that same location must be identified in the other image; and (3) the disparity between the two corresponding image points must be measured.

If one could identify a location beyond doubt in the two images, for example, by illuminating it with a spot of light, the first two steps could be avoided and the problem would be easy. In practice, we cannot go around carefully shining a spot of light on a surface and noting where its image falls in the two eyes, so we must somehow find a way of identifying a location by the more passive means of sensing the environment.

The reason why the task of identifying corresponding locations in the two images is difficult is because of what is called the false target problem. This occurs in what may be its extreme form in Julesz's random-dot stereograms (see Figure 1–1), and the nature of the problem is illustrated in Figure 3–5. The question is, Which dot corresponds to which? The left eye here sees four dots, and the right eye sees four, but which corresponds to which? A priori, all of the 16 possible matches are plausible candidates but when we observe such a stereo pair, we make the correspondences shown by the filled circles and not any of the correspondences shown by the open circles, which are called false targets.

Although this obviously makes some kind of sense, it is nevertheless surprising. How do we know which matches are correct and which should be ignored? What is more, there is another solution to this particular correspondence problem that seems just as valid. Look at the figure for a moment and try to see what it is. The other answer is the four central vertical matches, in which R_1 is paired with L_4, R_2 with L_3, R_3 with L_2, and R_4 with L_1. But we never see this match perceptually, which would appear as a set of squares in a receding line. Why not? Why only the other one, in which the squares line up, all about the same distance away?

From reading Chapter 2, the reader will immediately suggest using higher-level descriptions of the image—for example, matching first the two rows of dots as units and then, going on to match the individual squares and finally the edges of each square. And I think that something like this happens, but the first point to be clear about is that such a suggestion on its own is only a mechanism. The real question to ask is *Why* might something like that work? For the plain fact is that if we look just at the pair of images in Figure 3–5, there is no reason whatever why L_1 should not match R_3, L_2 match R_1, and even L_3 match R_1.

What we need is some additional information to help us decide which matchings are correct by constraining them in some way, and to do this we have to examine the basis in the physical world for making a correspondence between the two images.

The constraints that we need are the following, and they look deceptively simple; (1) A given point on a physical surface has a unique position

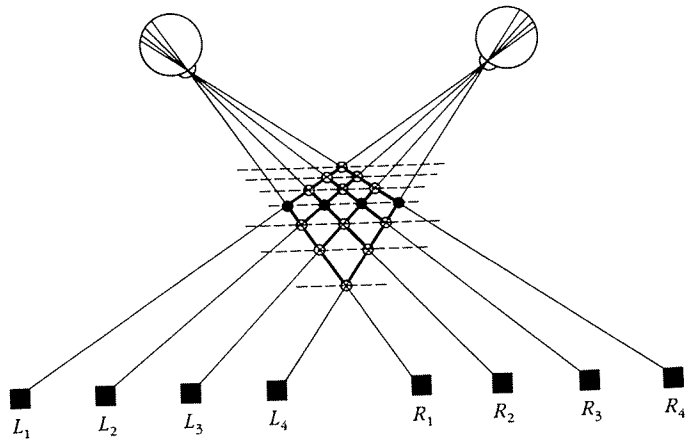


Figure 3–5. Ambiguity in the correspondence between the two retinal projections. In this figure, each of the four points in one eye's view could match any of the four projections in the other eye's view. Of the 16 possible matchings, only 4 are correct (filled circles); the remaining 12 are false targets (open circles). Without further constraints based on global consideration, such ambiguities cannot be resolved. The targets (filled squares) are assumed to correspond to matchable descriptive elements obtained from the left and right images. (Reprinted by permission from D. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science 194,* October 15, 1976, 283–287. Copyright 1976 by the American Association for the Advancement of Science.)

in space at any one time; and (2) matter is cohesive, it is separated into objects, and the surfaces of objects are generally smooth in the sense that the surface variation due to roughness cracks, or other sharp differences that can be attributed to changes in distance from the viewer, are small compared with the overall distance from the viewer.

These observations are properties of physical surfaces, and they constrain the behavior of the surface position. Hence, if we want to use these observations to help us establish a correspondence between two images of a surface, we must ensure that the items to which we apply them are in one-to-one correspondence with well-defined locations on a physical surface. To do this, we must use image predicates that correspond to surface markings, shadows, discontinuities in surface orientation, and so forth.

These physical considerations were precisely the motivation for the primal sketch, as we saw in Chapter 2, and that is why the primal sketch can be used, because the descriptive items in it—line and edge segments,

blobs, terminations and discontinuities, and tokens obtained from these by grouping—usually correspond to items that have a physical existence on a surface. And it is perhaps worth pointing out here that since the grouping processes have to be rather catholic in what they are prepared to group together, the larger and more abstract tokens tend to be less reliable than the very early and primitive things in the raw primal sketch. This is particularly relevant to stereopsis for another reason: Large-scale tokens are quite large, perhaps several degrees, whereas useful disparities tend to be rather small, on the order of minutes. To make accurate measurements, therefore the smaller, more primitive descriptors are preferred. On the other hand, clear statistical effects are likely to be quite a reliable indication of a physical change even at quite high levels so that high-level boundaries of the kind I called texture discrimination boundaries are probably more useful for stereopsis than aggregates at the same high level. We shall meet what I think are some consequences of this later on.

We can therefore rewrite the physical constraints as matching constraints, which restrict the allowable ways of matching two primitive symbolic descriptions, one from each eye. For the matching constraints to be valid, the elements in the matched descriptions must correspond to well-defined locations on the physical surface being imaged. We can think of these elements as carrying only position information, like the black dots in a random-dot stereogram, although for a full image, rules will exist that specify which matches between descriptive elements are possible and which are not. These rules will again be deducible from the physical situation; if the two descriptive elements could have arisen from the same physical marking, then they can match. If they could not have, then they cannot be matched. This is our first matching constraint, which I shall call the *compatibility* constraint.

The second and third matching constraints come from the two physical constraints. The uniqueness constraint means that, except in rare cases, each descriptive item can match only one item from the other image. The exceptions can arise as a result of the imaging process when two markings lie along the line of sight from one eye but are separately visible from the other. The third constraint, *continuity*, means that disparity varies smoothly almost everywhere. This constraint follows because the second physical constraint implies that the distance to the visible surface varies continuously except at object boundaries, which occupy only a small fraction of the area of an image.

These three restrictions, then, are our constraints. We now turn them to our purposes by making what I shall call the *fundamental assumption of stereopsis: If a correspondence is established between physically meaningful primitives extracted from the left and right images of a scene that contains a sufficient amount of detail, and if the correspondence satisfies the three matching constraints, then that correspondence is physically correct.* It follows immediately from this assumption that the correspondence must be unique.

But this is all very well, the skeptical reader will say. The matching constraints look perfectly reasonable and even quite powerful. But to turn them into a fundamental assumption which asserts that they are not only necessary consequences of the physical world but also actually *sufficient* to determine uniquely the correct correspondence—now that is an altogether different matter.

To say this is absolutely correct and hits fairly and squarely upon a philosophical point that constitutes one of the foundations of the approach. For to isolate this fundamental assumption and to establish that it is valid is precisely what I mean by the computational theory of a process. Establishing the sufficiency of this assumption here is more difficult than establishing the sufficiency of the spatial coincidence assumption that we met in Chapter 2, because that is a rather simple assumption which follows quite directly from the structure of the physical world.

However, we can establish validity for a wide range of situations. I shall try to show here in more general terms how the argument runs, because the underlying methodological point is so important. We shall meet it at the heart of the theory of every process.

As formulated, the fundamental assumption of stereopsis contains phrases like "scene that contains a sufficient amount of detail" and "physically meaningful primitives," which are too imprecise for mathematical demonstrations. So I will replace the phrase "physically meaningful primitives" by employing the special case of a physical surface that is white with black dots on it, and the first phrase by specifying the condition that the density—call it v—of the dots be sufficiently high; specifically, we shall need v to be at least 2% or so for our demonstration to work. By these somewhat devious means, analogous to spraying the world with black paint spots, I have converted the real-world situation into images that bear an uncanny resemblance to one of Julesz's random-dot stereograms. The matching conditions now obtain between the two binary images, and when translated, they read as the following three rules:

Rule 1: *Compatibility.* Black dots can match only black dots.

Rule 2: *Uniqueness.* Almost always, a black dot from one image can match no more than one black dot from the other image.

Rule 3: *Continuity.* The disparity of the matches varies smoothly almost everywhere over the image.

Our task now is to prove that these rules force a unique correspondence between the two images, and we can do this in the following way. First, note that because the two eyes lie horizontally, we need consider only all the possible matches along horizontal lines; therefore, we can reduce the problem to the simple one-dimensional case illustrated in Figure 3–6(a). L_x shows all possible positions for dots on the left retina, and R_x for dots on the right retina. The continuous vertical and horizontal lines represent the lines of sight from the left and right eyes, respectively; the dotted diagonal lines, marking traversals at the same rate across the left and right images, therefore represent planes of constant disparity.

Our proof is now easy, at least in conception. Rule 1 tells us to consider only black dots. Rule 3 tells us that, on the whole, the correct matches cluster along or close to these diagonal lines, and Rule 2 tells us that, at each point, only the matches along one of these planes should be chosen. The density of dots in each image is ν, so on the correct plane the density of possible matches is ν. On the incorrect planes it is only ν^2. Hence, provided the disparity changes slowly enough so that the area A spent on each disparity plane is big enough for $A\nu$ to be significantly different from $A\nu^2$, the three rules will yield a unique solution. Hence, since the solution is unique (following the $A\nu$ matches), it is physically correct, since the physically correct situation will yield one solution. That is the gist of the argument. Of course, this version is somewhat baldly stated, and various subtleties have to be attended to.

The arguments I have given have established two things. First, the fundamental assumption of stereopsis is valid, and this is why the constraints that it incorporates were derived from arguments based on the structure of the physical world. And second, the fundamental assumption provides a sufficient basis for defining the matching process, since a matching that satisfies it is guaranteed to be correct. Furthermore, there will always be such a match in normal physical situations. This completes the computational theory of stereopsis.

Algorithms for stereo matching

A cooperative algorithm

In order to drive home the point that more than one algorithm can be designed to implement a given process, I shall give two algorithms for the stereo matching process. The first one (Marr and Poggio, 1976) follows naturally from the thinking of the last section, and it can be understood most easily from the diagrams in Figure 3–6.

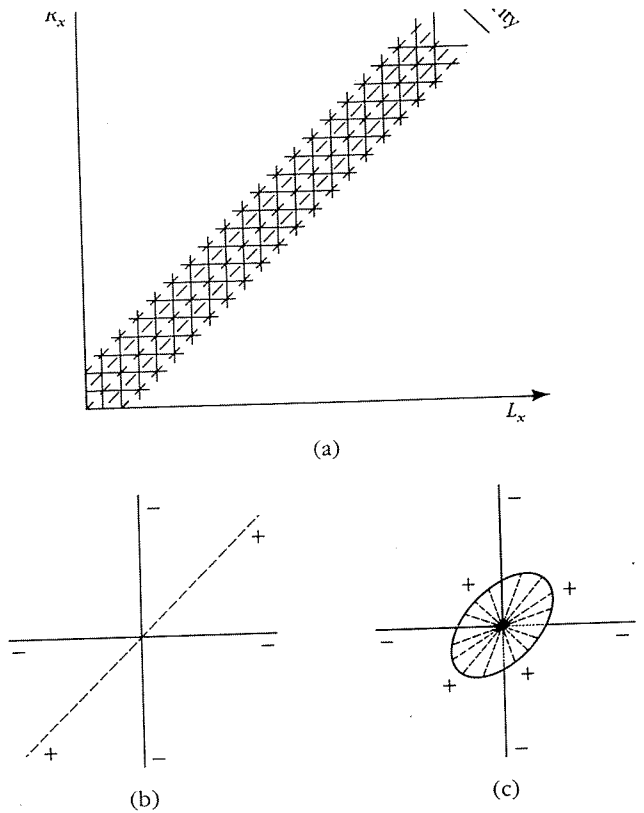


Figure 3–6. In (a), L_x and R_x represent the positions of descriptive elements in the left and right images. The continuous vertical and horizontal lines represent lines of sight from the left and the right eye. The intersections of these lines correspond to possible disparity values. The dotted diagonal lines are lines of constant disparity.

In the cooperative algorithm described in the text, a cell is placed at each node; then solid lines represent inhibitory interactions, and dotted lines excitatory. The local structure at each node of the network in (a) is given in (b). This algorithm may be extended to two-dimensional images, in which case each node in the corresponding network has the local structure shown in (c). The oval in this figure represents a two-dimensional disc rising out of the plane of the page. (Reprinted by permission from D. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science 194,* October 15, 1976, 283–287. Copyright 1976 by the American Association for the Advancement of Science.)

As we saw above, Rules 2 and 3 determine the solution to the matching problem. Rule 2 says in effect that only one match is allowed along any of the small vertical or horizontal lines in Figure 3–6(a). Rule 3 says that the correct matches tend to lie along the dotted diagonals.

What we do now is to make a parallel, interconnected network of processors that implements these two rules directly. At each intersection, or node, in Figure 3–6(a) we place a little processor. The idea is that if the node represents a correct match between a pair of black dots, then it should eventually have the value 1. If it represents an incorrect match—a false target, as we called it earlier—then the processor should have the value 0.

We implement the rules by interconnections between the processors. As we saw, Rule 2 tells us that only one match is allowed along each horizontal or vertical line. So, we make all the processors at the nodes along each vertical or horizontal line inhibit each other—the idea being that, in the resulting competition along each line, only one processor will survive to be 1, all the others will be 0, and so Rule 2 will be satisfied. Rule 3 says that correct matches tend to lie along the dotted lines, so we insert excitatory connections between processors in these directions. This gives each local processor the structure shown in Figure 3–6(b). Each such processor sends inhibitory connections to processors along the horizontal and vertical lines shown there, which correspond to the lines of sight from the two eyes, and excitatory connections along the diagonal line, which is the line of constant disparity. We can even extend the algorithm to two-dimensional images, in which case the inhibitory connections remain the same but the excitatory ones cover a small two-dimensional neighborhood of constant disparity. This situation is diagrammed in Figure 3–6(c).

The idea now is to load the network of processors by taking the two images and putting a 1 wherever two black dots could match—false targets and all—and a 0 at all other places. Then we let the network run. Each processor adds up the 1's in its excitatory neighborhood, adds up the 1's in its inhibitory neighborhoods, and subtracts the resulting figures (after multiplying one of the sums with a suitable weighting factor). If the result exceeds a certain threshold, the processor takes the value 1; if it does not, the processor is set to 0. Formally, this algorithm can be represented by the iterative relation

$$C^{t+1}_{x,y;d} = \sigma\left\{ \sum_{x',y';d' \in S(x,y;d)} C^{t}_{x',y';d'} \; - \; \varepsilon \sum_{x',y';d' \in O(x,y;d)} C^{t}_{x',y';d'} + C^{0}_{x,y;d} \right\}$$

where $C^{t}_{x,y;d}$ denotes the state of the cell corresponding to position (x,y), disparity d, and time t in the network of Figure 3–6(a); $S(x,y,d)$ is the local excitatory neighborhood, and $O(x,y,d)$ the inhibitory neighborhood. The Greek letter ε is an inhibition constant, and σ is a threshold function. The initial state C^{0} contains all possible matches, including false targets, within the prescribed disparity range; here it is added at each iteration. (It does not have to be, but the algorithm converges faster if it is.) Notice how Rules 2 and 3 are implemented through the geometry of the inhibitory and excitatory neighborhoods O and S.

This algorithm successfully solves random-dot stereograms, and an example is shown in Figure 3–7 of how the network gradually organizes itself into the correct solution. The stereograms themselves are labeled Left and Right, the initial state of the network as 0, and the state after n iterations is marked as such. To understand how the figures represent states of the network, imagine looking at the network from above—that is, from the direction of the top of Figure 3–6. The different disparity layers in the network lie in parallel planes, so that the viewer is looking down through them. In each plane, some nodes are on and some are off. Each of the seven layers in the network has been assigned a different gray level, so that a node that is switched on in the top layer (corresponding to a disparity of +3 pixels) contributes a dark point to the image, and one that is switched on in the lowest layer (disparity of −3) contributes a light point. Initially (iteration 0) the network is disorganized, but in the final state the order has stabilized (iteration 14), and the inverted wedding-cake structure has been found. The dot density of this stereogram is 50%.

The algorithm defined by the iterative relation above with the parameter values used for the example of Figure 3–7 is capable of solving random-dot stereograms with dot densities from 50% down to less than 10%. For this and smaller densities, the algorithm converges increasingly slowly. If a simple homeostatic mechanism is allowed to control the threshold σ as a function of the average activity (number of on cells) at each iteration, the algorithm can solve stereograms whose density is very low. In the second example, Figure 3–8, the density is 5% and the central square has a disparity of −2 pixels relative to the background. The algorithm fills in those areas where no dots are present, but it takes several more iterations to arrive near the solution than in cases where the density is 50%. When we look at a sparse stereogram, we perceive its shapes as being cleaner than the shapes found by the algorithm. This seems to be due to subjective contours that arise between dots that lie on shape boundaries.

We can see intuitively how the algorithm works from these examples. It never seems to have any trouble with stereograms, but this alone is not sufficient evidence for placing confidence in it. We did, however, manage to make it intellectually respectable; in a mathematical analysis of the algo-

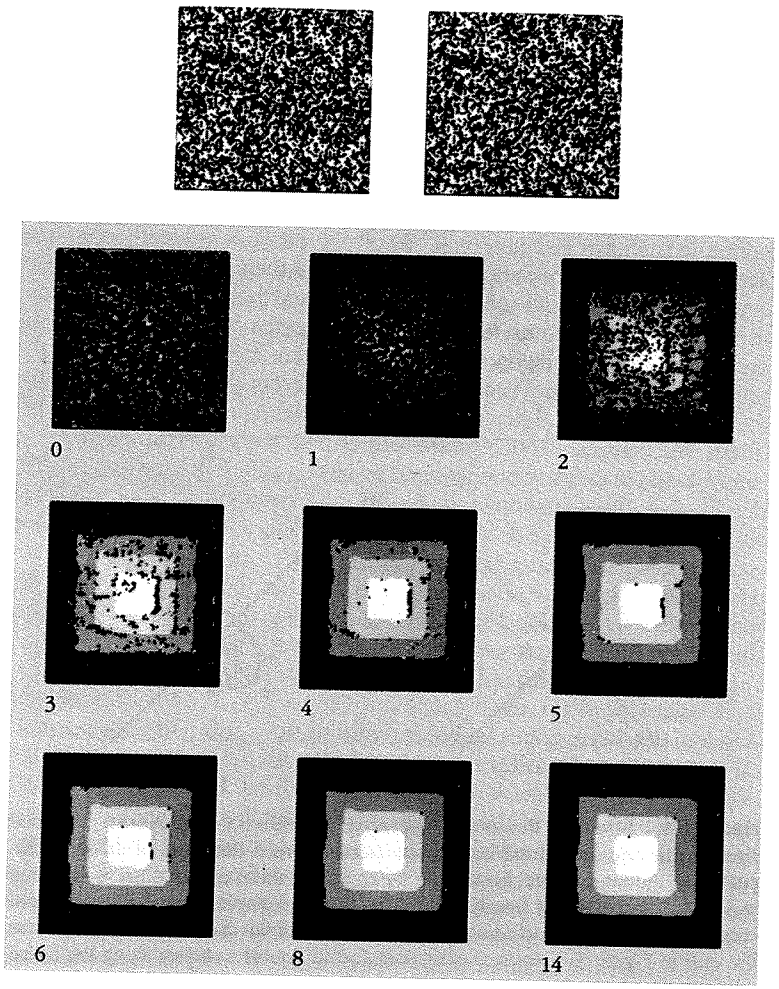


Figure 3–7. The decoding of a random-dot stereogram pair by the cooperative algorithm described in the text. The stereogram appears at the top, and the initial state of the network, which includes all possible matches within the prescribed disparity range, is labeled 0. The algorithm runs through a number of iterations, as shown, and gradually the structure is revealed. The different shades of gray represent different disparity values.

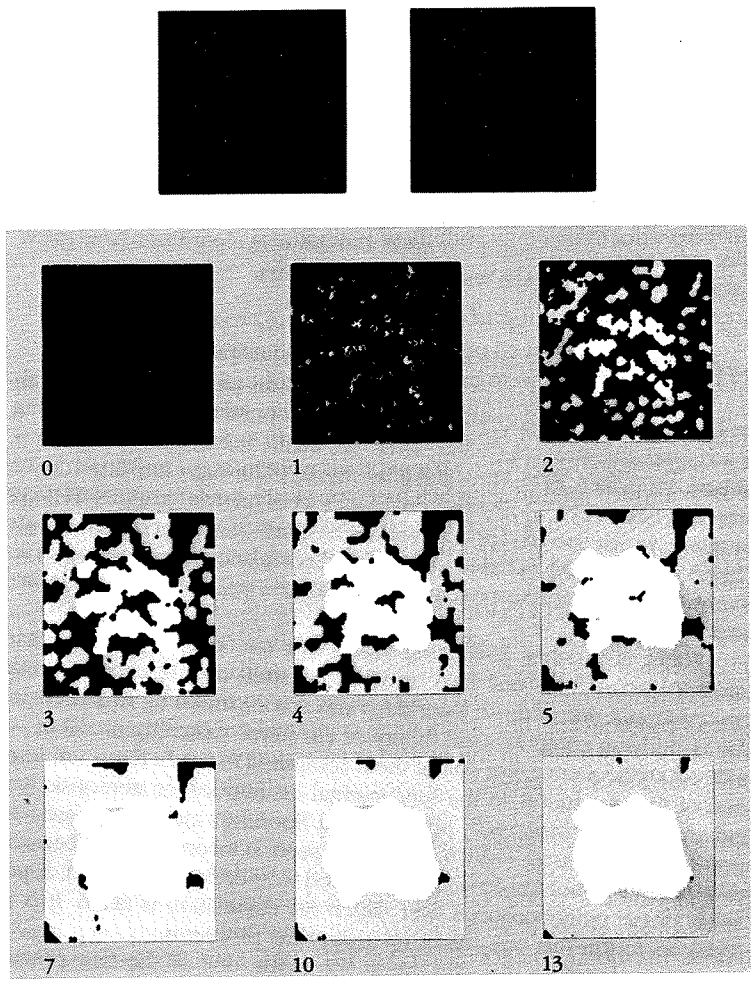


Figure 3–8. The algorithm used in Figure 3–7 can also decode and fill in very sparse stereograms. This one has a density of 5%.

rithm (Marr, Palm, and Poggio, 1978), we demonstrated that states obeying Rules 2 and 3 were stable states of the algorithm, and we showed that the algorithm converges for a wide range of parameter values.

This is an example of a cooperative algorithm, so-called because of the way in which local operations appear to cooperate in forming global order in a well-regulated manner. Cooperative phenomena are well-known in physics; for example, the Ising model of ferromagnetism, superconductivity, and phase transitions in general. Cooperative algorithms have many characteristics in common with these phenomena.

Cooperative algorithms and the stereo matching problem

Until 1977, almost all of the stereo algorithms put forward as models for human stereopsis were based on Julesz's proposal that stereo matching is a cooperative process (Julesz, 1971, pp. 203ff.; Julesz and Chang, 1976; Nelson, 1975; Dev, 1975; Hirai and Fukushima, 1976; Sugie and Suwa, 1977; Marr and Poggio, 1976). The two exceptions were Julesz's (1963) AUTOMAP program, which used an approach based on cluster-seeking, and Sperling's (1970) model, which is based on gray-level correlations but does make an interesting point of the connection between stereopsis and vergence movements.

There is a rather fascinating moral that one can draw from these attempts: Apart from our own, which was based on the computational approach, not one of these algorithms was accompanied by an analysis of the underlying computational theory of the stereo matching problem. As a direct consequence, not one of them computed the right thing—at least one of the constraints in the fundamental assumption of stereopsis was either missing or incorrectly implemented. Sperling's model was based on gray-level correlation—which, as we have seen, is incorrect—and because this model was not implemented, he failed to specify the area and disposition of the neighborhoods over which the correlation is taken. It is in trying to do this that one comes up against the problems.

Dev's algorithm deserves credit for being one of the first precise attempts to embody Julesz's ideas (Dev, 1975, eqs. 1 and 2). The algorithm realizes Rule 3 but employs an incorrect version of Rule 2. Instead of two lines of inhibition, one down each line of sight, she has one that bisects the angle between the lines of sight. This algorithm, illustrated in Figure 3–9, should be contrasted with the geometry of Figure 3–6. Physically, the connections in Figure 3–9 correspond to something like the rule that any direction out from the viewer meets only one surface. This is not true in general; for example, when one looks into a shallow lake, one sees two surfaces, the lake surface and its bottom. The correct version, shown in

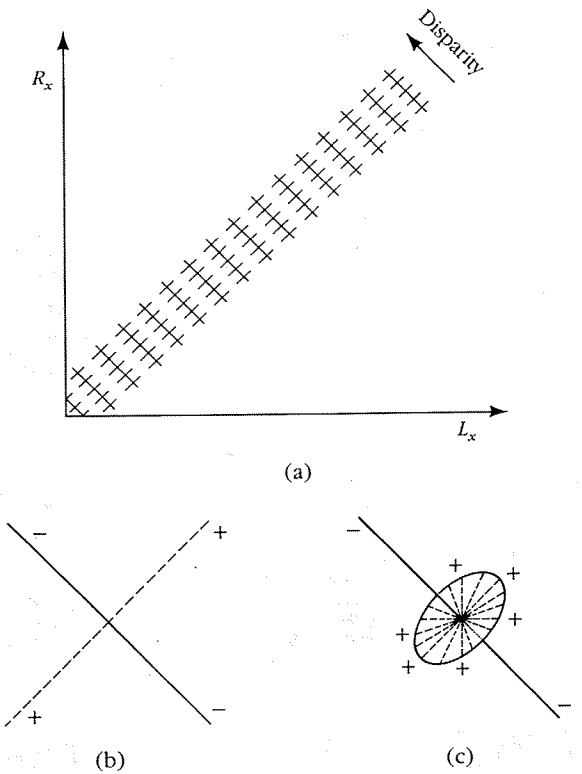


Figure 3–9. Several of the cooperative stereo algorithms that have been proposed include just one set of inhibitory connections between detectors of different disparities at the same retinal position. If we represent these connections in the same way as in Figure 3–6, it becomes obvious that they implement slightly different constraints. Instead of forbidding double matches down each line of sight, as was the case in Figure 3–6, these connections forbid double matches along the radial out from the viewer. It is incorrect to formulate the stereo correspondence process in this way.

