely rare, these properties of association would enable the generic view of a transparent surface to call forth the perception of transparency even though the number of pairings of an accidental view and an opaque surface was actually more frequent.
27. Our neglect of prior probabilities should also be contrasted to the very different approach taken by H. B. Barlow [*Vision Res.* **30**, 1561 (1990)], where prior probabilities are explicitly registered by changes in neural connection strengths.
28. F. Metelli, *Sci. Am.* **230**, 90 (April, 1974).
29. J. Anderson, *Math. Biosci.* **14**, 197 (1972); T. Kohonen, *IEEE* **C-21**, 353 (1972).
30. In a similar vein, we have previously proposed the importance of associative learning for the perception of subjective contours and surfaces arising from binocularly unpaired image points. The same principle of generic image sampling can apply to these varieties of surface phenomena [K. Nakayama and S. Shimojo, *Vision Res.* **30**, 1811 (1990)].

31. This research was supported in part by grant 83-0320 from the Air Force Office of Scientific Research. S.S. was supported by the Ministry of Education, Sciences, and Culture, Japan.

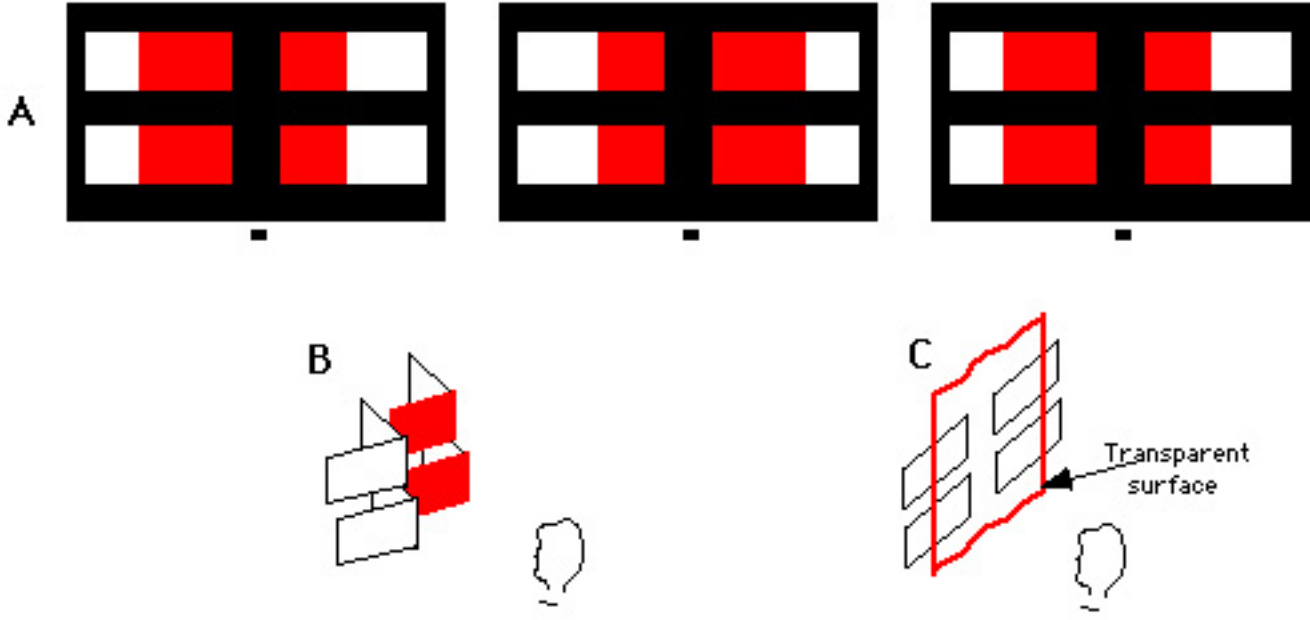
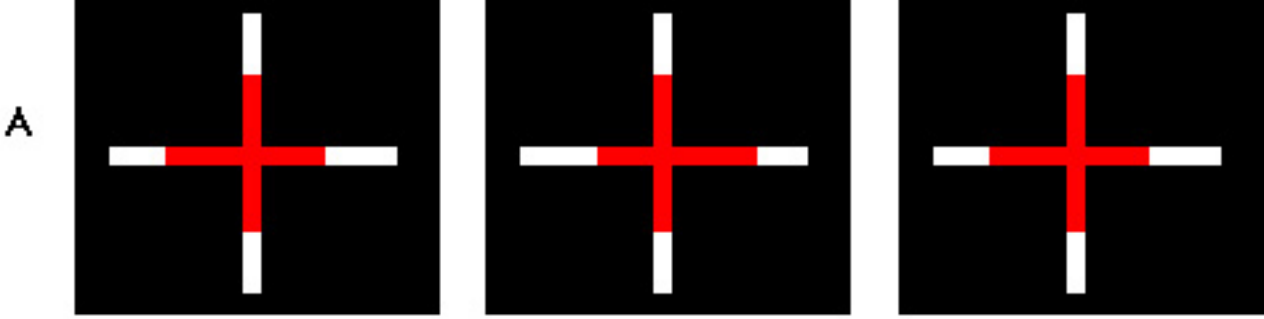
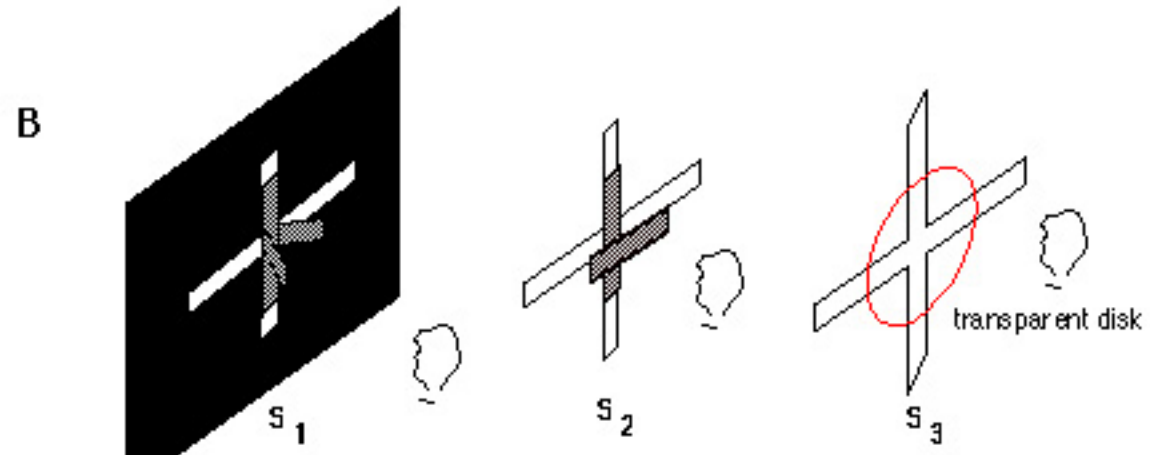