Figure 3–6, says that any particular visible marking will lie either on the lake's surface or on the bottom (or perhaps on a fish swimming by), but only on one of these.

Sugie and Suwa's (1977) algorithm implements only a part of Rule 3 and the same, incorrect version of Rule 2. Nelson (1975) gave no precise algorithm, nor did he implement any form of his ideas, but he also seems

to mean an algorithm that implements the wrong form of Rule 2. Hirai and Fukushima (1976) correctly implemented Rule 2 (p. 48, function [1]) but did not implement Rule 3, preferring instead a network that favored solutions with lower parallax.

Julesz's (1963) AUTOMAP fails to implement Rule 2 but implements Rule 3 implicitly in the way it detects clusters. Julesz's dipole model is more interesting. It is defined as a mechanical analogy in which the left and right stereo images are each represented by a network of compass needles (magnetic dipoles), one for each image marking to be matched. The needles are oriented so that they can point to nearby locations in the opposite image's network when the two networks are overlayed. The endpoints of neighboring needles on each side are coupled together by springs and the polarity of each needle (north or south) is chosen according to the intensity of the image (black or white) at that location. The idea is that when the left and right networks are overlayed in rough registration, the magnetic attraction between similarly arranged groups of needles on either side will cause the network to settle into a stable state with the needles pointing towards their correct matches on the other side. While the relation between the polarity of the magnets and the retinal intensity values is unclear except for random-dot stereograms, the dipole model implicitly implements uniqueness, Rule 2, because a given dipole can have only one orientation at a time. Spring coupling between the tips of adjacent dipoles implements the continuity of Rule 3. This model therefore comes the closest to meeting our requirements, but it has the interesting feature that, unlike the other cooperative models, it does not represent explicitly all possible nodes in the diagram of Figure 3–6(a). That is, there is really only one processor for each vertical or horizontal line in that diagram, the different nodes along them being represented by different angular positions of a single dipole. It would be interesting to see whether such a model could be made to work.

The reason for elaborating upon this point is simply to help my overall argument that intellectual precision of approach is of crucial importance in studying the computational abilities of the visual system. Unless the computational theory of a process is correctly formulated, the algorithm will almost certainly be wrong.

Finally, none of these algorithms has been shown to work on natural images. Gray-level correlation works some of the time, but it makes mistakes that a human operator has to correct. The other proposals make no specific suggestions about what their input representations should be, although Marr and Poggio (1976) suggested that the primal sketch is suitable.

Biological evidence

All of these algorithms are designed to select correct matches in a situation where false targets occur in profusion. Consequently, apart from early versions of Julesz's dipole model perhaps, they do not critically rely upon eye movements, since in principle they have the ability to interpret a random-dot stereogram without them. However, eye movements seem to be important for human stereo vision. Without them, in fact, one can see very little depth—the range over which one can fuse two images (called Panum's fusional area) is small, about 6′–18′ of arc (Fender and Julesz, 1967; Julesz and Chang, 1976)—and almost no structure can be perceived (Richards, 1977), except for small disparities (Mayhew and Frisby, 1979). For complex stereograms such as Julesz's spiral (1971, fig. 4.5–4), eye movements are probably essential (Frisby and Clatworthy, 1975; Saye and Frisby, 1975). In fact, in view of Fender and Julesz's early findings, it is quite surprising that so little psychophysical attention has been given to eye movements until very recently.

There are several other psychophysical phenonema that would be difficult to explain in terms of the type of algorithms we have been discussing. Some subjects, for example, can tolerate a 15% expansion of one image (Julesz, 1971, fig. 2.8–8). If one severely defocuses one of the pair in a stereogram, fusion is easy to obtain (Julesz, 1971, fig. 3.10–3). This is only the most striking demonstration of a phenomenon that can be shown in several other ways. In fact, one can simultaneously experience both binocular rivalry and fusion of different spectral components in a stereogram, as the reader may experience in Figure 3–10 (Kaufman, 1964; Julesz, 1971, sec. 3.9 and 3.10; Julesz and Miller, 1975; Mayhew and Frisby, 1976). Such findings raise the interesting possibility that disparity information is conveyed at some stage by independent stereopsis channels that are tuned to different frequencies and are roughly one and a half octaves wide—very reminiscent, in fact, of the different-sized $\nabla^2 G$ operators that we met in Chapter 2.

Other interesting findings are the physiological, clinical, and psychophysical evidence about Richards' two-pools hypothesis (Richards, 1970, 1971; Richards and Regan, 1973; Poggio and Fischer, 1978; Clarke, Donaldson, and Whitteridge, 1976). Richards' basic finding was that stereo blindness manifests itself as a blindness to all convergent disparities, all divergent disparities, or both—and some kind of stereo incapacity, incidentally, is extraordinarily common, having an incidence of about 30%. In other words, stereo detectors seem to be organized into two pools, one dealing with convergent and the other with divergent disparities, with perhaps a third pool dealing with zero disparity. The neurophysiologists report some-

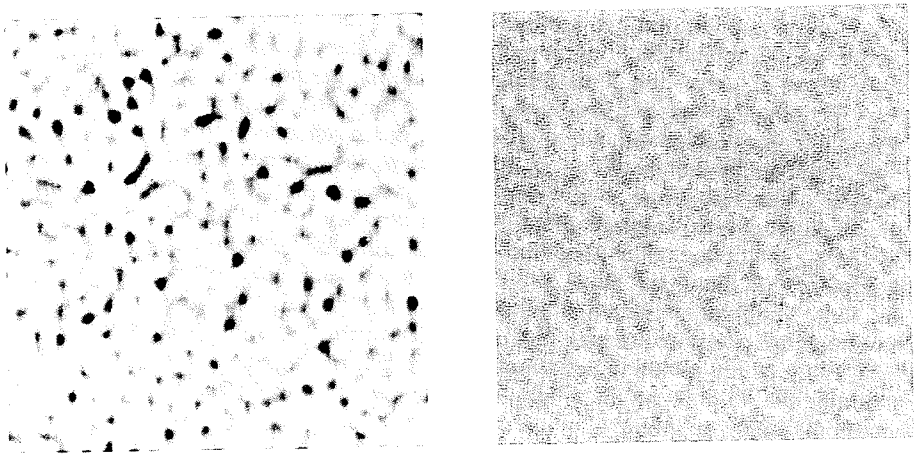


Figure 3–10. The high-frequency spectral components of this stereogram are rivalrous, yet the low-frequency components are not and can be fused. This suggests that independent spatial-frequency-tuned channels are involved in stereopsis. (Reprinted, by permission, from B. Julesz and J. E. Miller, "Independent spatial-frequency-tuned channels in binocular fusion and rivalry," *Perception 4,* 1975, 125–143, fig. 6.)

thing similar—roughly three classes of disparity-tuned neurons, one class broadly tuned to convergent (the so-called near neurons), and another broadly turned to divergent (far neurons), and a third sharply tuned to near-zero disparities. This goes against what one would expect of a neural implementation of the algorithms I discussed above, since, apart from the dipole model, all require many "disparity-detecting" neurons, whose peak sensitivities cover a range of disparity values that is much wider than the tuning curves of the individual neurons.

Finally, a remark about the motivation for the cooperative algorithm approach. As I have mentioned, these ideas were all inspired by Fender and Julesz's (1967) exhibition of hysteresis in stereopsis. In their experiment, they stabilized the images against eye movements and showed that once fusion was achieved, the two images could be "pulled" apart by up to about 2° of disparity before fusion "broke." However, once fusion had broken, the images had to be brought back to the 6′–14′ range before they would refuse. Hysteresis is one property of cooperative algorithms, and so is filling-in, which also seems to occur in stereopsis—as the reader has already seen, sparse stereograms like Figure 3–8 give the appearance of a smooth, solid surface, not of a few dots hanging isolated in space. Hence

everybody, including Julesz and ourselves, searched for a cooperative algorithm.

But not very sensibly. After all, the critical point of the Fender and Julesz experiment was that the hysteresis occurred over 2° of disparity, whereas matching only occurred under 20′. It therefore seems unlikely that the hysteresis is a consequence of the matching process, and much more likely that it is due to a cortical memory that stores the results of the matching process but is distinct from it. Fender and Julesz even suggested such a thing. Of course, this does not forbid the presence of cooperativity in the matching process, and the so-called pulling effect, described later by Julesz and Chang (1976), is probably evidence for its existence; however, the lesson is that we should probably deemphasize our ideas about cooperative processes and look instead for a rather different approach to the problem of stereopsis.

A second algorithm

The basic problem to be overcome in binocular fusion is the elimination or avoidance of false targets, and its difficulty is determined by two factors: the abundance of matchable features in an image and the disparity range over which matches are sought. If a feature occurs only rarely in an image, the search for a match can cover quite a large disparity range before false targets are encountered, but if the feature is a common one or the criteria for a match are loose, false targets can occur within quite small disparities.

For a given disparity range, then, if we want to simplify the matching problem, we have to decrease the incidence of matchable feature pairs; that is, we have to make features rare. There are two ways to do this. One way is to make them quite complex or specific, so that even if their density in the image is high, there would be so many different kinds that there would seldom be a compatible pair. The other way is to reduce drastically the density of all features in the image, for example, by decreasing the spatial resolution at which it is examined.

We know from Julesz's work on random-dot stereograms that the prospects for the first approach are rather slim. We know that the matching is carried out locally, yet all the edges are exactly vertical or horizontal and all have the same contrast, so even forcing very specific criteria onto them would not help us much. Furthermore, doing so would severely impair performance on real images, for which the orientations and contrasts of two corresponding edges can differ by surprising amounts. The reader can see for himself that stereograms with different contrasts can be fused by

Figure 3–11. The left and right images have different contrasts, yet fusion is still possible.

looking at Figure 3–11. The contrasts must, however, have the same sign. The criteria for orientation are also quite lax.

However, the other possibility is more promising. Indeed, the existence of independent spatial-frequency-tuned channels in binocular fusion now acquires a new and special interest, because it suggests that several copies of the image, obtained by successively finer filtering, are used during fusion, providing increasing and, at the limit, very fine disparity resolution at the cost of decreasing disparity range.

A notable feature of a system organized along these lines would be its reliance on eye movements for building up a comprehensive and accurate disparity map from two viewpoints. The reason for this is that the most precise disparity values are obtained from the high-resolution channels, and eye movements are therefore essential so that each part of a scene can ultimately be brought into the small disparity range within which high-resolution channels operate. The importance of vergence eye movements is also attractive in view of the extreme precision with which they may be controlled (Riggs and Niehl, 1960; Rashbass and Westheimer, 1961a).

These observations suggest the following scheme for solving the fusion problem: (1) Each image is analyzed through channels of varying coarseness and matching takes place between corresponding channels from the two eyes for disparity values of the order of the channel resolution; (2) coarse channels control vergence movements, thus causing fine channels to come into correspondence.

This scheme contains no hysteresis and therefore does not account for the observations of Fender and Julesz (1967). According to our emerging theory of intermediate visual information processing, however, a key

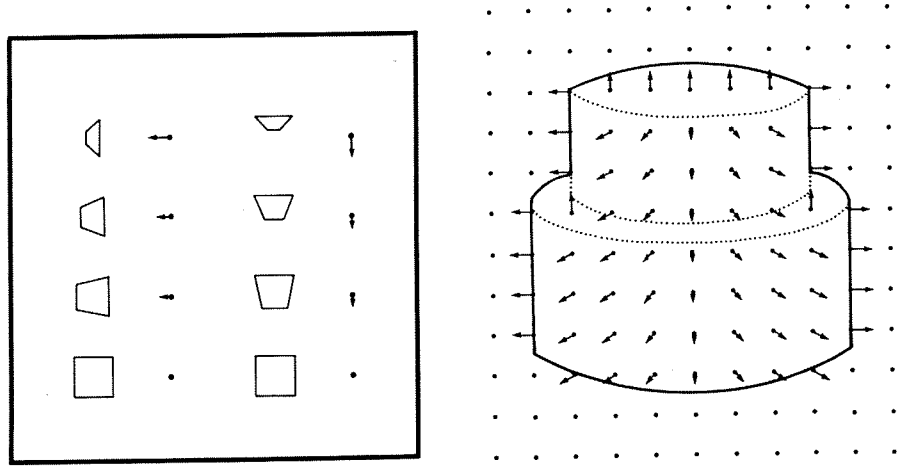


Figure 3–12. Illustration of the 2½-dimensional sketch. In (a), the perspective views of small squares placed at various orientations to the viewer are shown. The dots with arrows symbolically represent the orientations of such surfaces. In (b), this symbolic representation is used to show the surface orientations of two cylindrical surfaces in front of a background orthogonal to the viewer. The full 2½-dimensional sketch would include rough distances to the surfaces as well as their orientations; contours where surface orientations change sharply, which are shown dotted; and contours where depth is discontinuous (subjective contours), which are shown with full lines. See Chapter 4 for more details. (D. Marr and H. K. Nishihara, 1978.)

goal of early visual processing is the construction of something like an orientation-and-depth map of the visible surfaces around a viewer (see Chapter 4). In this map, information is combined from a number of different and probably independent processes that interpret disparity, motion, shading, texture, and contour information. These ideas are illustrated by the representation shown in Figure 3–12, which Marr and Nishihara (1978) called the 2½-D sketch.

Suppose now that the hysteresis that Fender and Julesz observed was not due to a cooperative process during matching but was in fact the result of using a memory buffer, like the 2½-D sketch, for storing the depth map of the image as it is discovered. Then the matching process itself need not be cooperative (even if it still could be): it would not even be necessary for the whole image ever to be matched simultaneously, provided that a depth map of the viewed surface was built and maintained in this intermediate memory.

Our scheme can now be completed by adding to it the following two steps: (3) When a correspondence is achieved, it is held and written down in the 2½-D sketch; (4) there is a reverse relation between the memory and the channels, acting through the control of eye movements, that allows one to fuse any piece of surface easily once its depth map has been established in the memory.

The idea of matching coarse, widely separated features first, and then with the information so obtained, repeating the matching process at successively finer scales of resolution sounds promising, but what features should we match at these different resolutions? We have seen enough of early visual processing to suggest various possibilities. Are they zero-crossings, the raw primal sketch, the full primal sketch, or some combination of them all? Poggio and I proposed that the input representation for the stereo matching process consists of the raw zero-crossings, labeled by the sign of their contrast change and their rough orientation in the image, and of terminations—local discontinuities—also labeled by contrast and perhaps very rough orientation.

The matching process. The choice of input representation leads to the matching algorithm illustrated in Figures 3–13 and 3–14. These figures show Eric Grimson's computer implementation of the algorithm running on a pair of random-dot stereograms, which represent one of the most difficult kinds of input for the algorithm.

The left and right images, forming a random-dot stereogram with density 50%, appear at the top of Figure 3–13. The first step in the algorithm is to apply a large $\nabla^2 G$ filter to each image and obtain the zero-crossings, just as we did in Chapter 2. Although in theory the elements to be matched between images include both zero-crossings and terminations, it is only the zero-crossings that cause difficulties with false targets. Thus Figure 3–14 shows only the zero-crossings and in fact horizontal segments are ignored, since they cannot be easily matched.

In addition to their locations, the zero-crossings have been given a sign and a rough orientation. The sign corresponds to the sign of the contrast change from left to right across the zero-crossing, and it is indicated by the shade of the zero-crossing in the figure. Two zero-crossings are matchable if they have the same sign and their local orientations are within 30° of each other. Matching itself is carried out point by point along the zero-crossings.

The convolution values and signed zero-crossings for three sizes of the $\nabla^2 G$ filter appear in Figure 3–14. The reader can see that far more zero-crossings are obtained from the smallest channel than from the largest, which means that the disparity range considered can be greater for the larger channels without any increase in the incidence of false targets.

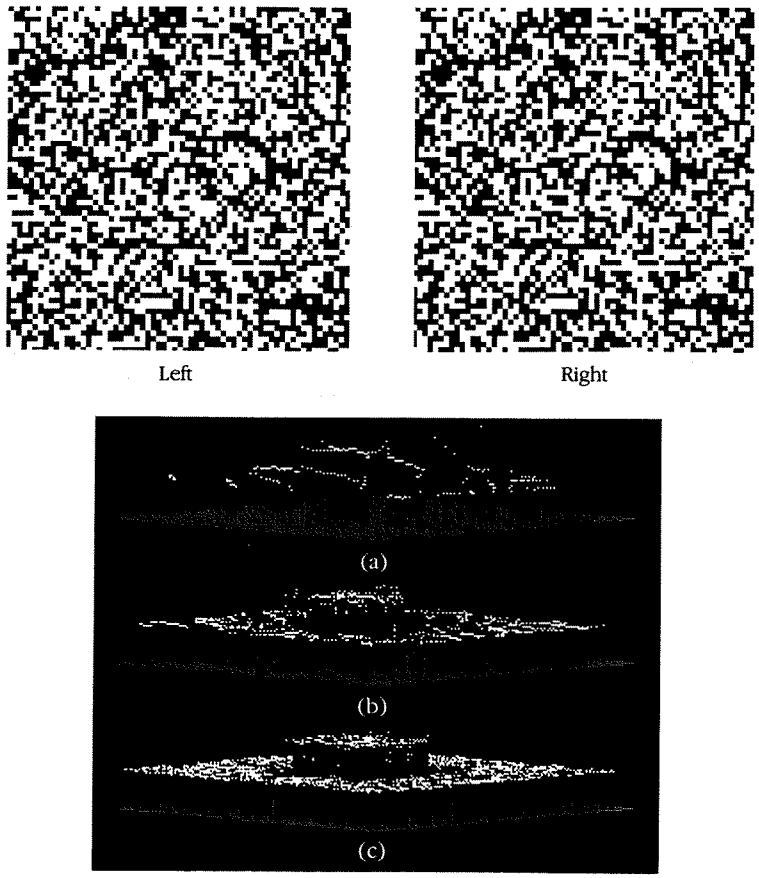
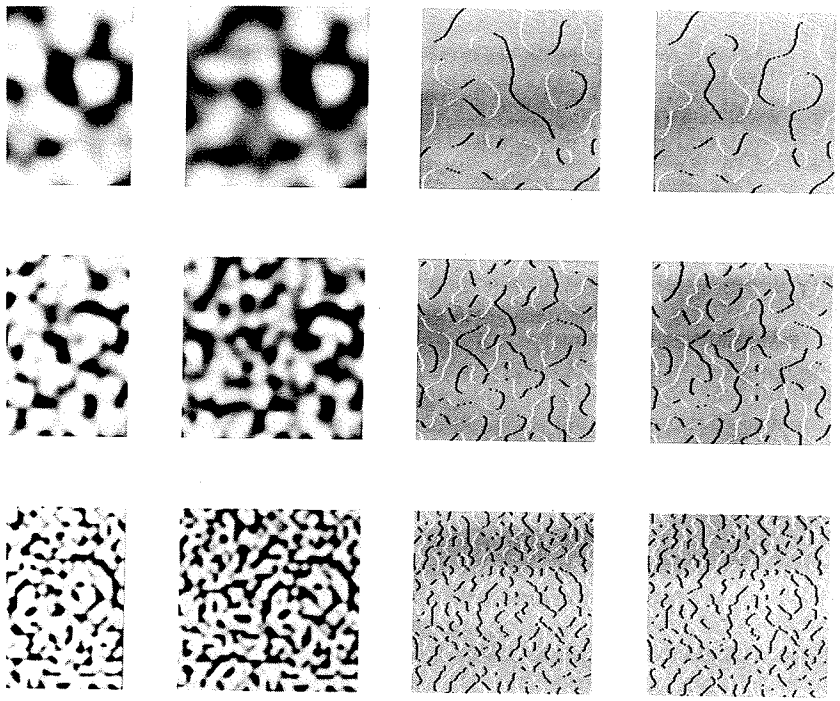


Figure 3–13. The solution of a 50% random-dot pattern. The left and right images are shown at the top. The three lower figures indicate an orthographic view of the disparity maps obtained by matching the zero-crossing descriptions of Figure 3–14. A point in the image with coordinates (x,y) and an assigned disparity value of d is portrayed in this three-dimensional system as the point (x,y,d). Here the heights of the bright points above the plane indicate their disparity values.

In general terms, then, the overall structure of the algorithm is clear from Figures 3–13 and 3–14. First, the coarse images are matched; the results of this are illustrated in Figure 3–13(a), which shows an orthographic view of the resulting disparity map. This rough result is used as the starting point for the same matching process applied to the medium-sized channel. The decrease in the allowed disparity range is offset by the knowledge, obtained from the large channel, of its approximate value. This



3–14. The convolutions and zero-crossings involved in solving the stereogram of Figure he two left columns indicate the convolutions of the left and right images with masks of $_{-D}$ = 35, 17, and 9, respectively, from top to bottom. The two right columns indicate the ɔssings obtained from the convolutions in the left two columns. Notice how much more ıe smaller masks reveal.

gives the disparity map shown in Figure 3–13(b). Second, the smallest channel is considered, yielding the accurate disparities made possible by its small disparity range, and the results appear in Figure 3–13(c). In this example, the central square has a disparity of 12 pixels, and each black square is 4 × 4 pixels. In the final disparity map, less than 0.1% of the points are incorrectly matched, and these all occur at the borders of the square.

More properties of zero-crossings. In this algorithm, the false target problem is solved essentially by evasion, but exactly how it is solved is

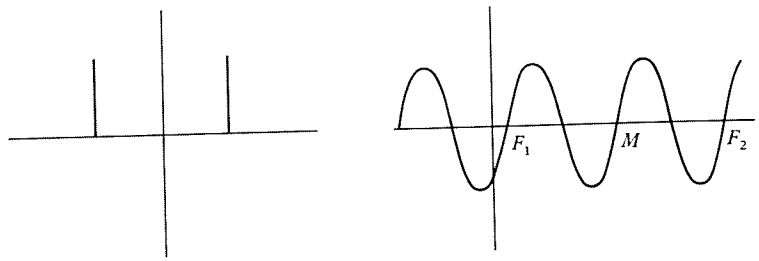
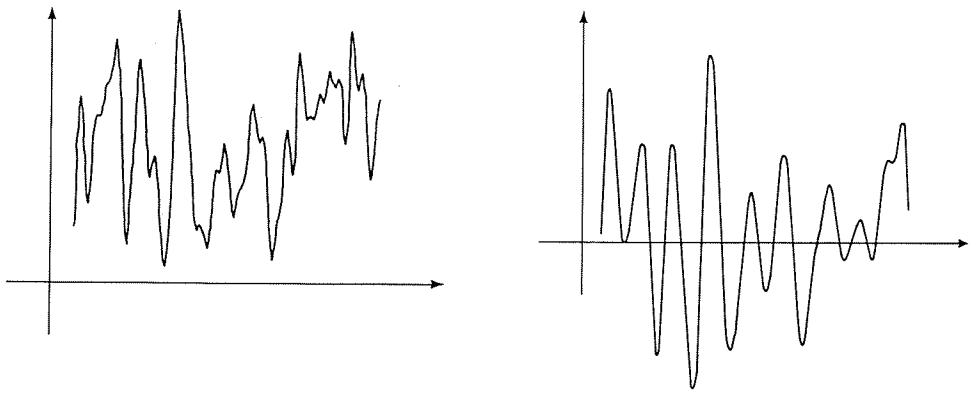


Figure 3–15. The positive (or negative) zero-crossings of a pure sine wave are guaranteed to be λ apart, where λ is the wavelength. See discussion in text.

interesting and, from the point of view of psychophysics, quite important. I shall not give the proofs here, but the general argument can be conveyed without much technical detail.

The central idea is illustrated in Figure 3–15. Suppose, for sake of argument, that the intensity variation in the image was purely sinusoidal, consisting solely of a vertically oriented sinusoidal grating. Such a signal has the Fourier transform shown in Figure 3–15(a) and passes unscathed through $\nabla^2 G$, giving the same curve shown in one-dimensional cross-section in Figure 3–15(b). Now the problem is to match the zero-crossings between the two filtered images, so let us suppose that we have fixed on a particular positive-going zero-crossing from the left image whose true match is the one marked M in Figure 3–15(b). Then F_1 and F_2 are false targets. But since they also have to be positive-going zero-crossings, they must be at least a distance λ away, where λ is the wavelength of the sinusoid. Hence, provided that we restrict our search for possible matches to a disparity range of at most λ, we are guaranteed to find only one possible match, and provided that we know by some other means roughly where to carry out the search, we can be sure that the one match we find will be correct.

That is the basic idea, but the real world is not restricted to pure sine wave gratings. A sine wave is, however, only the extreme case of a band-pass function, in which the bandwidth is zero. The same qualitative argument holds for wider bandwidths, and this can be seen roughly from Figures 2–19 and 3–16. For example, consider the case of an ideal one-octave band-pass filter of the type whose Fourier transform appears in Figure 2–19(b). A portion of a typical signal from such a filter is illustrated in Figure 2–19(c). The average value of this signal is zero, so the signal



(a) (b)

Figure 3–16. The signal in (a) varies randomly in the range 0 to 100. After being passed through the filter $\nabla^2 G$, it has the appearance shown in (b), with more or less regularly occurring zero-crossings. A similar example is given in Figure 2–19 for a pure one-octave band-pass filter. For general band-pass signals, like those passed by $\nabla^2 G$ or a pure one-octave filter, the zero-crossings cannot on average occur too closely together or too far apart. The intervals between zero-crossings are governed by the statistical rules illustrated in Figure 3–17.

crosses zero quite frequently, like the sine wave. However, because it is a band-pass signal, its zero-crossings cannot occur too far apart. On average, they occur at the frequency corresponding to the middle of the filter's range.

The important point for us is that zero-crossings cannot on average occur too close together, and this is true for any band-pass filter. The filter $\nabla^2 G$, however, is also roughly a band-pass filter—the reader may care to look once again at its one-dimensional Fourier transform, shown in Figure 2–9(c). The results of passing a random one-dimensional signal (Figure 3–16a) through $\nabla^2 G$ are shown in Figure 3–16(b), and the reader can see that it has the same qualitative features as Figure 3–15, its average value is zero, and the zero-crossings lie neither very close to nor very distant from their neighbors.

The general lines of the argument are now quite straightforward, and they are the same as those of the argument for the sine wave. Since $\nabla^2 G$ is roughly a band-pass filter, its zero-crossings are usually separated by some minimum distance. Provided we know approximately where to look for a match, and provided we do not search over too large a range, we shall find a unique candidate for the match and it will be correct.

This shows us a promising approach to the matching problem, but it also raises another rather exciting possibility. From the point of view of psychophysics, $\nabla^2 G$ is monocular, but matching is binocular. That is, the parameters of the $\nabla^2 G$ filters—their widths w_{1-D}, for instance—are obtained by purely monocular measurements. The disparity range for matching, usually called Panum's fusional area and which I shall denote by ∇, is essentially a binocular phenomenon. If our theory is true, it will predict a clear and unexpected relationship between these a priori unrelated quantities, which are measured in completely different ways. This will therefore provide an excellent way of testing the theory.

It is therefore important to derive the precise quantitative relationship that we expect should hold between w_{1-D} and ∇. In order to do this, we need a quantitative model for the channels used in early processing and some way of estimating the probable distances between zero-crossings. The idea of using zero-crossings, it should perhaps be said, came from early work on the primal sketch (Marr, 1976) in which many of the cells early in the visual pathway were thought of not as feature detectors but as differential operators. Hubel and Wiesel's (1962) definition of a cortical simple cell as linear led us to think, for example, of a bar-shaped receptive field as an oriented second-derivative operator from which one subsequently found zero-crossings. Only later did we come to realize that the simple cells themselves are probably the zero-crossing detectors, as in Figure 2–18 (see also Section 3.4). This slight confusion does not matter from a mathematical point of view, because under only very weak assumptions the two points of view are equivalent (see Marr and Hildreth, 1980, app. A). With respect to their implementation and consequently to psychophysics, the two things are rather different. I shall return to this point later on.

For our analysis, then, we need a quantitative hold on the distances between zero-crossings for the filters that the visual system actually uses. At the time the present stereo theory based on matching at different scales of resolution was formulated, we did not know that $\nabla^2 G$ was the optimal filter to use, but we knew something just as good, because Hugh Wilson at Chicago had just formulated his four-mechanism model for the structure of the channels. He described their structure using DOG's—differences of Guassians—which are almost indistinguishable from $\nabla^2 G$, as we saw in Figure 2–16.

We were also very lucky with the mathematics of the problem because obtaining estimates of the probable distances between the zero-crossings of band-pass signals turns out to be very difficult. Various mathematicians had already worked on it, starting with Rice in 1945 and more recently M. Longuet-Higgins (1962) and Leadbetter (1969). The problem itself is inter-

esting, because it relates to a number of physical phenomena, some quite important and some less so. The important ones include the effects of Brownian noise due to the random motion of electrons in electrical circuits—and some amplifiers, for example, switch as the voltage crosses zero—and the analysis of the distribution of wave heights in the sea, which is of particular interest now that people are trying to tap this source of energy. On a more frivolous note, the same type of mathematics is involved in the study of twinkles, which are the places in the sea that happen to reflect the sun back into your eyes, causing the surface to glitter and, well, twinkle.

We can therefore analyze the spatial distribution of zero-crossings, at least for one-dimensional band-pass signals. The results are illustrated in Figure 3–17 for two cases: first, the example shown in Figure 2–19 of a pure one-octave band-pass filter (left column), and second, the case, illustrated in Figure 3–16, of a $\nabla^2 G$ filter which closely approximates the filters that Wilson concluded are present in the early stages of the human visual system (right column).

The legend explains the details, but the important graphs are the two in Figure 3–17(c). They show the probability given a zero-crossing at the origin, of encountering another zero-crossing of the same sign at distance ξ away. The units in which ξ are plotted, in the biologically interesting case of the right-hand column, are such that w_{1-D} has the value 2.8. Two values of this probability are worth remembering: at distance w_{1-D} it is about 5%,

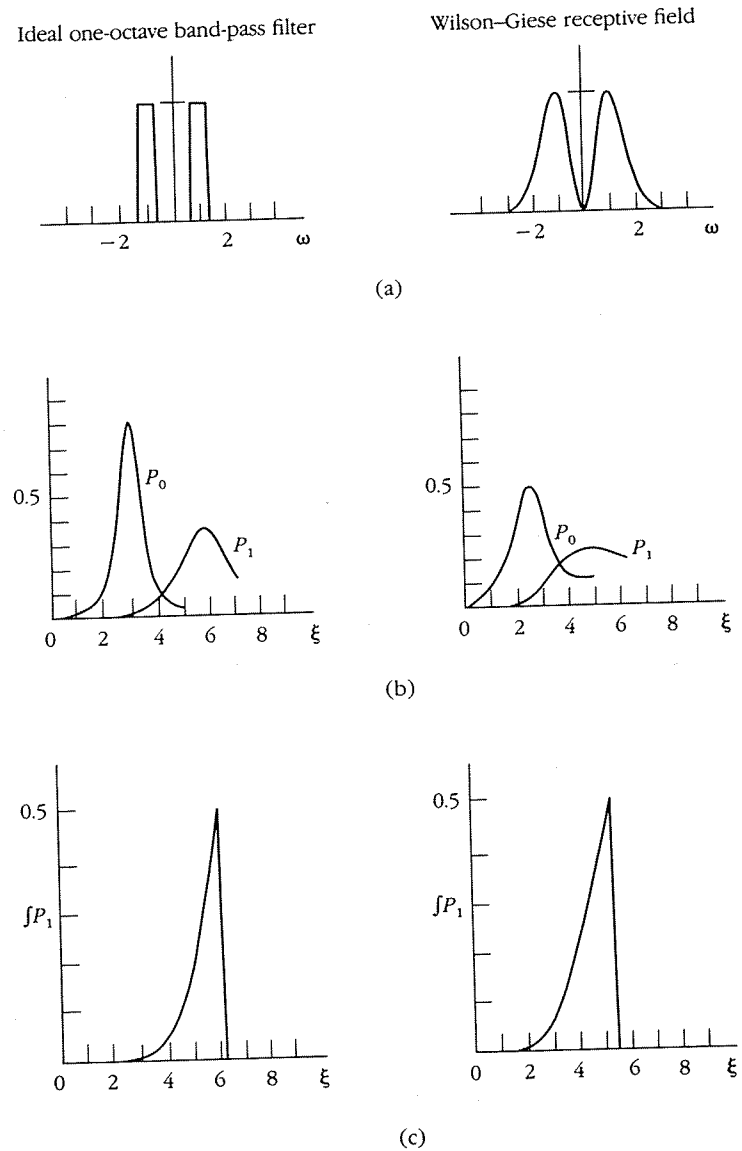


Figure 3–17. (opposite) Interval distributions for zero-crossings. A "white" Gaussian random process is passed through a filter with the frequency characteristic (transfer function) shown in (a). The approximate interval distribution for the first (P_0) and second (P_1) zero-crossings of the resulting zero-mean Gaussian process is shown in (b). Given a positive zero-crossing at the origin, the probability of having another within a distance ξ is approximated by the integral of P_1 and shown in (c). In the left column, these quantities are given for an ideal band-pass filter one octave wide and with center frequency $\omega = 2\pi/\lambda$; in the right column, these quantities are given for the case of the receptive field described by Wilson and Giese (1977). The ratio of space constants of excitation and inhibition is 1:1.5. The width w of the central excitatory portion of the receptive field is 2.8 in the units in which ξ is plotted. For the case portrayed in the left column, a probability level of $\int P_1 = 0.001$ occurs at $\xi = 2.3$ and a probability level of 0.5 occurs at $\xi = 6.1$. The corresponding figures for the case illustrated in the right column are $\xi = 1.5$ and $\xi = 5.4$. If the space-constant ratio is 1:1.75, the values of $\int P_1$ change by not more than 5%. (Reprinted by permission from D. Marr and T. Poggio, "A computational theory of human stereo vision," *Proc. R. Soc. Lond. B 204,* 301–328.)

and at distance $2w_{1-D}$ it is about 50% and increasing rapidly. Moderate changes in the shape of the underlying filter do not change these numbers by very much.

The matching algorithm. Given this background, we can now formulate the matching algorithm and prove that it will work. Let us first examine a simple case, in which false targets are essentially avoided. It is best explained by looking at Figure 3–18(a). Here we have a zero-crossing from the left image, marked L, which matches another of the same sign in the right image, which is displaced by a disparity of amount d. The correct match is labeled R, and a possible false target F, shown dotted, is shown lurking nearby. Provided that we consider only the disparity range $w/2$, however, we are safe, because even if R is right at one end of the range—if $d = w/2$, for example—our statistical analysis assures us that, with a probability of 95%, it will be the only zero-crossing of its type within a disparity range that extends over w. Even if we ignore all cases in which two candidates are present, we shall still succeed over 95% of the time.

This assumes, of course, that R *is* the correct match, that is, that the correct match lies in the range of $w/2$ that the procedure examines. However, we can tell when the correct match does not lie in this range, because if the visible surface has disparity in this range, almost all zero-crossings from the left image will find matches in the right image, and all of them will find at least one candidate match. If the surface has a disparity lying outside this range, then the probability that a zero-crossing from the left image will find a candidate match within range from the right image is, for all intents and purposes, simply the probability that a zero-crossing of the appropriate sign falls by chance within the particular spatial interval $w/2$ in the right image. This probability is about 40%. Hence, if the surface lies outside the disparity range, only 40% of the matches will be achieved versus nearly 100% if the surface falls within this range. It is therefore easy to tell when the matching process is succeeding. And notice, incidentally, that we rely on the third constraint, continuity, of our fundamental assumption, since it is assumed that we can look over a neighborhood in the image that is large enough to enable us to measure the difference empirically between a situation with a 40% probability of matches and one of, say, 95%. Such a neighborhood does not have to be very large, but it has to exist, and this is why we need the continuity assumption.

Now that this simple algorithm has given us the basic idea, we can improve on it, and by doing so increase the allowed disparity range from $w/2$ to w. Figure 3–18(b) shows our zero-crossing L in the left image, but this time its match R in the right image has a disparity d that can be as much as w. The first point to note is that if d is positive, then by the same

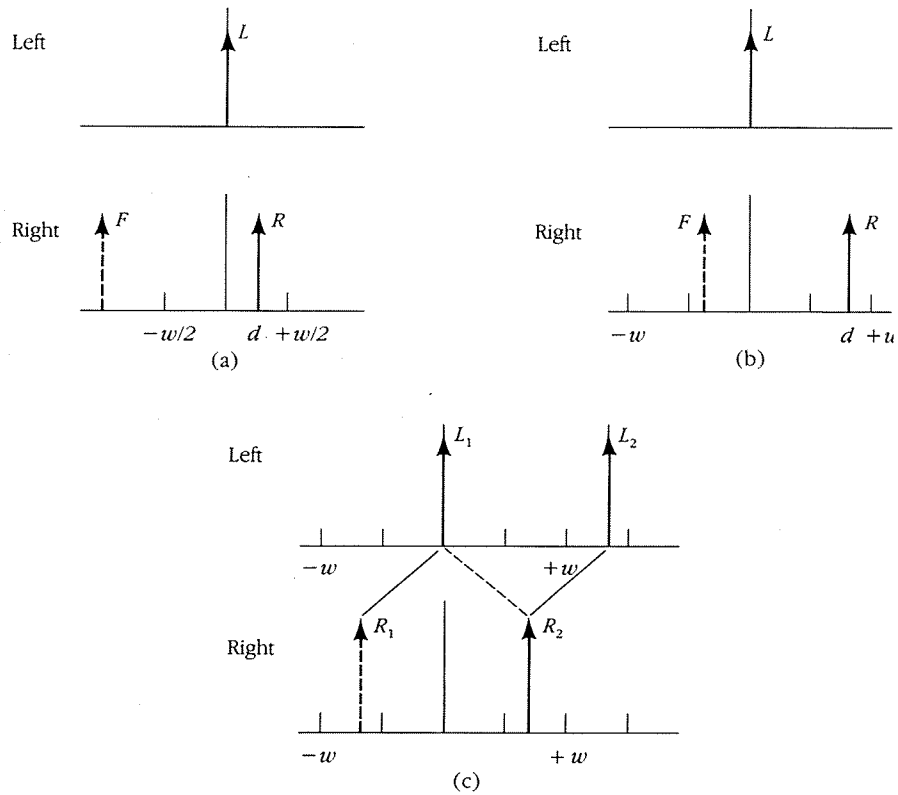


Figure 3–18. The matching process driven from the left image. A zero-crossing L in the left image matches one R displaced by disparity d in the right image. The probability of a false target within w of R is small, so provided that $d < w/2$ (a), almost no false targets will arise in the disparity range $w/2$. This gives the first possible algorithm. Alternatively, all matches within the range w may be considered (b). Here false targets, designated F, can arise in about 50% of the cases, but the correct solution is also present. If the correct match is convergent, the false target will with high probability be divergent. Therefore, in the second algorithm, unique matches from either image are accepted as correct, and the remainder as ambiguous and subject to the pulling effect, illustrated in (c). Here L_1 could match R_1 or R_2, but L_2 can match only R_2. Because of this and because the two matches have the same disparity, L_1 is assigned to R_1. (Reprinted by permission from D. Marr and T. Poggio "A computational theory of human stereo vision," *Proc. R. Soc. Lond. B 204,* 301–328.)

arguments as before, R is at least 95% certain to be the only candidate in the disparity range 0 to w. Second, we know from our statistics that the likelihood of a false target in the $2w$ disparity range from $d = -w$ to $d = +w$ is at most 50%, even when the correct match lies at one extreme of this range. Putting these two facts together, we see that at least 50% of the time the match will be unambiguous and correct and that remaining cases will be ambiguous, consisting mainly of two alternatives, one convergent (in the range $(0,w)$) and one divergent (in the range $(-w,0)$), one of which will be correct. In the ambiguous cases, selection of the correct alternative can be based simply on the sign of neighboring matches (note the use of continuity here). Notice, incidentally, that if a match is near zero disparity, it is likely ($p > 0.9$) to be the only candidate, again according to the statistics of the situation. Hence the notion of three disparity ranges—one convergent, one divergent, and one around zero—follows naturally from this matching technique.

Once again, if the surface lies within the disparity range, nearly 100% of the zero-crossings will find matches; if it does not, the figure in this instance is 70% instead of 40%, but this is still different enough from 100% to enable us to tell when matching is succeeding.

We cannot improve much on the range w without resorting to more powerful techniques for removing false targets, because the probability of false targets occurring increases quite sharply above the range $2w$. The percentage of unambiguous matches, for example, is already down to 20% at $1.5w$.

Uniqueness, cooperativity, and the pulling effect

Eric Grimson (1981) made the important point that matching can be carried out from either image or from both images. In Figure 3–18(c), for example, if matching is initiated from the left image, the match for L_1 is ambiguous, but for L_2 it is unique. From the right image, matching is unique for R_1 but ambiguous for R_2. Together the two unique matches provide the correct solution.

That the two unique matches should be correct rather than contradictory is a consequence of the uniqueness property embedded in the fundamental assumption of stereopsis. As a result, the algorithm can be designed to accept unambiguous matchings by starting from either image. However, this design does have some fascinating consequences, for it means that the uniqueness assumption is no longer internally verifiable by the algorithm, whereas the continuity assumption is.

This fact is determined in the following way. We have already seen that the algorithm needs to check the proportion of local candidates that are matched in order to tell whether the surface lies within the disparity range under examination. If the proportion is near 100%, everything is satisfactory. If it is not (in which case it is probably 70%), the solution is rejected. It is extremely difficult to fool this test, and since it relies on continuity for its validity, it amounts to an internal check that continuity is being locally satisfied by the visible surfaces.

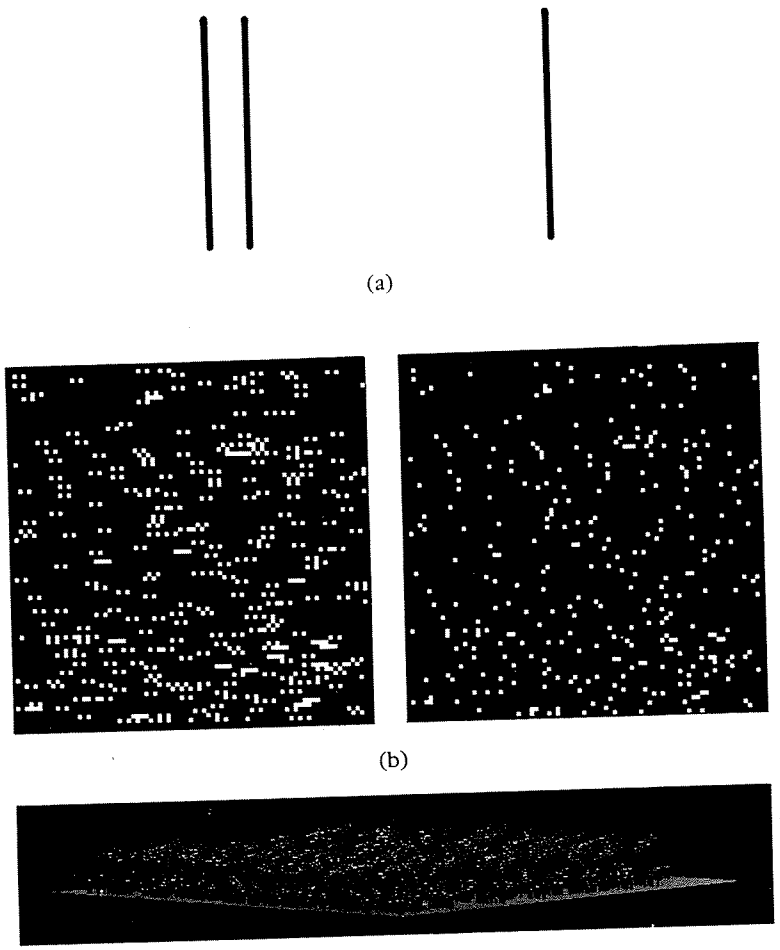
Not so uniqueness. If the algorithm accepts unique solutions from either image, this allows it to fuse patterns such as the Panum's limiting case example (Figure 3–19) not only for rare occurrences across an image, as in Figure 3–19(a), but also for frequent ones. Oliver Braddick investigated this point by constructing stereograms like Figure 3–19(b), in which each dot from the right image matches two from the left. Matching initiated from the left image is unique, so one accepts it, and the resulting percept is of two planes, one behind the other. The visual system is not particular about which eye it operates from, and one can mix the doublets up so that some of them are in the right image, and some are in the left. It makes no difference.

Physically, of course, this situation is effectively impossible to produce with two real surfaces, which is perhaps why we have not evolved an internal check for uniqueness. The general point here is an interesting one, though; some assumptions can be and are checked internally, like continuity here; some could be but are not, like uniqueness; and some cannot be even in principle. We shall meet some examples of this later on, but it may be worth mentioning here that the Ames room illusion may be one. Without stereopsis or motion cues, the assumptions of right angles cannot be tested internally.

Finally, there are situations in which matching is ambiguous from both eyes. In this case, the ambiguity can be resolved by consulting the signs of the neighboring matches and choosing the matches with the same sign. There is, however, an important distinction between the two most obvious ways of doing this. Either we consult the signs of the neighboring matches that were unambiguous from the start, or we consult the signs of the neighboring matches that have so far been assigned. The second scheme introduces cooperativity, the first does not.

To see this, imagine a stereogram cleverly constructed so that every match is ambiguous except for an unambiguous region located, for example, at the border. With the first scheme, none of the matches in the interior region of the stereogram will ever be disambiguated, because there are never any unambiguous matches to start from. With the second scheme, however, the disambiguation will gradually propagate from the borders, where the matches are determined, into the interior, where matches will eventually be chosen whose signs are those of the matches at the border.

Julesz and Chang (1976) did just this experiment, and an example of the type of stereogram they used appears in Figure 3–20. It transpired that

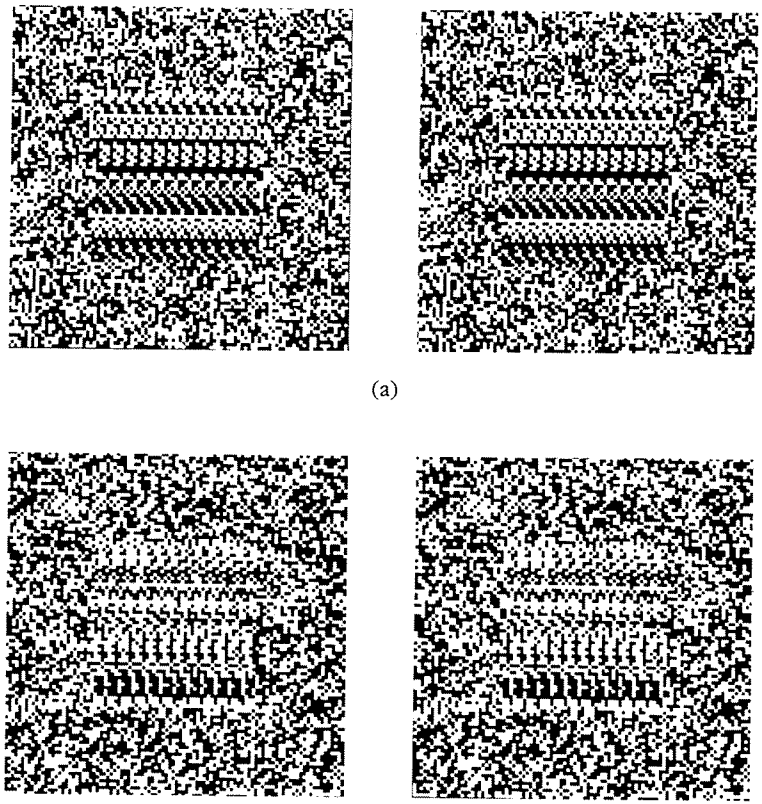


(a)

(b)

(c)

Figure 3–19. (a) Panum's original limiting case. When fused, the impression is of two lines separated in depth. In (b), each dot in the right image is paired with two in the left image. When fused, the viewer sees two planes. The doubling does not have to be restricted to one image. (c) The results of running the stereo algorithm on (b), disparity being displayed according to the same conventions as were used for Figure 3–13. Two planes are found.



(a)

(b)

Figure 3–20. (a) There are many possible ways of matching the center of this stereogram, but usually only the matches having the smallest disparity are perceived. However, the particular match found can be biased by inserting unambiguously matchable dots at a particular disparity. In (b), 6% of the dots in the upper half of the square have unambiguous matches at a two-dot crossed disparity—shifted in the nasal direction, whereas the lower half is biased with a two-dot uncrossed disparity. Even a bias inserted into the border will pull fusion to one of the possible solutions in the center. This is evidence for some cooperativity in the human stereo matching algorithm. (Reprinted by permission from B. Julesz and J. J. Chang, "Interaction between pools of binocular disparity detectors tuned to different disparities", *Biol. Cybernetics 22,* 1976, 107–120, figs. 1, 2.)

information from the border could pull the matching going on in the interior one way or the other. This suggests that our visual systems use the second of the two alternatives outlined above.

Panum's fusional area

By using the second of the above schemes, matching may be assigned correctly for a disparity range of w. The precision of the disparity values thus obtained should be quite high and a roughly constant proportion of w (which can be estimated from stereoacuity results to be about $w/20$). For Wilson's foveal channels, this means 3′ disparity with a resolution of 10″ for the smallest and perhaps up to 20′ for the largest with a resolution of 1′. At 4° eccentricity, the range is 5.3′ to about 34′.

Under these assumptions, the predicted values apparently correspond quite well to available measures of the fusional limits without eye movements. Mitchell (1966) used small, flashed line targets and found, in keeping with earlier studies, that the maximum amount of convergent or divergent disparity without diplopia is about 10′–14′ in the fovea and about 30′ at 5° eccentricity. The extent of the so-called Panum's fusional area is therefore twice this.

Under stabilized image conditions, Fender and Julesz (1967) found that fusion occurred between line targets (13′ by 1° high) at a maximum disparity of 40′. This value probably represents the whole extent of Panum's fusional area. Using the same technique on a random-dot stereogram, Fender and Julez arrived at a figure of 14′ (6′ displacement and 8′ disparity within the stereogram). Since the dot size was only 2′, we expect more energy in the high-frequency channels than in the low, which would tend to reduce the fusional area. Julesz and Chang (1976), using a 6′ dot size over a visual angle of 5°, routinely achieved fusion up to 18′ disparity. Taking all factors into account, these figures seem to be consistent with our expectations.

A critical prediction of the theory is that the maximum fusible disparity should scale with the spatial frequency of the stimulus, since the lower spatial frequencies will be detected by only the larger channels. There are already hints that this might be so (Felton, Richards, and Smith, 1972).

Impressions of depth from larger disparities

We have assumed that Panum's area corresponds to pure stereoscopic fusion. One still gains some impression of depth outside this disparity range, however, although this impression does not accurately reflect the disparity that is present. There are two interesting cases to examine.

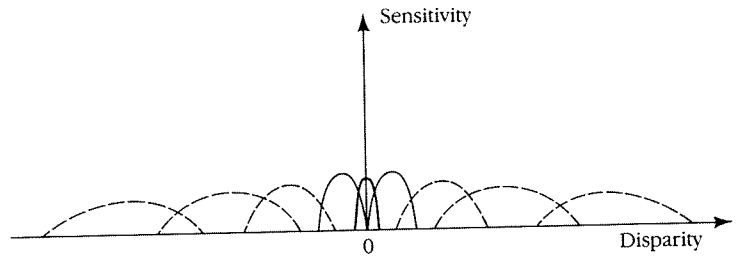


Figure 3–21. In addition to the three narrow disparity pools (solid lines) required by the matching algorithm, sets of outlying disparity detectors may exist, as illustrated here by the dotted lines. Their function would be to estimate whether the fusion plane lies convergent or divergent, so that vergence eye movements can be initiated in the appropriate direction.