Figure 2


A


B
S 1
S 2
S 3
transparent disk

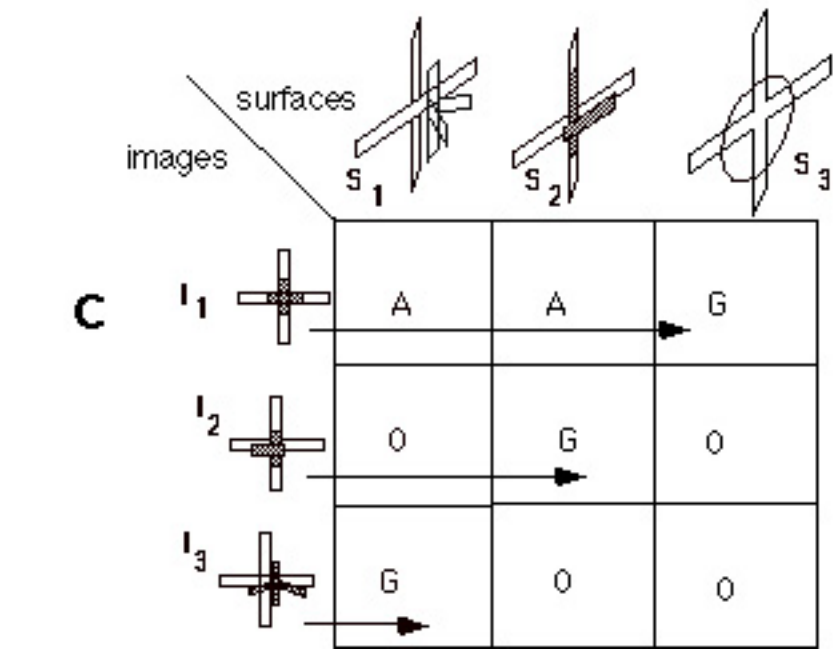

surfaces
images
S 1
S 2
S 3
C
I 1
I 2
I 3
A
A
G
0
G
0
G
0
0

Figure 6