The first is diplopia, in which one sees double but still senses depth. The stereo matching algorithms I described above are designed to work when the images are complex. When they are very sparse, there is no real trouble with matching them, because there are no false targets to be avoided. If, for example, there are no possible matches at all in the range w, detectors operating outside this range, possibly sensitive to any match over a broad interval may be consulted. The idea would be that if some indication of the sign of the disparity was available, this would be enough to initiate vergence eye movements in the correct direction so as to bring the images into the fusible range.

There is another way in which such detectors could be used. As we saw in the subsection on the computational theory of stereopsis, if the image contains matchable features with a density of ν, the density of matches at the correct disparity is ν, whereas at incorrect disparities it is only ν^2. If there is a range of disparity detectors and we want only to extract the sign of the disparity where the correct matches lie, we could conceive of a scheme in which the total number of convergent matches—false targets and all—is summed and compared with the corresponding number of divergent matches. We can think of various ways of doing this. For example, adding up over the whole convergent and divergent disparity ranges simultaneously would be the simplest, but just conceivably the range of summation might be gradually extended until a significant difference is obtained. In any case, in a biologically plausible implementation of the kind illustrated in Figure 3–21, we would expect the number of detectors to decrease as the disparity increases. This would, for statistical reasons,

produce a psychophysical interdependence between the disparity in an unfused stereogram and the area needed to detect the disparity's sign.

Interestingly, Tyler and Julesz (1980) have reported that such a relationship holds for dynamic random-dot stereograms. These are stereograms that change in their patterns but not necessarily their disparity at rates around 30 frames per second. The sign of disparity can be detected, but not, for example, the shape of the disparate pattern at up to several degrees of disparity. Their finding, that detection ability depends on the square root of the area, $\sqrt{A}$, could be explained by the kind of scheme I suggested, in which the density of disparity detectors falls off with the inverse of disparity, $1/d$. This produces a $\sqrt{A}$ dependence (Marr and Poggio, 1980). Of course, there are other possible explanations of these findings, based on things like motion cues or possible nonlinear, temporal summation at the receptor level between successive frames.

Finally, we shall return to what I still regard as something of a puzzle about stereopsis; namely, Why should one use zero-crossings as the input representation for the matching process? Why not wait and use the raw and full primal sketches, using a scheme that has the same general characteristics but which replaces the low-spatial-frequency zero-crossings by the rough, large-scale primitives in the primal sketch, and the high-spatial-frequency zero-crossings by the raw primal sketch. The findings of Julesz and Miller (1975), for example, about the independent fusion of different spatial frequencies seem the best evidence for the pure zero-crossings approach, but they can probably be explained by this other scheme. The reason is that since, in Julesz and Miller's patterns, like the one reproduced in Figure 3–10, information from different regions of the spatial-frequency spectrum does not come from a common source, the spatial coincidence assumption will be violated, and so independent descriptions for each will appear in the primal sketch.

In addition to this, we have the evidence of Kidd, Frisby, and Mayhew (1979), which I described in Chapter 2, that some texture boundaries can drive vergence movements in stereopsis. This is definite evidence that some of the later primal sketch descriptions are used for stereo vision.

On the other hand, however, the same group found that, in some sense, stereo fusion can preempt and therefore probably precede texture vision discriminations (Frisby and Mayhew, 1979, figs. 1b, c, and d). Figure 3–22 shows some examples. When viewed monocularly, the differently textured regions are clearly visible. When viewed binocularly, however, they disappear. This is slight but not incontrovertible evidence for the zero-crossings approach.

My own view is that some combination of the two is in fact used, although it is based mainly on the zero-crossings approach. The decisive [illegible] since they are the first

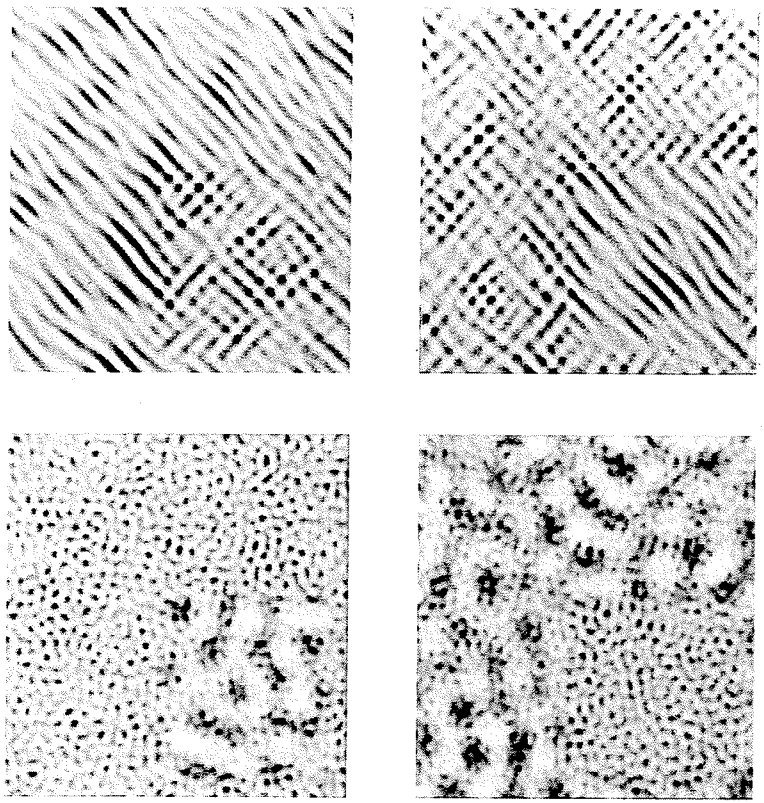


Figure 3–22. The texture differences, which are clearly visible monocularly, disappear when the two images are fused stereoscopically. (Reprinted by permission from J. P. Frisby and J. E. W. Mayhew, "Does visual texture discrimination precede binocular fusion?" *Perception 8,* 1979, 153–156, figs. 1, 2.) Figure 3–22 continues on next page.

things to be obtained, and precision, since they can be very accurately localized. The theoretical reservations one has about them—that they are only approximately and not strictly tied to physical changes—are not very strong points because zero-crossings are pretty physical (much more so than gray levels, for example). In fact, we know that they are sufficiently physical, because the computer implementation of the zero-crossings theory works well on natural images (Grimson and Marr, 1979; Grimson, 1981).

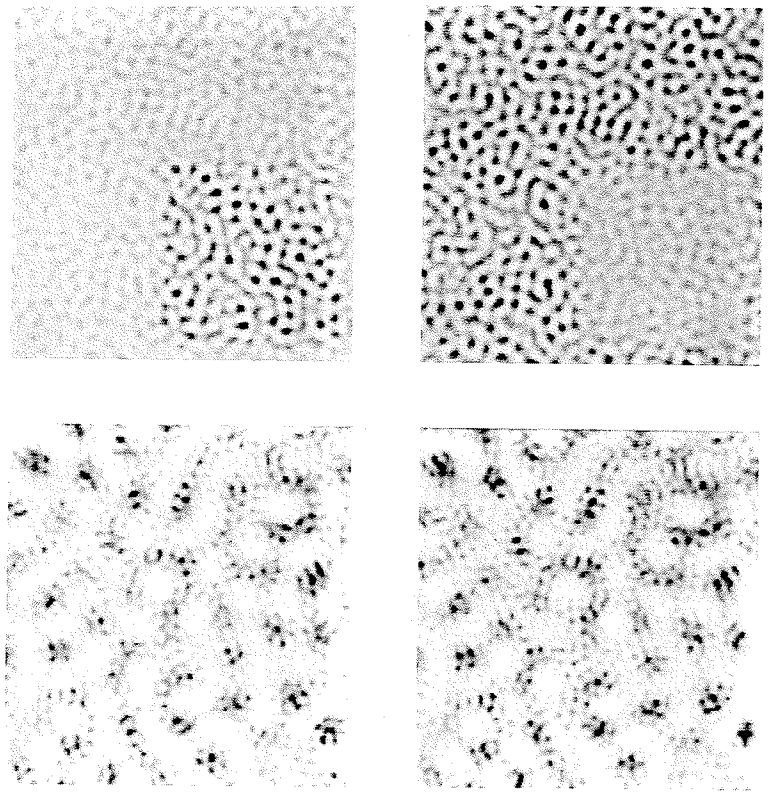


Figure 3–22 (continued).

Have we solved the right problem?

The basic issue that faces the designer of a stereo matching algorithm is, What are the difficult problems and what are the easy ones? A neurophysiologist could, with some justification, object that the matter of stereo fusion is not very difficult at all, and that the really remarkable thing about our stereoscopic vision is its precision, which can be as great as 2″ of arc for a 75% success rate, that is, roughly one-twelfth the diameter of a foveal cone (Berry, 1948). The false target problem, he might argue, is not difficult if we match special features that occur only rarely.

I disagree with these arguments for the following reasons. In stereo matching, the critical questions are, of course, How rare is rare, and how is rarity related to the disparity range that is consulted? The psychophysical evidence is that the features that can be matched are low level and not very specific for contrast or for orientation. Thus, random-dot stereograms must create false targets, yet we can fuse them. The theory of our second algorithm is in fact largely devoted to precisely the question of how rare is rare, and it is specifically tied to the suggestion that the input representation for stereo fusion is the roughly oriented, signed zero-crossings.

Stereoscopic acuity, on the other hand, although quite remarkable, is an engineering and not a theoretical problem. It occurs at the third of our three levels, the level of implementation mechanisms, because the only question that it raises is, How accurately are the zero-crossings localized? That they can be located to 2″ is remarkable but easy to incorporate in a computer program, for example. We simply have to calculate quite precisely the positions at which the $\nabla^2 G$ convolution passes through zero. No issue of principle is raised here. That neural hardware can do this calculation is remarkable, and it probably means that very many small cells are at some stage used to find and locate these positions, but this calculation is not a theoretical problem in the same way that stereo fusion is. I shall return to the problem of acuity in the neural implementation subsection.

Vergence movements and the 2½-D sketch

According to the second stereo matching theory, once zero-crossing matches have been obtained between $\nabla^2 G$ filtered images using masks of a given size, they are represented in a temporary buffer. These matches also control vergence movements of the two eyes, thus allowing information from large masks to bring small masks into their range of correspondence. The control of vergence could be direct, deriving from the matching neurons themselves, or it could be indirect, routed through the memory buffer or (most likely) through both paths.

The reasons for postulating the existence of a memory are of two kinds, those arising from general considerations about early visual processing and those concerning the specific problem of stereopsis. A memory like the 2½-D sketch (see Figure 3–12) is computationally desirable on general grounds, because it provides a representation in which information obtained from several early visual processes can be combined (see Chapter 4). The reason associated specifically with stereopsis is the computational simplicity of the matching process, which requires a buffer in which to preserve its results as disjunctive eye movements change the plane of fixation and as objects move in the visual field. In this way, the 2½-D sketch becomes the place where global stereopsis is actually achieved, combining the matches provided independently by the different channels, making the resulting disparity map available to other visual processes, and forming the

representational basis for the subjective impression that we obtain from stereograms of visible geometrical surfaces.

I shall discuss the 2½-D sketch in detail in the next chapter; here I shall make a few brief remarks about the control of eye movements during stereo vision.

Disjunctive eye movements, which change the plane of fixation of the two eyes, are independent of conjunctive eye movements (Rashbass and Westheimer, 1961b), are smooth rather than saccadic, have a reaction time of about 160 ms, and follow a rather simple control strategy. The (asymptotic) velocity of eye vergence depends linearly on the amplitude of the disparity, the constant of proportionality being about 8°/s per degree of disparity (Rashbass and Westheimer, 1961a). Vergence movements are accurate to within about 2′ (Riggs and Niehl, 1960), and voluntary binocular saccades preserve vergence nearly exactly (Williams and Fender, 1977). Furthermore, Westheimer and Mitchell (1969) found that tachistoscopic presentation of disparate images led to the initiation of an appropriate vergence movement but not to its completion. These data strongly suggest that vergence movements are not ballistic but rather are continuously controlled.

The hypothesis is that vergence movements are controlled by matches obtained through the various channels by means of the mechanisms described earlier that can give a rough sense of depth and by means of some higher types of boundary acting either directly or indirectly through the 2½-D sketch. This hypothesis is consistent with the observed strategy and precision of vergence control, and it also accounts for the finding that perception times depend to some extent on the distribution of disparities in a scene (Frisby and Clatworthy, 1975; Saye and Frisby, 1975). A stereogram of a spiral staircase ascending toward the viewer does not produce the long perception times associated with a two-planar stereogram of similar disparity range. This is to be expected within the framework of the theory, because scenes like a spiral staircase, in which disparity changes smoothly, allow vergence movements to scan a large disparity range under the continuous control of the outputs of even the smallest masks. On the other hand, two-planar stereograms with the same disparity range require a large vergence shift but provide no accurate information for its continuous control.

The long perception times for such stereograms may therefore be explained in terms of a random search strategy by the vergence control system. In other words, vergence movement control is a simple, continuous, closed-loop process that is usually inaccessible from higher levels. The stereograms in Figure 3–23 will enable the reader to see for himself that this is at any rate subjectively true.

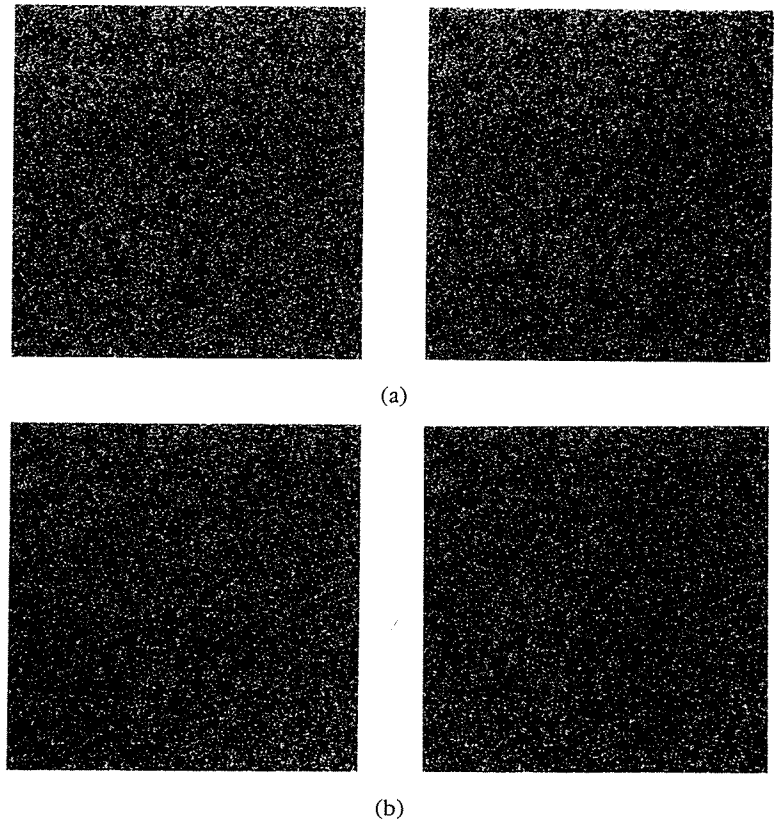


Figure 3–23. These two stereograms have about the same disparity range, but in (a) disparity varies continuously while (b) consists of just two disparity planes. It takes longer to see this second one, presumably because the vergence control system has less information about how to cover the disparity range.

Interestingly, there is some evidence that an observer can learn to make an efficient series of vergence movements (Frisby and Clatworthy, 1975). However, this learning effect seems to be confined to the type of information used by the closed-loop vergence control system. A priori, verbal or high-level cues about the stereogram are ineffective, as, incidentally, they seem to be at all levels of processing up to and including the 2½-D sketch.

A complete neural implementation of the second stereo matching algorithm just described has not yet been formulated. One reason is that such a formulation was not worth the considerable work involved until we were reasonably certain from implementation studies and psychophysics that the algorithm works and is roughly correct. However, the first steps have been taken in the analysis of possible neural mechanisms underlying the computation of $\nabla^2 G$ and of the detection of zero-crossings (Marr and Hildreth, 1980; Marr and Ullman, 1979).

The problem of the binocular combination is still an open question—the first of many that we shall be able to formulate. We can, however, allow ourselves some preliminary remarks on the topic. First, disparity sensitivity should not arise before zero-crossing detection. Hence, if the simple cells of area 17 (the striate cortex), which are the first cortical cells in the visual pathway, are disparity sensitive, as seems likely in the cat (Barlow, Blakemore, and Pettigrew, 1967), then they must also detect zero-crossings.

This can occur in several ways, and Figure 3–24 shows two examples. In the first, two independent zero-crossing detections are proposed to take place in the dendrites, each much as shown in Figure 2–18 and relying on local synaptic mechanisms of the Poggio and Torre (1978) type.

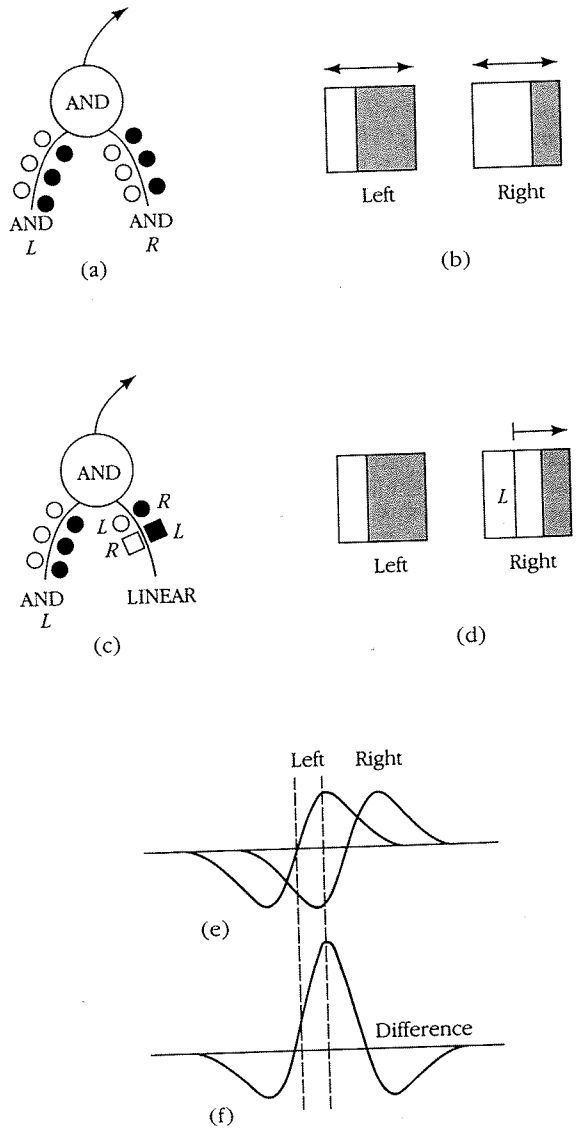


Figure 3–24. (opposite) Two possible neural implementations of disparity detectors. In the first, the cell (a) detects zero-crossings of a given sign independently in two dendrites, one driven by each eye. It then combines the result through an AND gate, which has the effect of making the cell fire whenever an appropriate zero-crossing simultaneously appears anywhere in the cell's left-eye and right-eye receptive fields. This is illustrated in (b). However, such a scheme can provide only a rather rough type of disparity detection; it has the disadvantage, for example, that the range of disparities to which it is sensitive varies with the position of the zero-crossing in the left-eye receptive field. Circles represent excitatory inputs; squares, inhibitory inputs. Open synapses (circles and squares) represent on-center inputs; filled ones, off-center inputs. *L* and *R* denote left-eye and right-eye inputs.

The second scheme does not suffer from this disadvantage, since it accurately signals the sign of the disparity, but it operates on only a small disparity range. A zero-crossing is detected in the left image by the AND dendrite of the cell in (c), and the sign of the disparity is assessed by examining the sign, at the zero-crossing, of the difference—computed by a linear process—in the values of the $\nabla^2 G$ convolution for the left and right eyes. This yields a detector of disparity sign that is independent of the position of the left-eye zero-crossing, at least for a small range (d). As illustrated in (e) and (f), if the difference at the zero-crossing is positive, the disparity has one sign; if it is negative, the disparity has the opposite sign.

Such a mechanism does have disadvantages. First, it is not very sensitively tuned to disparity, because the zero-crossings in each eye are located with an accuracy of not much better than $w_{1\text{-}D}$. Second, the range of disparities to which the mechanism responds depends upon the exact position of the zero-crossing in the left eye, because the range of positions in the right eye is also fixed by the geometry of the connections there.

The second model in Figure 3–24 shows another possibility. The cell is left-eye dominant, being driven by a zero-crossing from the left eye. However, it is gated by the difference between the left- and right-eye convolutions at the zero-crossing. If this difference is negative, the disparity will usually be of one sign, and if it is positive, the disparity will usually be of the other sign, as explained in Figure 3–24. For an edge that goes from light to dark as one moves from left to right in the visual field, a negative difference corresponds to divergent (near) disparities. This mechanism removes some of the imprecision associated with the first mechanism, since it measures quite directly whether the right image's zero-crossing (of fixed sign) is to the left or to the right of the left image's. It has its disadvantages, however, since for too closely occurring zero-crossings or for very different contrasts in the two eyes, it can be unreliable.

Unfortunately, the technical problems associated with the neurophysiology of stereopsis are considerable, and rather few quantitative data are currently available—certainly too few to enable us to rule out either or both of the mechanisms of Figure 3–24. Since Barlow, Blakemore, and Pettigrew's (1967) original paper, relatively few examples of disparity tuning curves have been published. Recently, however, Poggio and Fischer (1978) and von der Heydt and others (1978) have published properly controlled disparity curves for the monkey and cat, respectively. On the whole, these studies favor the idea that disparity detectors are organized into three pools—convergent, near zero, and divergent—and recently Clarke, Donaldson, and Whitteridge (1976) have found that, in the sheep, these detectors are organized into columns, as Hubel and Wiesel (1970) suggested they might be in area 18 of the macaque. However, the size of the disparities involved are surprisingly large—7° in the sheep and up to a degree or even several in the monkey. The precise role of these detectors in stereopsis is therefore not yet clear.

Curiously enough, even the owl, which diverged from the monkey probably before stereopsis evolved, appears to use an algorithm similar to the monkey's. Pettigrew and Konishi (1976) have found that although the anatomical organization of the owl's wulst is quite different from that of the monkey's visual pathway, the physiological responses of the cells are very similar. The owl, however, is unable to move its eyes very much, so at first it might be thought to be deprived of the ability to make the vergence

movements that are so essential for this approach to stereopsis. Nature, though, has found a way—the owl's horopter is sloped, passing through its feet at the bottom of the visual field and extending to infinity roughly straight ahead. The owl can therefore attain the effect of vergence eye movements, together with the simultaneous impression of a profound and grave wisdom, by the gentle but deliberate nodding of its head.

Finally, there is the problem of stereo acuity, which, like all human hyperacuity abilities, requires an underlying mechanism that is able to localize small, isolated features in an image to within about 5″ of arc for an average subject (Westheimer and McKee, 1977). Crick, Marr, and Poggio (1980) discussed the neurophysiological implications of these findings and suggested that one possible solution might be based on the high-resolution spatial reconstruction of the $\nabla^2 G$-filtered image as it enters the visual cortex from the optic radiations. Barlow (1979) made the suggestion first, and we amended it slightly, saying that the reconstruction need not be completely accurate. It will suffice to reconstruct accurately only those parts of the signal lying around the zero-crossings.

The natural candidate for performing the reconstruction is the granule cell population of layer IVCβ in area 17. Worst-case estimates suggest that, for each type (on center and off center) and each eye, there is easily one granule cell for every 5″ of arc for the smallest channel. David Hubel furthermore reports that these cells are all center–surround, so far indistinguishable from geniculate fibers, and that their spatial arrangement is very precisely retinotopic—nearby cells correspond to nearby points on the retina. These are all properties that we would expect of cells engaged in reconstruction. It would therefore be of great interest to know whether their responses differ physiologically in any way from those of the lateral geniculate fibers, for example in their spatial or particularly in their temporal characteristics.

Computing Distance and Surface Orientation from Disparity

Computational theory

Distance from the viewer to the surface

Suppose a point P lies at distance l from the viewer's left eye L and at angle ω to his forward line of sight, as illustrated in Figure 3–25. Let the distance between the viewer's eyes be δ_T; then, because the line of sight to P does not lie directly ahead, the effective distance between the two eyes

is only $\alpha = \delta_T \cos \omega$. Writing $\beta = \delta_T \sin \omega$, we see from the figure that ϕ, the angle between the lines of sight from the two eyes, is given by

$$\tan \phi = \frac{\alpha}{(l + \beta)} = \frac{\delta}{l}$$

For small values of ϕ, we can write

$$\phi \cong \frac{\delta}{l} = \frac{\alpha}{l + \beta}$$

Now take two points P and P' along the same line of sight from the left eye, with P at distance l and P' at distance l' as in Figures 3–25(a) and (b). It follows that the disparity $\Delta\phi$ between P and P' is $\phi' - \phi$. Hence, if we let

$$q = \frac{l' + \beta}{l + \beta}$$

then

$$\Delta\phi \cong \left(\frac{1}{q} - 1\right) \frac{\alpha}{l + \beta} = (1 - q)\frac{\delta}{l}$$

We can rewrite this as

$$(1 - q) \cong \left(\frac{l}{\delta}\right) \Delta\phi$$

In other words, the fractional change in distance for a given disparity depends upon the distance away. This fact can be important for depth-judging experiments and, as we shall shortly see, for the perception of surface orientation, because it shows that if the human visual system does its job properly, the proportional change in perceived depth obtained for a given disparity should depend on l, that is, on what the observer happens to think the current true depth is.

Surface orientation from disparity change

The trigonometry of the recovery of surface orientation is rather tedious. However, the resulting formulas are interesting, so I shall discuss them here. We need to consider two cases, one in which the surface slopes

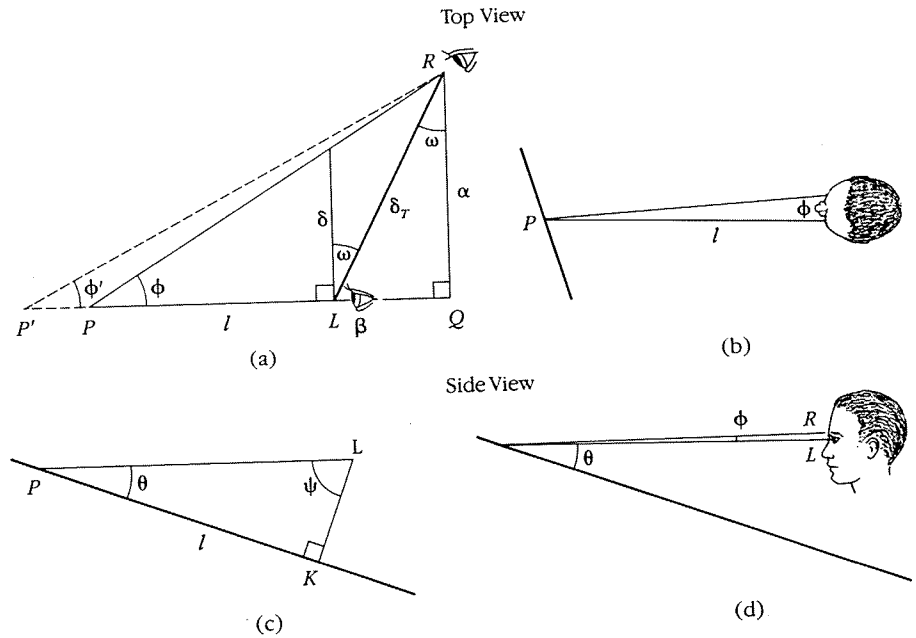


Figure 3–25. The trigonometry of recovering depth from disparity. (a) shows a top view of the geometry of the two eyes looking at a point P distance l from the left eye, as illustrated in (b). The line of sight is not necessarily perpendicular to the line joining the two eyes L and R, and the difference is described by the angle ω as illustrated. The true interocular distance is δ_T, and the effective interocular distance for this line of sight is $\delta_T \cos \omega$. The angle between the lines of sight from the two eyes is ϕ, and it is the differences in the values of ϕ for different points P' that are normally called disparities. The lengths $\alpha = \delta_T \cos \omega$ and $\beta = \delta_T \sin \omega$ are useful geometrical quantities.

(c) shows a side view of the same situation, illustrated in (d). The point P is shown lying on a plane that slopes vertically, and its slope at P is described by the angle θ. Only the left eye L is shown in this diagram, and again the distance l refers to the distance from the left eye. In order to recover surface orientation, it is necessary to recover the angle θ.

in the horizontal direction, as in Figures 3–25(a) and (b), and one in which it slopes away in the vertical direction, as in Figures 3–25(c) and (d). These situations differ because our eyes are positioned horizontally, not vertically. In both cases, we need the formulas that relate surface orientation, which I denote by θ, to the rate of change of disparity ϕ with visual angle ψ, which I write $\partial\phi/\partial\psi$. The formulas are as follows:

For surfaces changing in depth in the vertical direction:

$$\frac{\partial\phi}{\partial\psi_V} = \frac{-\alpha l \cot\theta}{\alpha^2 + (\beta + l)^2}$$

For surfaces changing in depth in the horizontal direction (the formula is perforce more complicated):

$$\frac{\partial\phi}{\partial\psi_H} = \frac{\alpha^2 + \beta(\beta + l) - \alpha l \cot\theta}{\alpha^2 + (\beta + l)^2}$$

There are two points to be noted about these formulas. First, like estimates of fractional depth, they depend on the viewing distance l, roughly as $1/l$. Hence, if the brain is doing its task, a given rate of change of disparity should be perceived as an increasingly steep surface as its distance away is increased. The reader can see this by looking at the stereogram in Figure 3–26 from different distances. Disparity and viewing angle change together, so $\partial\phi/\partial\psi$ is constant for all viewing distances. Hence, the surface should appear to steepen as one moves the stereogram further away, and it does. This also shows, incidentally, that the brain has a pretty good idea of where the stereogram actually is and uses this information.

Second, when the horizontal rate of change of disparity $\partial\phi/\partial\psi_H$ reaches 1, the line of sight from the other eye must fall directly along or in front of the actual physical surface. The viewer then sees a discontinuity

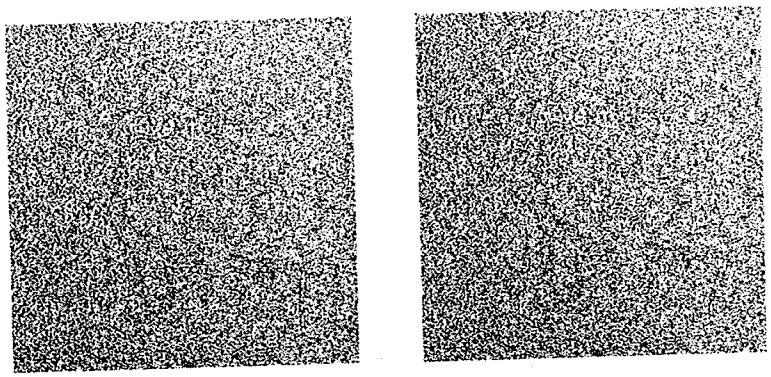


Figure 3–26. Notice that if the viewing distance from this stereogram changes, the perceived surface orientations change. This is to be expected if the visual system

in depth from the second eye. This can be checked by putting $\theta = -\phi$ into the horizontal disparity-change formula; then $\partial\phi/\partial\psi_H = 1$. In this situation, all the change in viewing angle from the first eye is a change in disparity, so $\partial\phi/\partial\psi_H$ remains equal to 1 until the other eye starts seeing the surface again. This fact can be used to help us to find discontinuities in viewing distance from stereopsis.

Algorithm and implementation

Nothing is known about how these formulas are implemented, although the example of Figure 3–26 suggests that approximations to them are and that the approximations may be quite accurate. It is perhaps worth emphasizing that the effects I pointed out, of a dependence of perceived depth and surface orientation on viewing distance and direction, are wholly to be expected and are not some strange psychophysical phenomena that need complex explanations.

3.4 DIRECTIONAL SELECTIVITY

Introduction to Visual Motion

Motion pervades the visual world, a circumstance that has not failed to influence substantially the processes of evolution. The study of visual motion is the study of how information about only the organization of movement in an image can be used to make inferences about the structure and movement of the outside world. Again there are two basic parts to the problem: How are the raw measurements of the changes produced by motion made, and is this information used? Neither is at all easy to solve, and perhaps because the first is so difficult, the second is to some extent a study of the minimum information necessary from the first part in order for subsequent computations to deliver any sort of useful results.

The psychophysical study of visual motion is old. Most people would probably trace its origins to members of the Gestalt movement (Wertheimer, 1923; Koffka, 1935), who, like their followers Gibson and Julesz (Gibson et al., 1959; Julesz, 1971, ch. 4), were interested in the effects of motion on the separation of figure and ground and on eye movements. Miles (1931) and Wallach and O'Connell (1953) introduced the problem of determining three-dimensional structure from motion, a problem dealt with at length in the recent and remarkable book by Shimon Ullman (1979b). Gibson (1966) was interested in the problem of optical flow, a problem that has only recently received the mathematical attention it deserves (Longuet-Higgins and Prazdny, 1980).

Table 3–1. Determinants of apparent motion found with two perceptual criteria.

Criterion of segregation in random-dot display	Criterion of smooth apparent motion for isolated element
Spatial displacement must be 15′ arc or less (Braddick, 1974).	Spatial displacement may be many degrees (e.g., Neuhaus, 1930; Zeeman and Roelofs, 1953).
ISI must be less than 80–100 ms (with 100 ms stimulus exposure) (Braddick, 1973).	ISI may be at least 300 ms (e.g., Neuhaus, 1930).
Segregation abolished by bright uniform field in ISI (Braddick, 1973).	Motion perceived whether ISI is bright or dark.
Successive stimuli must be delivered to the same eye or to both eyes together (Braddick, 1974), as must bright field for effective masking (Braddick, 1973).	Successive stimuli may be delivered to the same or different eyes (Shipley, Kenney, and King, 1945).
Pattern defined by chromatic but not luminance contrast is inadequate (Ramachandran and Gregory, 1978).	Stimuli may be defined by chromatic contrast alone (Ramachandran and Gregory, 1978).

Note: ISI = interstimulus interval.

The first important psychophysical finding I wish to emphasize, however, is quite recent, and it bears upon the question of how many different motion modules or processes there are, what they do, and how rich the information is that they run on. Following Julesz's (1971, ch. 4) example, Braddick (1973, 1974) used random dots and lines to explore the psychophysical properties of apparent motion. For example, he found a number of strange differences between what happens over short times and small displacements and what happens over long times and large displacements. He concluded that there were two different processes characterized by different perceptual criteria, and which have the properties listed in Table 3–1 (from Braddick, 1979).

These properties were found in experiments of the following kind. Two patterns are used as displays, each composed of random dots or lines. Outside a central rectangle, the two patterns are uncorrelated, as illustrated in Figure 3–27. Inside the central rectangle, the dots are displaced in one pattern relative to the other, in the manner of Figure 3–28. The two patterns

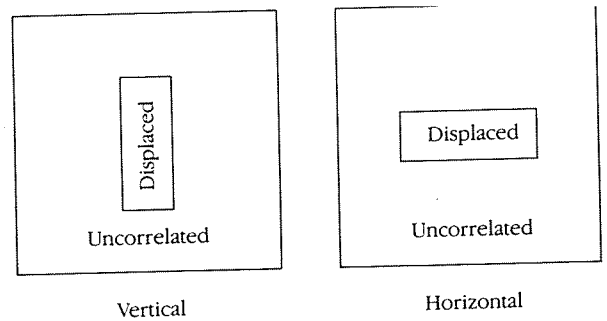


Figure 3–27. The discrimination task for Braddick's short-range phenomena. A vertical or horizontal rectangle has to be discriminated against an uncorrelated background.

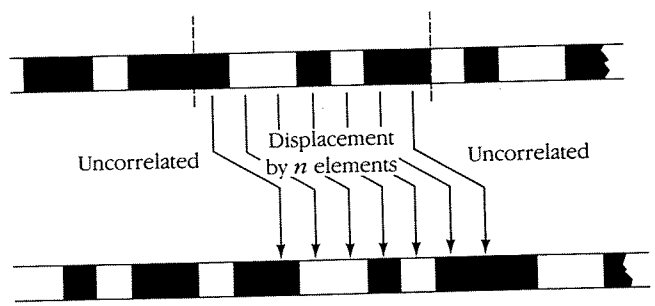


Figure 3–28. The rectangles in Figure 3–27 are created in a pair of successively presented random-dot displays by displacing a rectangular region by a few elements. The rest of the display is uncorrelated between frames.

are alternated at some rate with an interstimulus interval (ISI) during which other masking fields are sometimes shown. The question is, For what rates and displacements does the subject perceive the rectangular region well enough to say whether it is horizontal or vertical?

The second kind of experiment was like those extensively used by Ullman, in which one or a few lines are presented in frame 1, followed by the ISI, followed by a second few lines, as illustrated in Figure 3–29. Here the question is, Does the subject smoothly perceive one line mapping to another line or lines, and if so, how does the mapping go? Ullman's (1978) experiments have warned us to be wary of smoothness, but the actual mapping itself is a reliable and useful phenomenon.

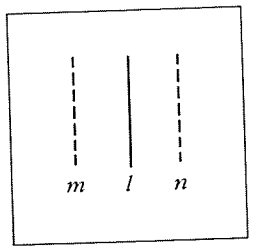


Figure 3–29. The second type of display, extensively used by Ullman, also consists of two frames, but they are much simpler than those in Figures 3–27 and 3–28. The first might consist of the line l shown here, and the second of two lines m and n. The observer is asked, Does l go to m, n, or both?

What Braddick found was that if one does various things to the two types of display—things like changing the displacement or the ISI or flashing a bright uniform field during the ISI—perceptions of the displays are very different. Conditions that easily disrupt the first experimental task do not disrupt the second. For example, to discern the rectangle successfully, the angular displacement must be small (less than 15′), the ISI short (less than 80 ms), and no masking field may intervene. Not so the second task; the angular displacement may be many degrees, the ISI may be 300 ms or more, and the masking field may be bright or dark. These and other differences are summarized in Table 3–1.

What could be the significance of these distinctions? Perhaps the key to the puzzle is that in the analysis of motion—more so, perhaps, than any other aspect of vision—time is of the essence. This is not only because moving things can be harmful, but also because, like yesterday's weather forecast, old descriptions of the state of a moving body soon become useless. On the other hand, the detail of the analysis that can be performed depends upon the richness of the information on which the analysis is based, and this in turn is bound to depend upon the length of time that is available to collect the information. In an instantaneous view, for example, everything is static, so no information about motion is available. After a 60 ms wait, information derived from observed changes may enable a much more thorough analysis, and in a third look in yet another 60 ms perhaps everything about the motion can be recovered, provided that computation is powerful enough.

Perhaps one of the most primitive types of motion analysis is the type concerned with noticing that something has changed, where the change is in the visual field, and perhaps something about the direction of movement involved, though this is arguably a more complex matter. Such analysis we have already met in our earlier discussion of the visual system of the housefly. Another case where similar mechanisms are thought to operate is in the directionally selective cells of the rabbit's retina (Barlow and

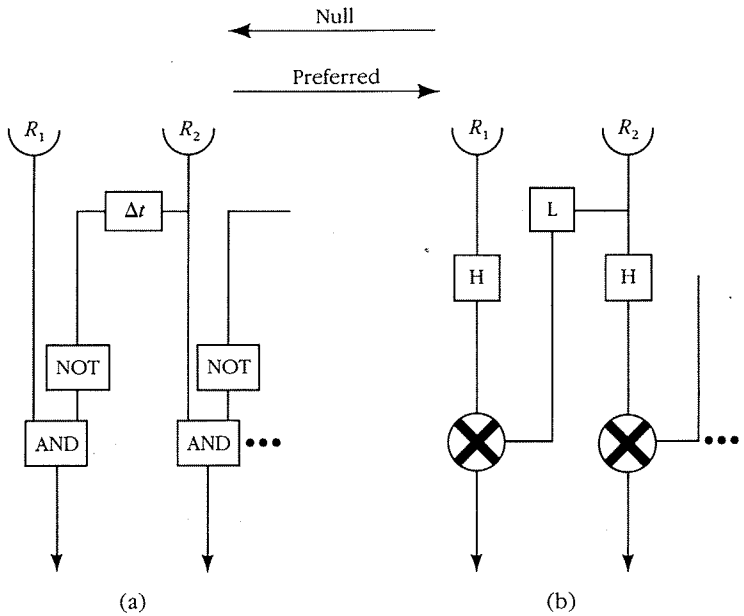


Figure 3–30. (a) Barlow and Levick's (1965) model for directional selectivity connects two detectors to an AND–NOT gate, one via a delay. Thus the network does not respond to stimuli moving with roughly the right speed in the null direction. (b) Hassenstein and Reichardt's (1956) model operates on the same principle except that the delay is replaced by a temporal low-pass filter (L). H = high-pass filter.

Levick, 1965), the frog's retina (Barlow, 1953; Maturana et al., 1960), the pigeon's retina (Maturana and Frenk, 1963), and perhaps the mammalian retinal W cells.

These mechanisms all have various things in common. They probably all operate at the earliest possible stage—that is, directly on the gray-level image intensity values—and their underlying mechanism is something equivalent to combining a time delay (or temporal low-pass filter) and an AND–NOT gate*. The basic idea is illustrated in Figure 3–30(a). Two receptors are connected to an AND–NOT gate, one directly and one through a delay. If a bright spot moves first across the right-hand receptor R_2, then

*A logical device that gives an output only when its first input is on and its second input is off.

across the other one, R_1, signals from the two will arrive at the gate roughly simultaneously, causing it to remain silent. This is called the null direction. A white spot moving the other way will cause the gate to fire.

If the intensity detectors are replaced by a center–surround operator, this difficulty goes away—we get a directionally selective bug detector or edge detector—but it still has characteristic problems. First, if a stimulus is moved very slowly in the null direction or is stopped and restarted halfway between the two receptors, the gate will give a response. Second, and again relating to the delay, the range of spatial frequencies over which the device operates reliably depends on how fast the pattern moves. To the device, a thick sinusoidal grating moving fast looks like a thin one moving slowly. Our own visual systems exhibit similar properties (for example, Kelly, 1979). To maintain reliability, we must make sure that the mechanism looks only at the appropriate portion of the range of spatiotemporal possibilities.

The reason that detectors of the type shown in Figure 3–30 fail to be reliable is a deep one. Fundamentally, they are reading a receptor in one place at one time and another in a nearby place a little later; if anything happens at one and then at the other the correct interval later, the detector implicitly assumes that the two changes are due to the same physical cause. This, in fact, is our first real introduction to the *correspondence problem of apparent motion.* The unreliability of these detectors arises for the same basic reasons that make a fast, clockwise-turning wagon wheel in a Western movie seem to be turning slowly counter-clockwise. The implicit assumption, that the nearest spoke in the next frame is the same one as in the last frame, is wrong because the wheel is turning too fast relative to the movie frame rate.

Such schemes, as I indicated, are still useful for saying where in a visual field a relative movement has occurred and for giving some information about its direction, if one is careful. However, if we also wish to analyze the shape of a moving patch, it seems more sensible to try to combine the analysis of movement with the analysis of contours (Marr and Ullman, 1979). This view, incidentally, is diametrically opposed to current physiological and psychophysical thinking, according to which the sustained and transient channels in human early vision are separated into two parallel systems, one concerned with the analysis of form or pattern, and the other with movement (Tolhurst, 1973; Kulikowski and Tolhurst, 1973; Ikeda and Wright, 1972, 1975; Movshon, Thompson, and Tolhurst, 1978). For eye movement control, of course, there is no need to combine them, but to see the shape of moving patches, it would seem sensible to do so.

We have now discussed the two types of information that can be gleaned from motion—(1) noticing a movement and finding its position in the visual field and (2) determining its two-dimensional shape. As we might have expected, neither requires very sophisticated measurements, and in principle they can both be carried out very quickly given reasonably accurate measurements. What, then, about determining three-dimensional structure? This is clearly more valuable, but intuitively we would have thought that more information from the images would be necessary.

In fact, more information is required, and the basic improvement needed is a good solution to the correspondence problem, rather than the half-baked guess at it which suffices for the simpler tasks. To recover three-dimensional structure, we need to be able to say that point A in the image at time t_1 corresponds to point B in the image at time t_2 for the equivalent of three frames in Ullman's (1979a) style of analysis, or, almost equivalently, we need the exact instantaneous positions and velocities in the image for the simpler task of analyzing the optical flow induced by the observer's movement through a rigid environment. Whether either or both of these theoretical possibilities are incorporated into the human visual system is a matter for psychophysics. As we shall see, the evidence for Ullman's scheme is strong; that for a Gibson-style analysis of optical flow is somewhat weaker, but the theory is nevertheless interesting.

This and the next section in this chapter deal with the different parts of the motion analysis problem. In this section we look first at directional selectivity from the point of view of using it to separate figure from ground and recovering the two-dimensional shape of the figure. We shall then explore Ullman's theory of the interpretation of three-dimensional shape from visual motion in Section 3.5, and shall briefly discuss the problem of optical flow.

Computational theory

The theory of directional selectivity is the theory of how to use partial information about motion—specifically, only its direction defined to within 180°—in order to discern the two-dimensional shapes of regions in the visual field based on their relative movement.

The background to this problem from a computational point of view comes from asking, How much of this information can one gain from motion without solving the full correspondence problem—that is, without being provided with the full instantaneous position and velocity field for the whole image? The motivation for studying what direction alone can tell us comes from something that we call the *aperture problem,* illustrated in Figure 3–31. If a straight edge is moving across the image in direction b, as indicated by the arrow in Figure 3–31, this fact cannot be discerned by local measurements alone. As the figure shows, the only motion that can

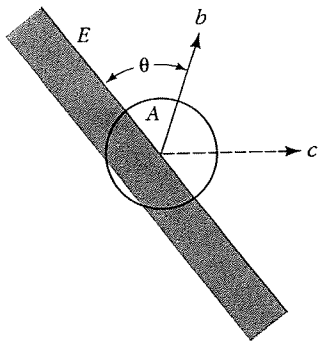


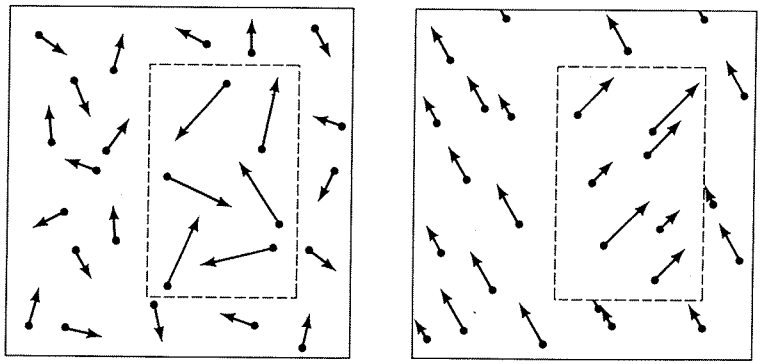
Figure 3–31. The aperture problem. If the motion of an oriented element is detected by a unit that is small compared with the size of the moving element, the only information that can be extracted is the component of the motion perpendicular to the local orientation of the element. For example, looking at the moving edge *E* through a small aperture *A*, it is impossible to determine whether the actual motion is in the direction of *b* or of *c*.

be detected directly through a small aperture placed over the edge is motion at right angles to that edge—just one bit of information, indicating whether it is moving forward or backward. Of course, if there is only a point or blob or a termination of some recognizable kind, more information can be recovered. And if one somehow knows θ, the angle between the edge and the direction of motion *b*, then the speed *s* can be recovered by measuring the component $s \sin \theta$ perpendicular to the edge. But the very simple case in which just the sign is available has at least a theoretical interest.

Various experiments suggest that this simple case is also of interest for understanding one of the visual system's ways of analyzing motion. The experimental situation is like that used by Braddick (1973, 1974), and the stimuli are shown in Figure 3–32. These experiments fall into the first of his two classes, being concerned with short-range, short-term phenomena.

In Figure 3–32(a), the individual dot speeds in the central square are all constant at twice the dot speeds in the surround, but the directions of movement are all random. The central square proves invisible, so we cannot use only speed of movement to separate the patches. Julesz (1971, ch. 4) described a similar effect. In Figure 3–32(b), the surround moves randomly, while the center dots all move in the same direction but with different speeds, spanning a factor of 4. The square can be seen clearly, and where the neighboring speeds are very different, the dots appear to have some relative movement as well.

The remarks about the aperture problem tell us what we want to measure and why we want to measure it. These psychophysical experiments suggest that the visual system uses information about direction alone to help carve up the visual field. We therefore explored algorithms for



(a) (b)

Figure 3–32. Two experiments showing that Braddick's (1979) short-range system uses only limited information to decompose the image. In (a), the speeds in the central rectangle and in the surround are different and uniform, but the directions of motion are random. Discrimination is not possible. In (b), the directions in the central rectangle are the same but the speeds differ. Discrimination is easy.

quickly detecting the sign of movement direction at the level of local edge segments or their precursors. The earliest stage at which this could be carried out is at the level of zero-crossing segments, and as we shall later see, the physiological data support this possibility.

An algorithm

To construct a directionally selective zero-crossing detector, we must somehow determine the direction of movement of an oriented zero-crossing segment of the type defined in Chapter 2. There we saw that a zero-crossing segment is defined as a locally oriented segment of the zero values of the convolution $\nabla^2 G * I$. A cross section of this convolution appears in Figure 3–33 for the image intensity profile illustrated there.

There are several ways of building a directionally selective unit from this, one of which is to use two zero-crossing detectors as the inputs to a device like Barlow and Levick's (1965). As we have already seen, however, such devices suffer from the stop–restart false response in the null direction, and directionally selective cortical simple cells are known not to do this (Goodwin, Henry, and Bishop, 1975). Marr and Ullman (1979) therefore suggested the following algorithm:

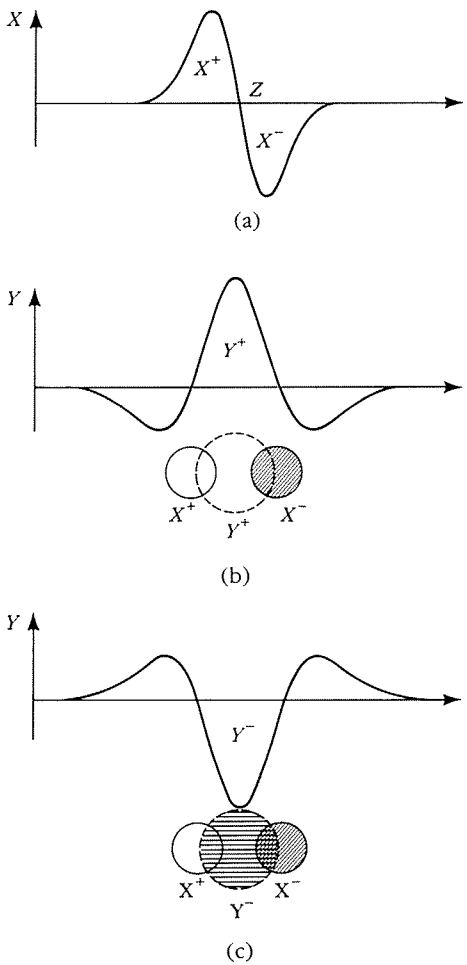


Figure 3–33. The value of $X = \nabla^2 G * I$ and of $Y = \partial/\partial t(\nabla^2 G * I)$ in the vicinity of an isolated intensity edge. (a) The X signal as a function of distance. The zero-crossing Z in the signal corresponds to the position of the edge. (b) The spatial distribution of the Y signal when the edge is moving to the right, and (c) when it is moving to the left. Motion of the zero-crossing to the right can be detected by the simultaneous activity of $X^+Y^+X^-$ in the arrangement shown in (b). Motion of the zero-crossing to the left can be detected by the $X^+Y^-X^-$ unit in (c).

Step 1. Measure the time derivative $\partial/\partial t(\nabla^2 G * I)$.

Step 2. If this is positive at Z, the zero-crossing is moving to the right; if it is negative, it is moving to the left. If the edge has opposite contrast, the directions are reversed.

The truth of these statements can be seen from Figures 3–33(b) and (c), which plots $\partial/\partial t(\nabla^2 G * I)$, the time derivative of Figure 3–33(a), for the two cases of movement to the right and to the left, respectively. The sign of the time derivative is constant over the whole width w_{1-D} between the peaks of the original convolution $\nabla^2 G * I$, so the algorithm is robust.

This scheme has several positive features. (1) It requires only local measurements. (2) No time delay is involved beyond that required to compute the derivative. (3) The method can be made extremely sensitive. The lower limit to the displacement that can be detected is set by the unit's sensitivity, and the upper limit, which depends on the temporal filter, is high if the time constants are small. Hence, a single unit can be made sensitive to a wide range of speeds, and since the only really important part of the measurement of $\partial/\partial t(\nabla^2 G * I)$ is its sign, this can be exploited by making the measuring unit extremely sensitive. It does not matter if it saturates early. (4) Finally, within this range and for a sufficiently isolated edge, the unit will be completely reliable.

The critical difference between the Barlow and Levick type of scheme and this one is that this system does not have to wait until the zero-crossing has passed from the first detector to the second. It can therefore respond instantaneously, and it is sensitive to very small displacements. In addition, unlike systems based on a pair of detectors, it does not have to "guess" that the zero-crossing exciting the left-hand detector now is the same one that excited the right-hand detector a short time ago; and so, at the price of delivering less information, it avoids the difficulties inherent in the full correspondence problem.

Neural implementation

I would not, of course, have suggested this scheme without an idea of how it might be implemented. We have already seen that the detection of zero-crossing segments (Figure 2–18) rests on the idea that the lateral geniculate X cells carry the positive and negative parts of $\nabla^2 G$ via on-center and off-center cells, respectively. Finding a zero-crossing is simply a matter of connecting the on- and off-center X cells via a logical AND gate.

But how to measure the time derivative?—here is an interesting and fascinating point. The psychophysical studies of the transient channels and

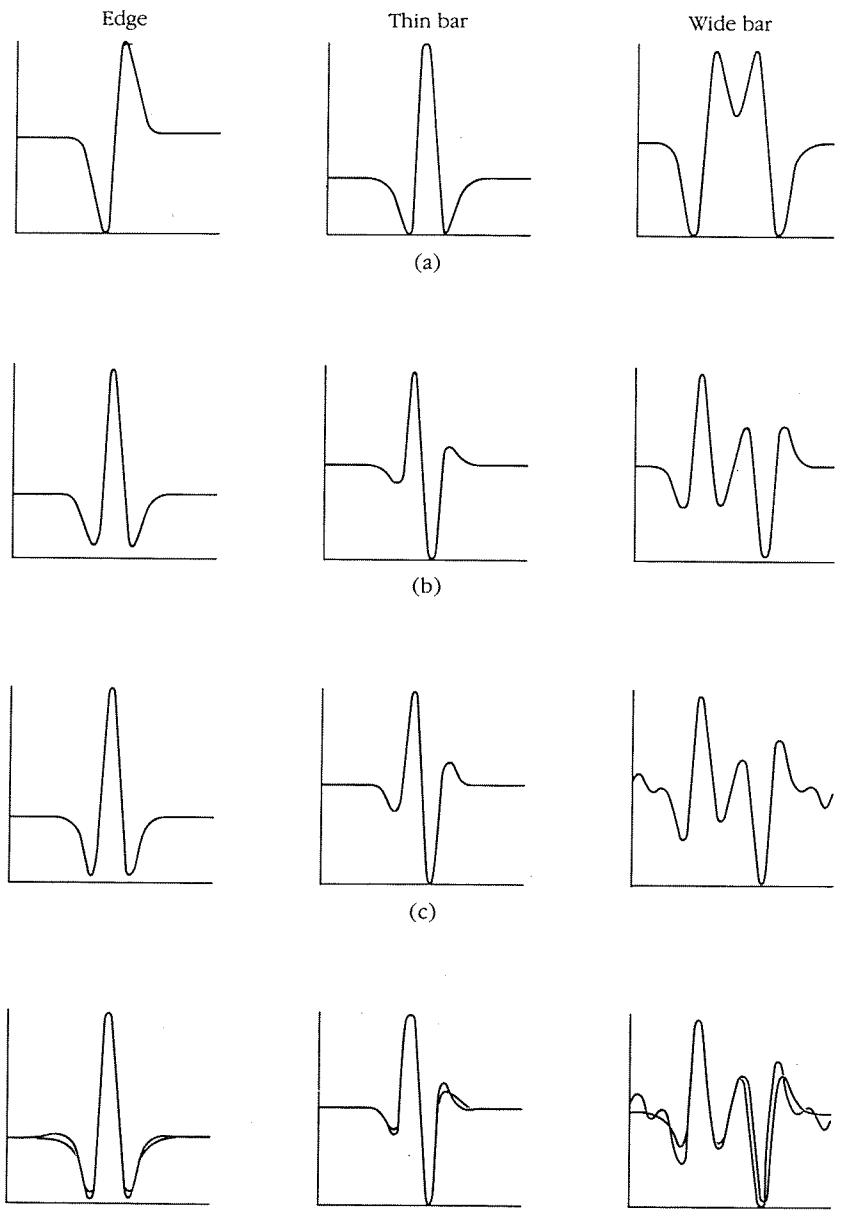
the neurophysiological recordings of the Y cells, to which the transient channels are thought to correspond, essentially demonstrate that these channels measure this time derivative, $\partial/\partial t(\nabla^2 G * I)$! Interestingly, so far as we are aware, the behavior of these channels has never been formulated as a time derivative, presumably because no one ever thought that such a thing might be a useful function so early in the visual pathway.

Let us look at the evidence a little more closely. Ideally, to obtain a time derivative, we subtract from the current value of a signal its value an infinitesimal time ago. In practice, these measurements must be taken over finite intervals of time. Hence, the impulse response of the device in the time domain should be composed of a positive phase followed by a phase of a similar shape but opposite sign. In the frequency domain, the power spectrum should be roughly linear in frequency over the range in which the device is to operate.

A temporal filter composed of about a 60-ms positive phase followed by a negative phase was explicitly suggested by Watson and Nachmias (1977) and further supported by Tolhurst (1975), Breitmeyer and Ganz (1977), and Legge (1978). The negative phase may be somewhat longer than the positive one, or it may be followed by damped oscillation of small amplitude (see Breitmeyer and Ganz, fig. 3) without significantly affecting the results.

In the frequency domain, the temporal modulation transfer function (MTF) measured by Wilson (1979) for the transient U channel can be accurately described up to range $\omega = 10$ Hz by $F(w) = 16\omega - \omega^2$. This is consistent with an operator that approximates the first derivative of its input, provided that the input signal has no significant power above 8 Hz. Since the U channel attenuates spatial frequencies above 3 cycles/deg, the channel will signal the derivative for edges and bars that drift across the retina with a velocity of up to about 3 deg/s. Figure 3–34 shows how closely

Figure 3–34. (opposite) The computed response of the transient U channel to a light edge, a thin bar, and a wide bar all moving at 3 deg/s. (a) The output of the spatial filter ($\nabla^2 G * I$) when the U channel parameters from Wilson and Bergen (1979) are used. The y-axis represents the normalized response, and the x-axis represents distance, the entire range being 3°. The x-axis in (b), (c), and (d) represents time; the entire range is 1 s. (b) The theoretically predicted output of the temporal filter if the transient channel carries $\partial/\partial t$ ($\nabla^2 G * I$). (c) The output of the temporal filter if Wilson's contrast-sensitivity curve is used and the filter is antisymmetric. (d) Comparison of (b) and (c). The thin bar is 2′ wide, and the thick bar is 40′ wide. In all cases the agreement between the curves derived from the time derivative hypothesis and the curves derived from the empirical observations is satisfactory. Hence for isolated bars and edges, the psychophysical evidence is satisfactory. Hence for isolated bars and edges, the psychophysical evidence is approximate the function $\partial/\partial t$



the measured characteristics of the transient channels match the expected behavior of the time derivative $\partial/\partial t(\nabla^2 G * I)$ for an isolated edge and a thin and a wide bar.

Turning to the neurophysiology, Rodieck and Stone (1965) described retinal ganglian cells whose response to a moving spot was "directly correlated with the gradient of the receptive field as defined by flashing lights" (p. 842). Of course, no physical device can take a perfect time derivative over the entire temporal frequency range. However, the published response curves of retinal and geniculate Y cells to bars and edges moving at moderate velocities closely agree with the predictions based on the time derivative operation $\partial/\partial t(\nabla^2 G * I)$. Figure 3–35 compares the predicted responses of on- and off-center Y cells with their observed responses to various stimuli. All the stimuli were light (that is, light edges and light bars); the thin bars were about ½° wide, and the thick bars 5°. The traces are taken from Dreher and Sanderson (1973). The predicted traces show pure values of $\partial/\partial t(\nabla^2 G * I)$, and as in Figure 2–17, the thickness of the thin and thick bars was, respectively, $0.5w$ and $2.5w$. The observed responses closely agree with the predicted ones, even in cases where both are elaborate (as with the wide bar).

The idea, that the X cells signal $\nabla^2 G$ and the Y cells its time derivative, which enables the construction of directionally selective, oriented zero-crossing segment detectors, offers a precise explanation for part of the function of the retina, and poses a fascinating challenge to the retinal anatomists and neurophysiologists—namely, How are these signals measured? Convolving with $\nabla^2 G$ is easy to imagine, but measuring $\partial/\partial t(\nabla^2 G * I)$ or even just determining its sign is quite a complicated task and requires both spatial and temporal comparisons: The center must be compared with the surround, and the result at a given time compared with the result a short time earlier, which means there must be a 60-ms memory there. In the retina, some of these components may be distorted, especially since comparing the values at two different times requires a delay. Hochstein and Shapley's (1976a) findings suggest, for example, that the Y-cell surround receives a delayed contribution from nearby units about the size of the centers of local X-cell receptive fields, and that this delayed input may be a major source of the observed nonlinearity. The nonlinear effects are induced primarily by gratings (Enroth-Cugell and Robson, 1966; Hochstein and Shapley, 1976a, 1976b). For isolated edges and bars moving at moderate velocities, however, the Y cells approximate $\partial/\partial t(\nabla^2 G * I)$ quite well, as we saw in Figure 3–35.

Provided that the Y channels deliver $\partial/\partial t(\nabla^2 G * I)$ and that positive and negative values are separated into different channels, the zero-crossing segment detector of Figure 2–18, reproduced in Figure 3–36, requires only

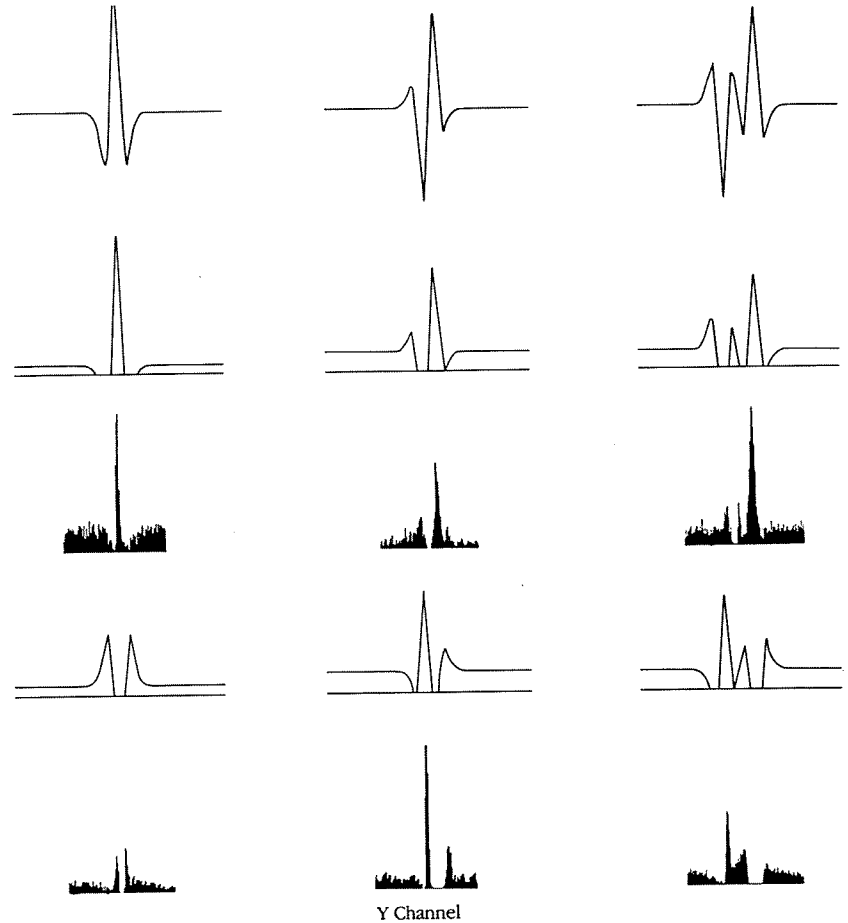


Figure 3–35. Comparison of the predicted responses of on- and off-center Y cells to electrophysiological recordings. The first row shows the response of $\partial/\partial t$ $(\nabla^2 G * I)$ for an isolated edge, a thin bar (bar width = $0.5w_{1\text{-D}}$, where $w_{1\text{-D}}$ is the width projected onto one dimension of the central excitatory region of the receptive field), and a wide bar (bar width = $2.5w_{1\text{-D}}$). The predicted traces are calculated by superimposing the positive (in the second row) or the negative (in the fourth row) parts of $\partial/\partial t(\nabla^2 G * I)$ on a small resting or background discharge. The positive and negative parts correspond either to the same stimulus moving in opposite directions, or stimuli of opposite contrast—for example, a dark edge versus a light edge—moving in the same direction. The observed responses (third and fifth rows) closely agree with the predicted ones, even in cases where both are elaborate (such as for the wide bar).

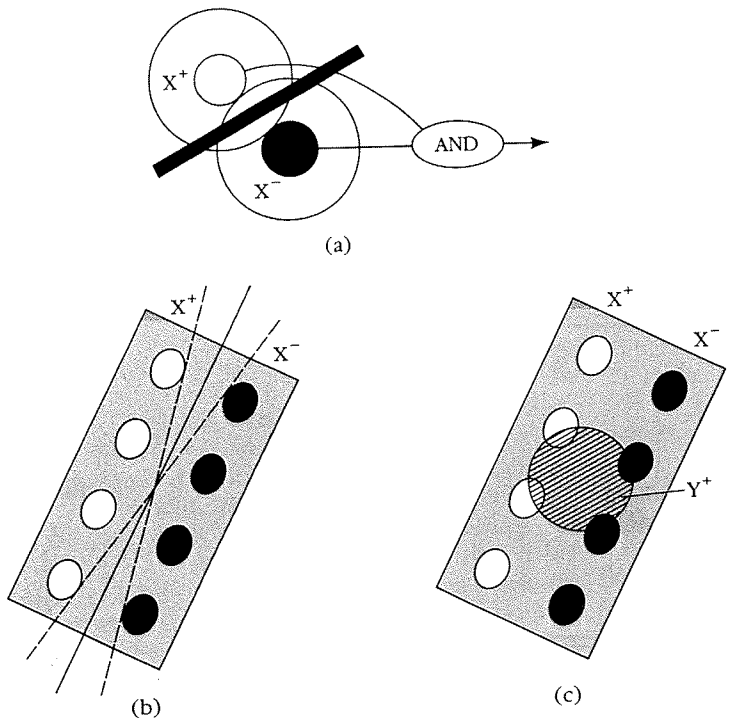


Figure 3–36. The detection of a moving zero-crossing. (a) X^- and X^+ subunits are combined through a logical AND operation. Such a unit would signal the presence of a zero-crossing of a particular sign running between the two subunits. A row of similar units connected through a logical AND would detect the presence of an oriented zero-crossing within the orientation bounds given roughly by the dotted lines in (b). In (c), a Y unit is added to the detector in (b). If the unit is Y^+, it would respond when the zero-crossing segment is moving in the direction from the X^+ to the X^-. If the unit is Y^-, it would respond to motion in the opposite direction.

the addition of one Y-cell input, again via an AND gate, in order to make it directionally selective.

The basic unit is shown in Figure 3–36(c), which is Marr and Ullman's (1979) XYX model for the simplest type of cortical simple cell. Its receptive field has three components, sustained on-center X inputs, sustained off-center X inputs, and a Y input. The X units need to be all the same size and

arranged in two parallel columns not more than $w_{2\text{-}D}/\sqrt{2}$ apart (where $w_{2\text{-}D}$ is the diameter of the central excitatory regions of the X-cell receptive fields). The Y-cell input can in principle be satisfied by a single input whose receptive field is positioned centrally or a little toward one side (toward the positive column for on Y units and the negative column for off Y units).

The ideal scheme requires a strict logical AND operation between the outputs of the subunits. In practice, this could be implemented by a strong multiplicative interaction between the columns and the Y input, and a weaker nonlinearity down the columns. Such a unit would respond optimally to a moving zero-crossing segment that extended along the entire length of the columns, but it would also respond to shorter stimuli and even to moving spots of light. More complicated receptive fields (for example, moving bars or slits) can be built up from these units. A critical empirical characteristic of such a unit would be that if its Y-cell input is abolished, the cell either fails to fire at all or, if it does fire, it loses its directional selectivity. It is not yet known whether this is true of directionally selective units. Otherwise, the model's properties are in overall agreement with the available facts (Hubel and Wiesel, 1962, 1968; Schiller, Finlay, and Volman, 1976a, 1976b [called S_1 cells there]). The paper by Marr and Ullman (1979) contains a fuller account of the properties of and predictions from this model.

Using Directional Selectivity to Separate Independently Moving Surfaces

Computational theory

The movement of an object against its background can be used to delineate the object's boundaries, and the human visual system is very efficient at exploiting this fact. If the complete velocity field is given (that is, speed and direction at each point of the image), object boundaries will be indicated by discontinuities in this field, since the motion of rigid objects is locally continuous in space and time. The continuity is preserved by the imaging process and gives rise to what I earlier called the principle of continuous flow, according to which the velocity field of motion within the image of a rigid object varies continuously everywhere except at self-occluding boundaries. Since the motions of unconnected objects are generally unrelated, the velocity field will often be discontinuous at object boundaries. Conversely, as we saw in Chapter 2, lines of discontinuity are reliable evidence of an object boundary.

Unfortunately, the complete velocity field is not directly available from measurements of small oriented elements. Because of the aperture problem, only the sign of the direction of movement is available locally. This means that an additional stage is necessary for detecting discontinuities in the velocity field. In this section, we ask how and to what extent the limited raw information (the sign of the direction only) may be used to detect these discontinuities.

The sign of the local direction of motion determines neither the movement's speed nor its true direction, but it does place constraints on what the true direction can be (see Figure 3–37). The constraint is that the true direction of motion must lie within the 180° range on the allowed side of the local oriented element (Figure 3–37a), or, alternatively, it is forbidden to lie on the other side (Figure 3–37b). The constraint thus depends on the orientation of the local element. Hence, if the visible surface is textured and gives rise locally to many orientations, the true direction of movement may be rather tightly constrained.

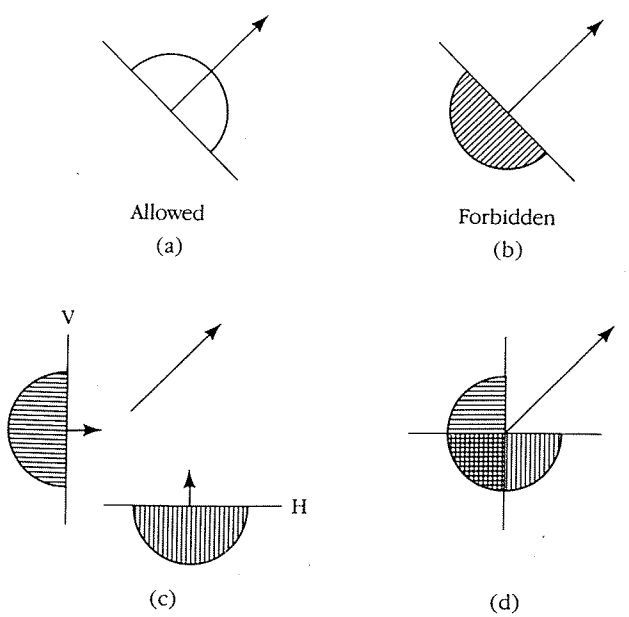


Figure 3–37. The combination of local constraints from directionally selective units to determine the direction of motion. The constraint placed by a single such unit is that the direction of motion must lie within a range of 180° on the allowed side (b). (c) The forbidden zones for two oriented elements (V = vertical; H = horizontal) moving along the direction indicated by the arrow. The forbidden zone horizontal) moving along the direction indicated by the arrow. The forbidden zone of their common motion is the union of their individual forbidden zones, as indicated in (d). The direction of motion is now constrained to lie within the intersection of their allowed zones, that is, the first quadrant.

Constraints can be combined as illustrated in Figures 3–37(c) and (d) for the simple case of two local elements. The true direction of motion is diagonal here. The vertically oriented directionally selective unit V sees motion to the right, and the horizontally oriented unit H sees motion upward. If these two units share a common motion, we can combine the constraints they place on the direction of that motion by taking the union of their forbidden zones (Figure 3–37d). The result is that the direction of motion is now constrained to lie in the first quadrant, as illustrated. Additional units can further constrain the true direction of motion by expanding the forbidden zone.

The diagram also shows how the motion of two groups of elements may be incompatible. If the allowed zone for one group of elements is completely covered by the forbidden zone of another, their motions clearly cannot be compatible. Notice in this connection that only the direction of movement, not its speed, is used here. A system that segments a scene in this way will be relatively insensitive to variations in speed.

The final observation that we need in order to use this scheme is that objects are localized in space. If the objects are also opaque, their images will have an interior within which the forbidden zones in diagrams like Figure 3–37(d) are consistent, provided that those forbidden zones draw their elements from small neighborhoods. Exceptions can occur, for example the center of a rotating disc, but only rarely. Hence, the method will be reliable. It is not, of course, exhaustive—if two surfaces are relatively stationary, this method will fail to separate them.

Algorithm and implementation

The diagrams of Figure 3–37 contain essentially all the information we need to know here, for the algorithm must consist of searching for neighborhoods with locally compatible directions of motion. Figures 3–38 to 3–40 show some results from a computer implementation of such an algorithm, written by John Batali. The first example, Figure 3–38, shows the detection of a moving pattern embedded in a pair of random-dot images. A central square in Figure 3–38(a) is displaced to the right in Figure 3–38(b), while the background moves in the opposite direction. Figure 3–38(c) depicts the zero-crossing contours of Figure 3–38(a) filtered through $\nabla^2 G$. Figure 3–38(d) represents the values of the transient channel if the two frames shown in Figures 3–38(a) and (b) are presented

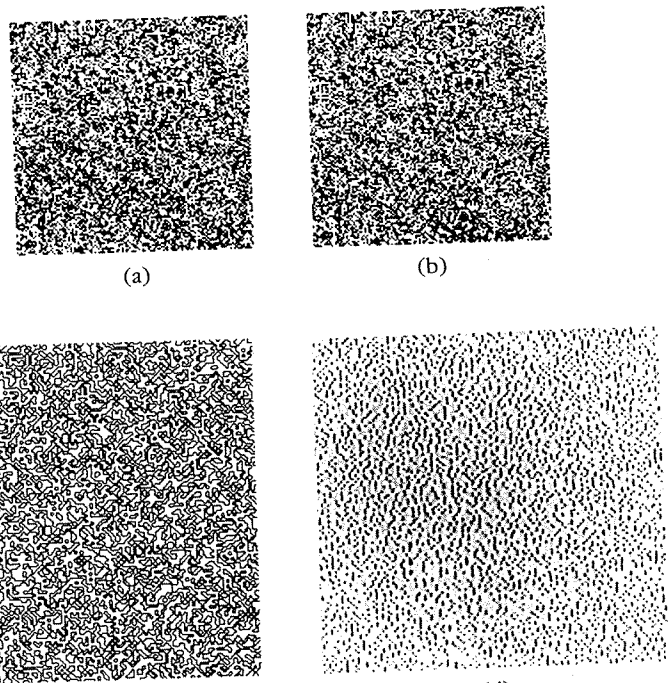


Figure 3–38. Separating a moving figure from its background by using combinations of directionally selective units. A central square in (a) is displaced in (b) to the right. The background in the two pictures moves the opposite way. (c) The zero-crossing contours of (a) filtered through $\nabla^2 G$. (d) The convolution of the difference between (a) and (b) with $\nabla^2 G$. If (a) and (b) are presented in rapid succession, the function shown in (d) approximates the value of $\partial/\partial t(\nabla^2 G * I)$. The images are 400 × 400 pixels, the inner square is 200 × 200, each dot is 4 × 4, and the motions are 1 pixel. (Courtesy John Batali.)

in rapid succession. Figure 3–40(a) shows the results of applying the XYX-motion-detection operation to the zero-crossings of Figure 3–38(c). The direction of movement has been coded, as indicated by the star in the figure. As can be seen, black represents motion to the right, and white represents motion to the left. The central square is clearly delineated by discontinuities in the direction of motion.

The same analysis was also applied to the natural images shown in [illegible] frames taken from a 16-mm film of

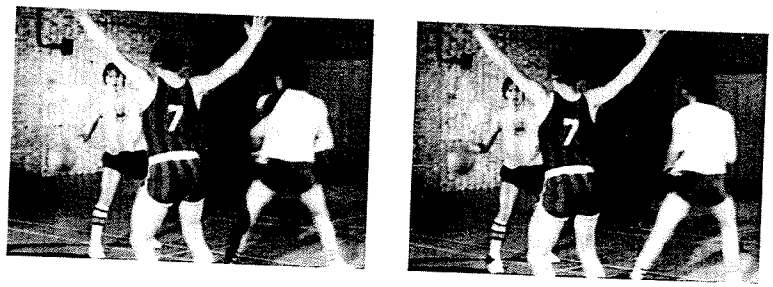
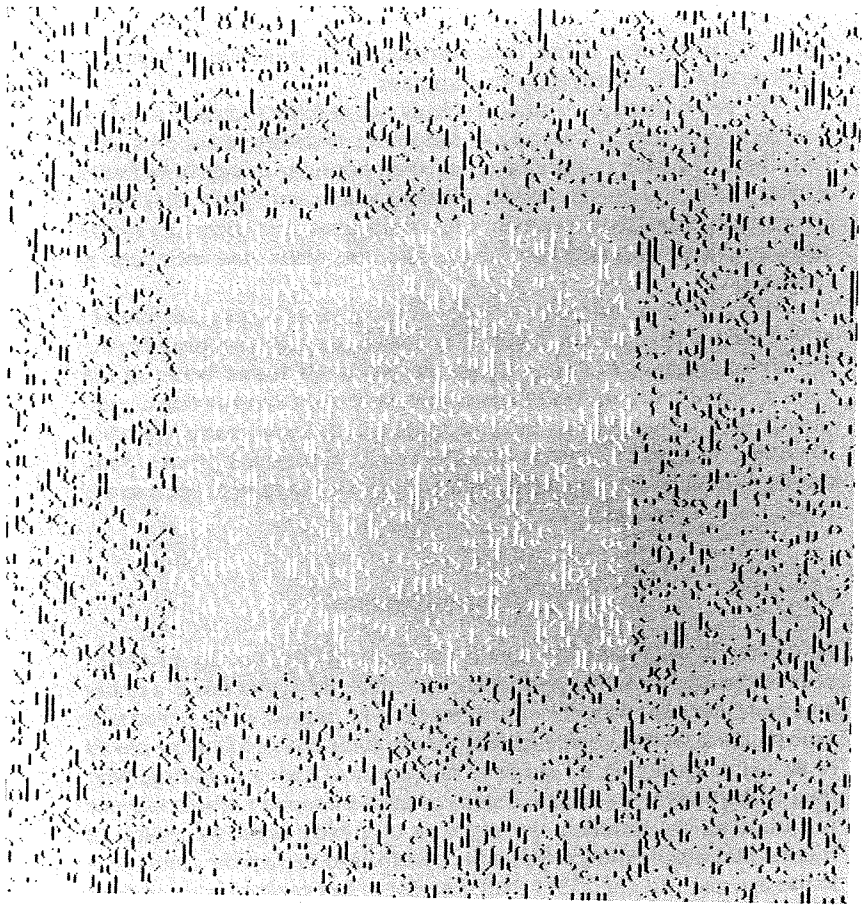


Figure 3–39. Two successive frames from a 16-mm movie of a basketball game. The same analysis was applied as to the random-dot patterns in Figure 3–38. (Courtesy BBC.)

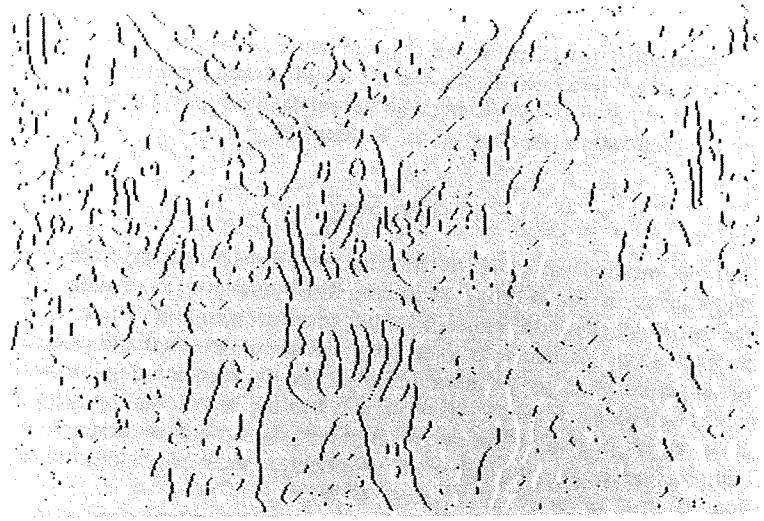
a basketball game. The results appear in Figure 3–40(b). For example, the left arm of player 7 moved downward and to the left, and the rightmost player moved to the right. Because of the extreme sensitivity of the method, small registration errors, more or less unavoidable because of the way the two images are digitized, sometimes give rise to spurious motion of the background.

Psychophysically, the XYX-motion-detection scheme fits well into the first of Braddick's two categories. For example, the phenomenon should occur only over short ranges (around $w/\sqrt{2}$ or 15′ at 5° eccentricity) and short ISI's (not more than the total time course of the temporal component of the transient channel, about 120 ms), according to Wilson's channel data. If speed and not direction were the only available discriminant, separation should be impossible, which we have found psychophysically (Figure 3–32).

In addition, the amount of information that can be obtained from directional selectivity depends on the direction of movement and on the orientation of the moved elements. Hence, the same velocity field may be seen as coherent or incoherent, depending on the orientations of the moved elements. The reason is that two nearby velocity vectors will produce the same directional sign on an element oriented roughly perpendicular to them but different signs on an element whose orientation bisects them. We also found this to be true psychophysically. Moreover, if the formation of coherent groups proceeds roughly in the manner of Figure 3–37, one might expect to see clusters of locally coherent motions in even purely random display sequences—and, in fact, one does. Such a mechanism also produces Anstis' (1970) reversed phi phenomenon, whereby simultaneous movement and contrast reversal can give rise to the illusion of movement in the opposite direction (see Marr and Ull[illegible]



(a)



(b)

Figure 3–40. Motion assigned to zero-crossings from the images of Figures 3–38 and 3–39. The direction of motion was assigned according to the rules described in the text and the result is displayed here using shades of gray. The keys below (a) and (b) indicate the shade of gray assigned to each direction. In (a), the central square clearly moves right, while the surround moves left. In the zero-crossings from the basketball game (b), the left arm of player 7 moves to the left and down, while the player to his right moves to the right. (Courtesy John Batali.)

Finally, the use of color but not luminance boundaries or the interposition of a white field during the ISI could disrupt the mechanism, as Braddick requires, by interfering with the retinal machinery for measuring the time derivatives traveling up the Y channels.

Looming

There is another way in which the outputs of directionally selective units might prove useful, because combining directionally selective units from the two eyes yields a different kind of information (Marr and Ullman, 1979). Suppose that a particular zero-crossing has been identified and assigned incompatible motions in the two images. Then the zero-crossing is moving in depth either toward the viewer if it is moving away from the nose on both retinas, or away if the motion is toward the nose. If motion is to the right on both retinas, the object will pass safely to the viewer's left, and vice versa (Regan, Beverley, and Cynader, 1979).

For this type of analysis, it is not necessary to combine constraints in the manner of Figure 3–37; the raw output of the directionally selective units can be used. The difficulty in this case lies in ensuring that both left and right detectors are looking at the same zero-crossing; establishing this match is the essence of the stereo matching problem. However, if inaccuracies from time to time are tolerable, a fast looming detector can be designed that does not have to wait for the results of stereo matching. For example, a simple looming detector can be constructed by comparing the signs of motion at corresponding retinal points. Such points will often but not always correspond to nearby points on the same moving object.

Such a scheme might rely at some point on a cell that has binocular receptive fields close by in the visual field, but not truly disparity sensitive, and whose preferred motions in the two eyes are opposite. There is some evidence for the existence of such cells (Regan, Beverley, and Cynader, 1979).

3.5 APPARENT MOTION

In the last section, we saw how very limited information about the motion in the visual field could be used at quite a primitive stage in the processing to provide certain rather rough information about how to decompose the scene into different surfaces. We also saw that this task can be done rather fast. With a little more time and care, however, visual motion can be made to yield a much richer harvest of information. Although the experiments of Miles (1931) and of Wallach and O'Connell (1953) preceded it, Ullman's [illegible] demonstration (illustrated later in Figure 3–52) is the most telling demonstration so far contrived of what our visual systems can obtain from visual motion.

The demonstration consists of a sequence of frames, each of which is a projection of a set of dots on two concentric, counterrotating cylinders. Only the dots appear in each frame, and their positions change from frame to frame. As in the case of random-dot stereograms, each individual frame has no visible structure. However, when the frames are shown as a movie sequence, a vivid impression of the two counterrotating cylinders is obtained.

From this demonstration, it is clear that our visual system has remarkable powers to recover the shapes of unknown structures simply from the way their appearances change in the image. In his recent book on the subject, Shimon Ullman (1979b) has gone far toward constructing a complete theory of how this may be done, and he includes supporting psychophysical evidence. This section consists of a summary of Ullman's work, together with one or two general points that I wish to raise about it in the context of vision in general.

Why Apparent Motion?

Movement is an inherently continuous process that usually produces smooth changes in an image. Indeed, one might think that this is a rather important intrinsic property of movement with regard to its perceptual analysis, since its very continuity should help in the task of following pieces of an object around in an image to find out how they are moving. Why, then, is this section based on the study of apparent motion, whose essence is a discrete, discontinuous presentation of a fairly rapid succession of frames? Surely something is lost in the transformation from the continuous to the discrete. The theories I shall describe in fact apply to both kinds of motion, continuous and framed (or apparent). But that is not quite a satisfactory answer, and it is worth a little discussion to see that for the type of situation of interest here, one probably can think in terms of framed stimuli.

The first point is that we are no longer dealing with almost instantaneous phenomena, as we were in the last section. We are out of the realm of detection tasks here. Instead of finding out something simple but possibly important within 50 ms, we can afford to take quite a long time—say, ¼ to ½ s, which is large by perceptual standards—to allow the image to change by a reasonable amount. For we want not just to detect the change but also to measure its extent and use this information. So our fundamental approach is to contrast the positions of items at one time in the image with their positions at a sufficiently later time for the differences to be [illegible]

reliably, and we shall use the differences to make calculations about the underlying shapes and movements.

Thus, it is in our interests to delay matters, at least up to a point, but not too much, or the image will have changed beyond recognition—visible portions of the surface may become occluded or may rotate out of view. But at least in principle, it is the changes over a period of time that we need here, and they must be determined quite accurately.

That may be, one can reply. But the fact is that even if we want only to know where things have moved after 100 ms or so, surely it is easiest to find this out by smoothly following them. And isn't this made more difficult by cutting the sequence into distinct frames? Well, up to a point this must be true. On the other hand, if the frame rate is sufficiently fast compared with the time constants in, say, the cones (which are of the order of 20 ms or so), the two situations will be indistinguishable. We all know, too, that we can watch movies perfectly well and that the motion there looks quite normal. Yet they are split into only 24 frames per second, and one cannot discern these facts from perceptual evidence alone. In addition, psychophysical presentations consisting of just two frames separated by as much as 300 ms can give the subjective impression of smooth motion.

So, although the continuous problem may be slightly simpler than the recovery of structure from apparent motion, it is probably not much simpler and we can certainly do the harder problem involving apparent motion. The apparent motion problem is also much easier to formulate and to investigate empirically, and its results can be applied to the continuous case. It therefore seems sensible to solve this problem first and then to take stock of where we stand.

The Two Halves of the Problem

Our goal here, then, is not so much to detect the changes induced by motion but to measure and use them to recover the three-dimensional structures in motion. Broadly speaking, this introduces two kinds of task that, at least superficially, look rather different and somewhat analogous to what we met in stereopsis. The first task is to follow things around as they move in the image and to measure their positions at different times. This is the *correspondence problem,* and at its heart is the question, Which item in the image at time t_1 corresponds to which item at time t_2? The second task is to recover three-dimensional structure from the measurements supplied by the results of the first task, and this is called the *structure-from-motion problem.*

Apparently, these two problems are solved independently by the human visual system, and it is a great stroke of good fortune that they are separate. The critical empirical evidence for this is that none of the measurements on which the correspondence process rests involve three-dimensional angles or distances—they are all two-dimensional measurements made on the image (Ullman, 1978). Thus, there is no deep need for any feedback from the later task to the earlier.

The two tasks may therefore be dealt with independently. We shall first examine the correspondence problem and then alternative approaches to the second task. By now the reader can formulate for himself the critical preliminary question—What are the primitives on which the process operates, or, in our earlier terms, what is the input representation for the process? And since the measurements of changes in position must refer to the changes in position of an identifiable surface location, these primitives need to be as physical as possible. So, the reader will not be surprised to learn that the elements in the primal sketch seem to be used, although various interesting side issues arise in the details.

Then we must formulate the relationships that should hold between the positions of the primitives in adjacent frames (remember, we shall be dealing with apparent motion). In general terms, it is not hard to see that the closer and more similar two items are in successive frames, the more likely it is that they correspond. This simply reflects some kind of a statistical rule of the universe, and it will hold provided that the interframe interval is not too long in relation to the velocities of and distances involved with the visible motions. It turns out that the human visual system incorporates a permanent or "hard-wired" table of similarities by which the similarities and dissimilarities in the various parameters may be compared. For example, in experiments that test the similarity of two lines of the same contrast in successive frames, a change in length by a factor of 3/2 produces the same change in similarity as a change in orientation of 45°.

This similarity Ullman called the affinity measure, and it is based on two-dimensional measurements. However, this does not by itself determine the correspondence process. In order to do so, one has to take account of extra factors. For example, suppose one has two lines A and B in the first frame, and two lines a and b in the second. There are four possible pairings.

(1) $A \rightarrow a$ and $B \rightarrow b$
(2) $A \rightarrow b$ and $B \rightarrow a$
(3) $A \rightarrow a$ and $B \rightarrow a$
(4) $A \rightarrow b$ and B → b

This list omits possibilities like $A \rightarrow a$, and B goes nowhere. The question is, How does one decide which of the possible pairings actually occurs?

The obvious answer is, take that solution which maximizes the overall similarity between the frames. This similarity can be measured by means of some standard cost function that gives a similarity value to each pairing in a given solution, the overall similarity being the sum of the values for each pairing. The cost function tells us roughly how many quite poor pairings should be accepted in order to avoid an abysmal pairing or to acquire an excellent one in the overall match.

An approach of this type, which involves finding a solution that achieves an overall or global minimum, is analogous to part of what the Gestalt movement became interested in during the first third of this century, although several different phenomena were probably involved in the experiments that the Gestaltists actually carried out. They had the idea of an attraction among elements that bound them into wholes and governed the interaction between successive frames, but they were unable to see how much such an approach could account for the complexity that they saw in the correspondence process. Their basic difficulty was this: In a display such as Figure 3–41, they saw that $A \rightarrow A'$ and $B \rightarrow B'$; but if A and B' were removed, $B \rightarrow A'$. Hence, they reasoned, movements of wholes are of critical importance, and the phenomenon cannot possibly be explained in a purely local way. In large measure, this type of argument killed the school, because the Gestaltists viewed the problem of the formation of wholes as intractable.

There are two fundamental misconceptions here, and I shall make a point of them in order to draw a moral. The first is the point of basic mathematical ignorance. Certainly examples like Figure 3–41 show that the correspondence process involves more than finding purely local minima; if the problem can be formulated in this way at all, the minimum one wants is a global minimum. But—and here is the first point—there are many systems in which global minima can be found using only local interactions, so the Gestaltists' findings did not force the conclusions they drew about the insufficiency of local interactions. In particular, the most obvious way out of the Gestaltists' problem with Figure 3–41 is to say that the total cost of $(A \rightarrow A') + (B \rightarrow B')$ turns out to be less than $(A \rightarrow B') + (B \rightarrow A')$. The idea seems even simpler when we observe that it is linear, and linear systems are extremely well-behaved—basically, they cannot get caught in local minima. In fact, Ullman's correspondence theory is essentially a linear one.

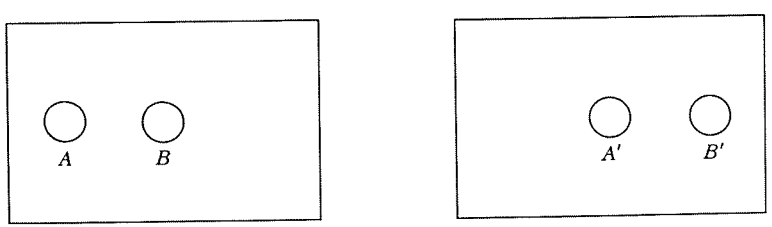


Figure 3–41. One of the patterns that puzzled the Gestaltists. (a) shows frame 1, and (b) frame 2. Perceptually, A goes to A' and B to B', so that B seems to move. (Courtesy of Shimon Ullman.)

The second misconception was that the Gestaltists lacked the idea of a process. They thought of groupings as being subject to various types of rules—the principles of closure, good continuation, regularity, symmetry, simplicity, and so forth (see Koffka, 1935, p. 110)—which were summarized as the Gestalt law of Prägnanz. This law was to them like a physical law. If they had had the idea of embodying such principles in a number of grouping processes—for example, as constraints on what should or should not be grouped together—they might not have abandoned the other half of their endeavor, the systemization of the formation of wholes.

The moral here is this. We saw in Chapter 1 something of the perils to purely computer vision workers of ignoring the biological evidence about how the human visual system is organized. The basic difficulty is that such an oversight can lead to trying to solve problems which are not really problems at all, but which happen to arise because of the particular limitations of sensors, or hardware, or available computer power. Here we see the opposite, in which mathematical ignorance (which could have been avoided) and a failure to think more in terms of processes (which is more excusable) led to the failure of a school of thought that had actually made a number of valuable insights. The moral is that ignorance in any of these three fields can be damaging. Just as the modern physicist has to know some mathematics, so must the modern psychologist, but the psychologist must also be familiar with computation and have a clear idea of its abilities, its limitations, the fruitful ways in which to think about processes, and, most importantly, what it takes to understand these processes.

This, then, is roughly the current state of affairs concerning the correspondence problem. Ullman formulated it as a linear minimization problem and showed how this can account for much of the psychophysics. We shall explore his ideas in some detail and see some even more recent ideas about their biological implementations based on the higher-level primal sketch primitives. As for the topic as a whole, it is not yet solved at any of our three levels; however, a substantial amount is known about it, and a complete computational theory of it cannot, I think, be too far off.

The second half of the problem, the structure-from-motion theory, is in better shape and has essentially been solved at the level of computational

theory (Ullman, 1979a). The form of the theory is by now familiar—it is the same as we saw in Chapter 2 (for the primal sketch) and earlier in this chapter, although chronologically Ullman's was one of the early theories. The critical additional constraint that he used was rigidity, and he made a very precise formulation of its use and showed how the recovery of three-dimensional structure may proceed from the measurements made available by a successful correspondence process. The underlying mathematics consists of a theorem stating essentially that three views of four rigid, noncoplanar points are sufficient for recovering their three-dimensional dispositions and motion. We shall see how this result may be used as the cornerstone of the interpretation of visual motion. Longuet-Higgins and Prazdny (1980) used a similar approach in their study of optical flow.

One final comment is perhaps in order as a conclusion to this brief survey. Although the geometry of three-dimensional space has been studied since the time of Euclid, some relatively simple theorems still appear to be unknown. The four-points, three-views theorem was one, and we shall meet another when we discuss the recovery of shape information from silhouettes (Marr, 1977a). It is difficult to believe that there are not others. These two have recently been formulated because the imaging process occurs in three dimensions, and hence certain types of geometrical relationships, if known and used, can be incorporated into processes for interpreting images. It may be well worth a mathematician's time to look again into the subject of three-dimensional Euclidean geometry.

The Correspondence Problem

Empirical findings

What is the input representation?

On general grounds, we require that the tokens on which the correspondence process operates, which we shall call *correspondence tokens*, be physically meaningful. This eliminates raw gray-levels, and one can directly demonstrate that, in the human visual system, gray-level correlation does not form the basis for the correspondence process. Figure 3–42 shows how. The maximum gray-level correlation between the two frames in Figure 3–42(a) occurs at zero displacement, as can be seen from the correlation graph, Figure 3–42(b). If the sharp edges are matched, however, one would expect edge *E* in frame 1 to jump to *F* in frame 2, and this is in fact what happens.

This demonstration establishes that the correspondence takes place above the level of gray-level intensity values, but how far above does one

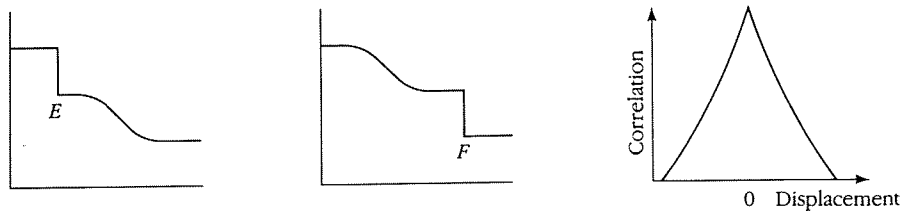


Figure 3–42. Correspondence is not established between gray-level images. If it were, two frames with the intensity profiles shown in (a), when presented in succession, would give no impression of movement, since the maximum value of the correlation between them occurs at zero displacement (b). Instead, edge *E* is seen to move to edge *F*, suggesting that edges, not gray-level images, are the tokens used in the correspondence process. (Reprinted from Shimon Ullman, *The Interpretation of Visual Motion* by permission of The MIT Press, Cambridge, Massachusetts, Figure 1.1, Copyright © 1979 by The Massachusetts Institute of Technology.)

go? Is the correspondence established between relatively small and simple parts of a scene, largely independently of shape and form, or are much more complicated descriptions involved, like the interpretation of the whole of a shape from one frame, before different frames are compared?

Figure 3–43 is one of a series of demonstrations that rules out the second alternative. The figure illustrates two successive frames, one denoted with full lines, and the other with dotted lines. If the whole pattern was analyzed from one frame, with the shape of the wheel extracted and then used to match the elements in the next frame, the observer should perceive the frames as a whole wheel rotating when they are presented in rapid succession. Notice, however, that the inner and outer parts of the wheel have their closest neighbors in one direction, whereas the central ring has its closest neighbors in the other direction. Because of this, if the matching is carried out purely locally, the observer should see the central ring rotating one way and the inner and outer rings rotating the other (as shown with arrows in Figure 3–43). When appropriately timed, this is in fact what happens.

This begins to suggest primal sketch elements, and the next demonstration shows that terminations play a role (as they do in stereopsis). In Figure 3–44(a), a correspondence is established between the ends of the two lines. This breaks down if the distances between the corresponding ends are much greater than that between the line segments, as they are in Figure 3–44(b), in which case a correspondence is established between the short line and only the nearest part of the long one. It has not yet been

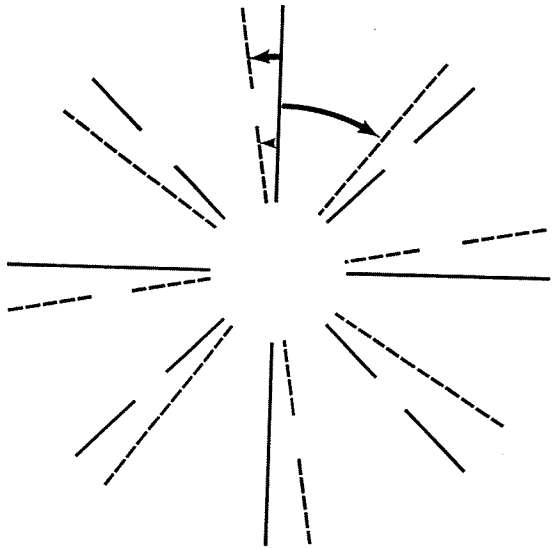


Figure 3–43. Evidence that the correspondence problem for apparent motion involves matching operations that act at a low level. Frame 1 is shown with full lines and frame 2 with dotted lines. Instead of appearing as a single wheel rotating, the wheel splits when appropriately timed, the outer and inner rings rotating one way and the center rotating the other as indicated by the arrows. This suggests that matching is carried out on elemental line segments and is governed primarily by proximity. (Reprinted from Shimon Ullman, *The Interpretation of Visual Motion,* by permission of The MIT Press, Cambridge, Massachusetts, fig. 1.3. Copyright © 1979 by The Massachusetts Institute of Technology.)

firmly established whether discontinuities of the type shown in Figure 3–44(c) are matched, but the question is obviously of interest.

Figure 3–45 adds to the evidence that correspondence is determined by quite low-level tokens and not by the shape or form of the corresponding figures. In Figure 3–45(a), the square *A* goes to the larger square *B*. In Figure 3–45(b), it goes to the larger triangle *B*, not to the smaller square *C*. Thus in these cases the motion of the constituent elements rather than the similarity between the complete forms governs the matching process. Ullman (1979b, p. 27) concluded that (1) differences in the tendency of different figures to fuse is consonant with the motions established between their components, and (2) there are no indications that structural figures are part of the basic elements or that the correspondence process is based on figural similarity.

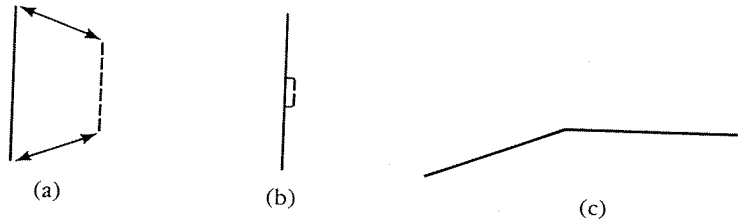


Figure 3–44. Terminations can also act as correspondence tokens, provided that the two lines in successive frames do not differ too much in length (a). If they are very different (b), correspondence is established between the short line and a segment of the long line. It is not yet known whether orientation discontinuities such as the one shown in (c) can act as correspondence tokens. (Reprinted from Shimon Ullman, *The Interpretation of Visual Motion,* by permission of The MIT Press, Cambridge, Massachusetts, fig. 2.10. Copyright © 1979 by The Massachusetts Institute of Technology.)

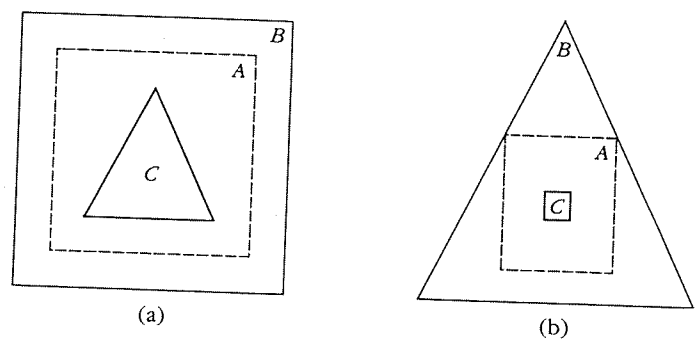
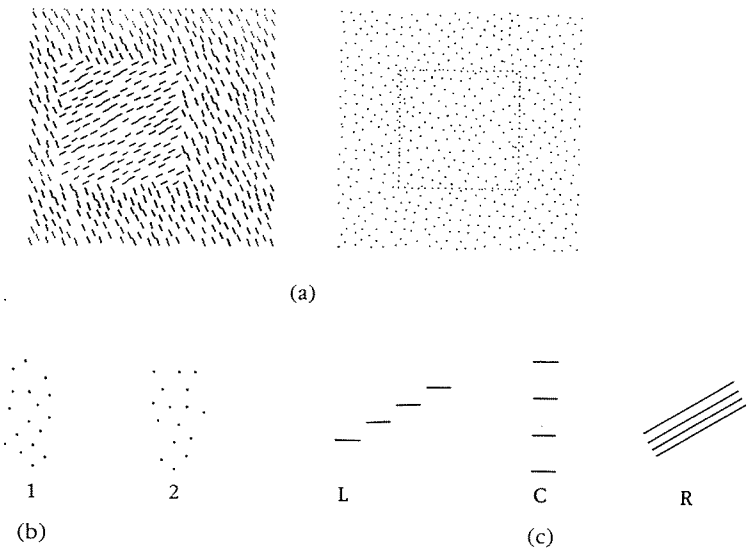


Figure 3–45. In (a), the square *A* goes to the larger square *B*, yet in (b) it goes to the larger triangle *B*, not to the smaller square *C*. This is more evidence that correspondence is governed by the motions of constituent elements, not by complete forms. (Reprinted from Shimon Ullman, *The Interpretation of Visual Motion,* by permission of The MIT Press, Cambridge, Massachusetts, fig. 1.6. Copyright © 1979 by The Massachusetts Institute of Technology.)

As a result of discussions between Shimon Ullman, Michael Riley, and myself, Riley has found that matches can, for example, be established between oriented clouds of dots or between groups of parallel lines, when in neither case do the constituents match. Two illustrations of this phenomenon appear in Figures 3–46(b) and (c). In such cases, the matching rules appear to be governed by parameters like the overall orientation and size of the group. Borders like those in Figure 3–46(a) can also be matched, although there can be no



–46. Matching can take place between higher-order borders or tokens even when the ents do not match. For example, correspondence can be established between the two kinds r surrounding the squares shown in (a). (b) An experiment in which one cloud of dots is d in frame 1, and two clouds in frame 2 (as marked), with the property that one of the ı the second frame is identical to the cloud in the first frame, whereas the other cloud is preference for the identical cloud is exhibited. In (c), this idea is carried further. The first nsists of group C, consisting of short horizontal lines. The second frame consists of two . comprised of short horizontal lines and R of long diagonal lines. The observer sees ence for motion from C to L, which proves that the correspondence in this case is not ried out between the constituents of the groups but between descriptions of their overall

The ISI's here are around 100 ms, much shorter than the ⅓ s or so required for shape to begin influencing matching.

So Ullman's conclusions may need slight modification so that these more abstract image descriptors from the full primal sketch can be included. However, his main point—that no elaborate form analysis precedes the correspondence process—still stands. And the limitations implied by the word *elaborate* here effectively allow the things that are allowed in the full primal sketch—overall length, size, orientation of a token, and so forth—but exclude the things that are excluded there—for example, any explicit representation of an internal angle in the token, the noticing of right angles, and so forth. It will be interesting to see how far the analogies can be taken between correspondence tokens and primitives in the full primal sketch.

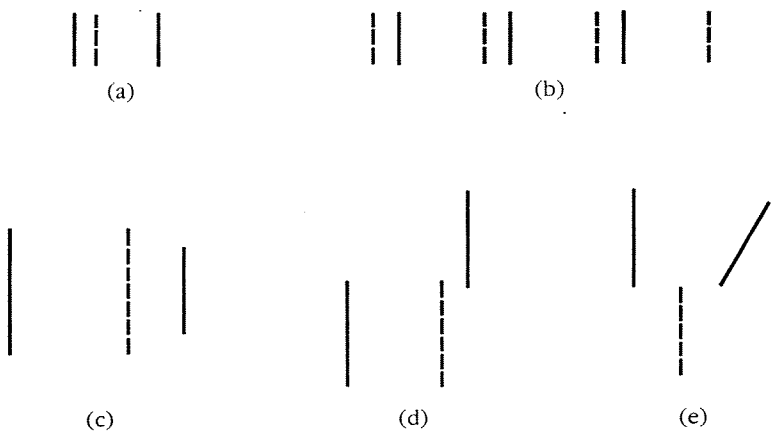
Two-dimensionality of the correspondence process

The local behavior of the correspondence process for small numbers of isolated elements can be studied in experiments like that shown in Figure 3–47(a), in which the first frame (dotted line) contains one element and the second, two (solid lines), and the observer is asked to which line in the second frame the line in the first frame appears to go. Riley recently modified this scheme to the form shown in Figure 3–47(b), which has many copies of the same problem. The extended display has the advantage of being somewhat more sensitive.

Figures 3–47(c), (d), and (e) show stimuli for these experiments; in each case, frame 1 is dotted and frame 2 is not. The examples shown all have approximately the same affinity for the original. Figure 3–47(c) shows how length trades off against distance, Figure 3–47(d) shows how vertical displacement trades off against distance, and Figure 3–47(e) shows how orientation trades off against displacement. The relative weights of the different parameters for a 3 line configuration are tabulated in Figure 3–47(f) (from Ullman, 1979b, table 2.1).

For our brief survey of this problem, the detailed values of the table in Figure 3–47(f) do not matter so much, but the fact that the process uses measurements made on the image and not measurements of objective, three-dimensional quantities is important. This was established by Ullman (1978) in the type of experiment shown in Figure 3–48. In frame 1 of the experiment shown in Figure 3–48(a), for example, all the lines had the same brightness except for C. In frame 2, only L and R were brighter, and motion was induced from C to L or R. In this example, the two-dimensional relations between C and L and between C and R are identical. Their three-dimensional distances apart, however, are very different. In Figure 3–48(b), an experiment along the same lines is shown in which the three-dimensional distances are the same but the two-dimensional ones are very different. Similarly, in Figure 3–48(c) the two-dimensional and three-dimensional angles are different.

From experiments like this, Ullman concluded that three-dimensional measures were irrelevant to the correspondence process; everything he found could be predicted from the two-dimensional configurations. He was also able to make another fascinating point, about the smoothness of apparent motion. When one looks at two frames, the transitions from one



	Relative Weight				
	0	1	2	3	4
Orientation difference (in degrees)	15	30	45	60	75, 90
Distance ratio	1.1	1.2	1.6	2.25	2.7, 3.8
Length ratio	1.04	1.13	1.5	2.1	2.5
1/cos α	1.04	1.15	1.41	2.0	(2.3)

(f)

Figure 3–47. (a) shows a typical two-frame experiment of the kind used to measure affinities, and (b) shows a more sensitive version of the same experiment. In (c)–(e), frame 1 is shown with dotted lines and frame 2 with full lines, and the two stimuli in frame 2 have about the same affinity for the original. (c) How length trades off against distance; (d) how displacement trades off against distance, and (e) how displacement trades off against orientation. The measured affinity values are tabulated in (f). (Reprinted from Shimon Ullman, *The Interpretation of Visual Motion,* by permission of The MIT Press, Cambridge, Massachusetts, figs. 2.5–2.9 and table 2.1. Copyright © 1979 by The Massachusetts Institute of Technology.)

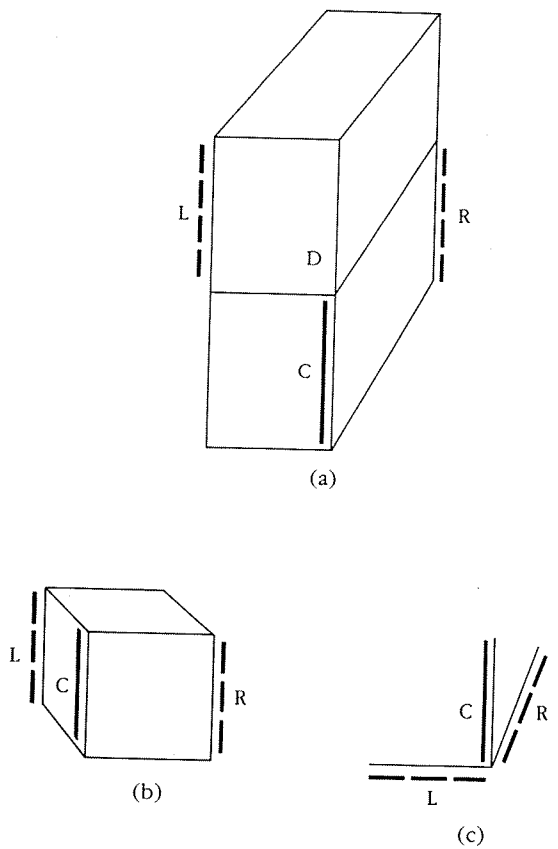


Figure 3–48. Only two- and not three-dimensional measures are used by the correspondence process. In (a), a correspondence is established between C (frame 1) and L and R (frame 2), which have identical two-dimensional relationships to C but different three-dimensional ones. They behave identically. In (b), L is preferred over R. (c) tests angles, and again it is two-dimensional angles that determine the correspondence. (Reprinted from Shimon Ullman, *The Interpretation of Visual Motion,* by permission of The MIT Press, Cambridge, Massachusetts, fig. 2.22. Copyright © 1979 by The Massachusetts Institute of Technology. Part reprinted by permission from Shimon Ullman, "Two dimensionality of the correspondence process in apparent motion," *Perception* 7 (1978), 683–693, fig. 1.)

to another sometimes seem to happen smoothly and sometimes not. Studies like those by Corbin (1942) and Attneave and Block (1973) had found that smoothness of motion was determined predominantly and perhaps entirely by perceived three-dimensional distance rather than by objective two-dimensional distance. Yet Kolers (1972, ch. 4 and 5) was only the most recent of a line of researchers who studied correspondence strength using smoothness of motion as a criterion.

Plainly, there was some inconsistency, because the three claims—(1) smoothness of motion depends on perceptual distance, (2) correspondence strength depends on two-dimensional distance, and (3) smoothness of motion reflects correspondence strength—are incompatible. Ullman (1978, experiment 5) resolved the dilemma by constructing a situation like Figure 3–47(a), in which motion one way was smoother but motion the other way was stronger and won. Smoothness and correspondence strength are therefore different phenomena, and the correspondence process relies on two-dimensional measurements only, probably after allowing for the effects of eye movements (Rock and Ebenholtz, 1962).

Ullman's theory of the correspondence process

In more complex displays, an element does not always map to the element with highest affinity, as we have seen in Figure 3–41. Mappings are affected by inter-element interactions as well. In his empirical approach to this, Ullman introduced the notion of *correspondence strength* (CS), which is derived from the local affinities but also incorporates the effects of various kinds of local competition and which determines the final mapping. Figure 3–49 illustrates this idea. First, the affinity between each pairing is measured, and then local interactions take place on these to produce the CS. The interactions weaken the CS when splitting or fusion occurs, for example, and so these conditions are avoided. In a particular numerical example (appendix 4 of his doctoral thesis), Ullman showed that this same simple scheme could account for several examples that were considered challenging to motion perception theories (Kolers, 1972; Attneave, 1974; Ullman, 1979b, sec. 2.4.1).

These points, though, primarily showed that the kind of thinking that had been used when examining the capabilities of local interactions was still often seriously flawed, sometimes in the same way as was the Gestaltists', by a failure to appreciate the complexity of functions that can be computed by local interactions. More interesting was Ullman's attempt to formulate a theory for the correspondence process, which he called the minimal mapping theory. It is, in fact, a maximum likelihood theory.

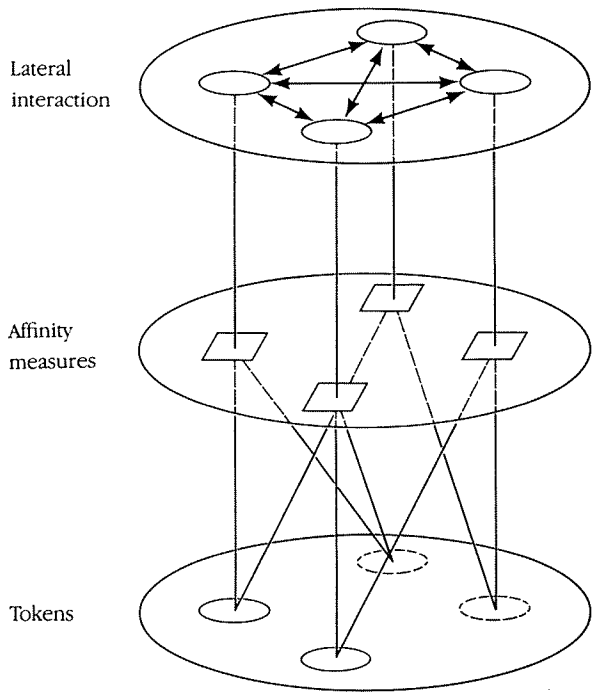


Figure 3–49. Ullman's approach to correspondence strength. Raw affinity values are measured between correspondence tokens, and then local interactions take place between them to obtain the final correspondence strengths.

There are three main assumptions behind the theory. The idea is to provide a way of judging the relative merits of pairing tokens between frames. Since the underlying argument is probabilistic, we need to assume that different pairing decisions are independent. That is the first assumption. The second is that each token in frame 1 is paired with at least one token in frame 2, and vice versa. We do not explicitly demand a one-to-one relationship (that is how splits and fusions are allowed), but since each pairing costs something, the final answer keeps splits and fusions to a minimum. Thus, the second assumption is that the set of pairings should cover both sets of tokens.

The third idea is the interesting one. Of course, the range of true velocities in the world varies widely—sometimes a viewer moves fast,

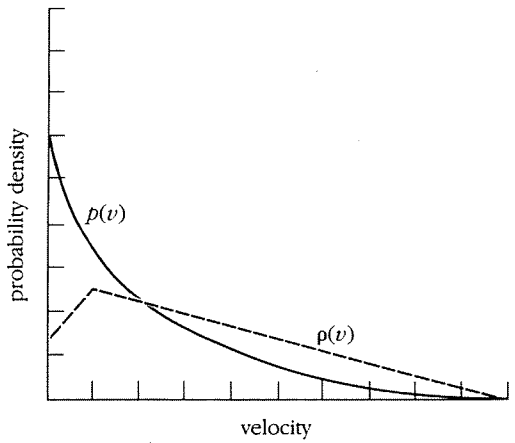


Figure 3–50. The average distribution of velocities in an image. For almost any reasonable velocity distribution for objects in the world such as $\rho(v)$, after projection into an image, $p(v)$, small velocities will predominate. See discussion in text. (Reprinted from Shimon Ullman, *The Interpretation of Visual Motion,* by permission of The MIT Press, Cambridge, Massachusetts, fig. 3.11. Copyright © 1979 by The Massachusetts Institute of Technology.)

sometimes slowly, sometimes objects move by quickly, sometimes not. But almost whatever is chosen for distribution of velocities in the world, the projections of those velocities in the image will usually be small rather than large, simply because of the imaging process. This point is illustrated in Figure 3–50. The dashed line $\rho(v)$ shows one choice for the probability distribution for true velocities in space. The solid line $p(v)$ shows the corresponding projected velocity distribution. Thus on only very general grounds, mappings that prefer nearest neighbors will be more likely.

The theory is now straightforward: The entropy $q(v)$ of a given velocity v is defined as $-\log p(v)$, where p is its probability. The maximum likelihood solution is the solution that minimizes the total entropy (just as in statistical mechanics), and we can find this simply by letting $q(v)$ be the "cost" of assuming velocity v and then discovering the mapping that minimizes the total cost. This is a linear problem that can be solved by a simple local network, in which one can incorporate additional penalties for deviations from one-to-one mappings if desired. The cost function is the affinity function that we discussed earlier, and the interactions of Figure 3–49 that produce the CS in effect find the minimum total cost, that is, the most likely

mapping given the statistics of the universe. This scheme generalizes naturally from the discrete case of successive frames to the continuous case, where the image is represented more as an incoming stream of tokens.

A critique of Ullman's theory

As a first attempt, this theory of the correspondence process is an extremely valuable contribution, and it provides a welcome and refreshing sip of clarity after the confusions and obfuscations of the preceding 50 years. Its importance is that it enables us to formulate a number of empirical questions that would not otherwise arise, and it opens the way for a rational investigation of the phenomenon rather than the confused cataloguing of its phenomenology.

Leaving aside the empirical aspects of the theory for the moment, there are a few points that should be raised, especially in a book whose primary business is the theory of the visual system. The first point is that the independence assumption, necessary for a probabilistic development, is empirically not quite true, at least in its simplest terms. In Figure 3–51(a) we do have independence—the unambiguous match of C_2 to R_2 does not affect the ambiguity of the behavior of C_1. In the situation shown in Figure 3–51(b), however, the behaviors of C_1 and C_2 are related—in fact, as Ullman pointed out, they behave as if they formed the endpoints of line C in Figure 3–51(c). They do not so behave when the induced grouping is different, as it is in Figure 3–51(d).

So it seems that the correspondence process can, to some extent, operate on groups as well as on their constituents. Although the grouping process does not involve explicit descriptions of the internal structure of the groups, and although matching between overall groups does not preclude additional matchings between their constituents, they can perhaps act to constrain those matchings. Specifically, matchings that are compatible with the grosser group matching are allowed, whereas those that are not are disallowed. This type of internal structure in a theory can be accommodated by a probabilistic framework, but it is awkward and indicates that we may not yet have found the most useful approach.

The second point we have already met—that correspondence may be established between groups without correspondences being established between their constituents. Ullman himself noted that this could happen (1979b, sec. 2.4.2), and more recent work with M. Riley has confirmed and extended this finding. Interactions such as these between higher-order units can, of course, be simply added on to the theory in the way that Ullman suggests, but they do not follow from it naturally and are not at all

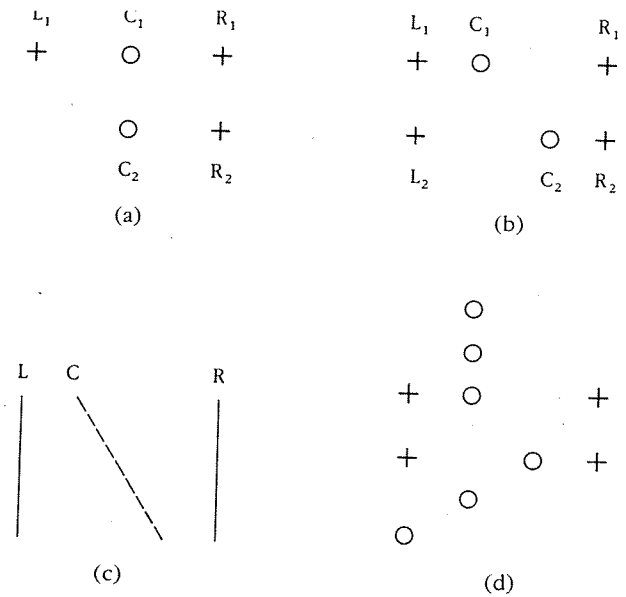


Figure 3–51. In this figure, frame 1 is shown with circles, frame 2 with crosses. In (a), the presence of C_2 does not affect the behavior of C_1. In (b) it does, however, the pair C_1C_2 acting like the line C in (c)—it goes either to L or to R. If the token configuration is disrupted by the presence of another organization as in (d), the central pair is no longer treated like the line C. (Reprinted from Shimon Ullman, *The Interpretation of Visual Motion,* by permission of The MIT Press, Cambridge, Massachusetts, fig. 2.20. Copyright © 1979 by The Massachusetts Institute of Technology.)

predicted by it. In fact, they run almost counter to it, since the whole thrust of the theory is to show how the sometimes confusing and complex behavior of the correspondence process on different patterns can arise from purely local interactions between simple processors that are associated with the constituents of the patterns.

For the third point, we need to adopt a slightly different perspective, that of the theory builder. What, we might ask, does the probabilistic approach buy? And the answer is, essentially, linearity. The practical consequence here is that purely local interactions are guaranteed to yield the global minimization that we seek. This is of great didactic value, because it shows that, as in our first cooperative stereo algorithm, the right global effects can be gained by purely local interactions. At first sight, this is exactly what we should try for, since, on the whole, the tangential connections in the cerebral cortex are known to be quite short (for example, Szentagothai, 1973).

Our experience with stereopsis and locally parallel organization has, however, warned us to beware of these arguments because of the problems associated with iteration. We must be careful here because Ullman's theory is not meant as an algorithm—it is a top-level theory—and there certainly are noniterative ways of implementing it. Nevertheless, the fact that it can be implemented with only local connections is an advantage only if it actually is implemented in this way. Unfortunately, if we take the theory at face value, which suggests an implementation, then I think a major objection must be that the rate of convergence for this type of calculation is slow—slower than, for example, the first stereopsis algorithm. To be sure, the rate depends on the starting point—and the rough grouping with large tokens could help here—but even so one would need, say, 10–70 iterations for reasonable convergence. This argument is not completely secure—one can usually patch up any particular convergence problem with special tricks to speed it up—but it does weaken the initial attractions of the theory being built around simple local network interactions.

The final point is much less easy for me to express, since it rests much more than the others on unsubstantiated intuition about how the brain works. Basically, my feeling is that at these rather low levels, probabilistic approaches such as the maximum likelihood principle are not used. Partly this feeling comes from having tried to use them myself a number of times—a probabilistic approach to stereopsis yields something like gray-level correlation, and I once tried to solve some problems related to the 2½-D sketch by using this approach—and partly from the general belief that a probabilistic approach is somehow not definite enough. For a problem of any complexity, the maximum likelihood solution is always pretty improbable (in the technical sense). Yet here the answers provided by the visual system are almost always correct and, moreover, are usually accompanied by a subjective feeling of certainty, rarely of doubt—much more certain and more often right than would be indicated by a rather low probability value. In similar situations, I have usually found that better constraints are available to describe how the world is put together, and these have often led to a much firmer basis for a computational theory.

In other words, if forced to answer the question posed at the end of the section on stereopsis—namely, Does this computational theory solve the right problem?—my answer would be more equivocal than it was for stereopsis or for the other half of Ullman's theory, the structure-from-motion problem. I do not yet have any very solid alternative, but the following remarks indicate the direction of my thoughts on this problem.

A new look at the correspondence problem

One problem or two?

The heart of any computational theory of a visual process is the answer to the question, What is the process for? In Ullman's framework, the goal of the correspondence process is to establish a relation between successive frames that allows measurements of the changes that have taken place. These measurements can then provide the input for subsequent processes that can recover the structures and their motions.

No doubt this is at least part of the job of the correspondence process, but is it the whole of it? Looking ahead a little, we shall see that the recovery of structure from motion incorporates (in an internally testable way) the assumption that the moving bodies are rigid. So we may ask about the correspondence problem first of all from the point of view of an observer in a world of moving, rigid bodies.

For small time intervals, the actual correspondence problem posed by this situation is essentially equivalent to the correspondence problem in stereopsis, because moving and rotating an object a little produces the same effect as moving and rotating one eye a little. Of course, different bodies may be moving in different ways, thus being equivalent to different pairs of eye positions, but the stereopsis matching theory is a local one, and it can be applied locally, provided that its assumptions are obeyed locally. These assumptions are that surfaces are smooth locally and matching is unique, because a given position always moves to only one other, and this nearly always means only one other in the image. Of course, some visible points will become invisible, and vice versa, but this is merely the analogue of the fact that, in stereoscopic depth changes, one eye can see parts of the surface that the other eye cannot see.

What, then, about the splitting and fusion phenomena of apparent motion, in which a single element in one frame splits to match two in the next (or conversely)? These are strong and well-known phenomena in apparent motion and have caused considerable theoretical problems. How often ought they to arise in the structure-from-motion situation? We have already seen that they can arise in stereopsis, both physically, in the rare instance that two surface markings that are distinct from one eye happen to lie along the line of sight from the other, and psychophysically, in Panum's limiting case. We have even seen from Braddick's stereograms of Figure 3–19(b) that the human visual system is very catholic about accepting double matches, provided that they are unique from one eye. But the reasons there were not fundamental ones; they had to do with the implementation and arise basically because the uniqueness condition is so strongly satisfied

by the physical world that the visual system can afford to assume that it holds without internally checking it.

Are the splitting and fusion phenomena of apparent motion of the same kind as in the stereo correspondence problem, or are they more fundamental? I think that if we are committed to the view that the sole function of the motion correspondence process is to solve the problems produced by rigid bodies in motion, then this problem can be solved in the same way as the stereo correspondence problem, which is equivalent. These phenomena would have to be explained away much as the examples of Panum's limiting case were in stereopsis.

However, this approach is not very satisfactory. One rather subjective reason is that the kinds of stereopsis achievable by the matching of pure texture edges are so rivalrous (see, for example, Mayhew and Frisby, 1976) and the impression of depth from them so poor that one has the feeling that "real" stereopsis is not happening at all—only some vague preliminary parts of it are (perhaps the vergence control system). In apparent motion, however, impressions are not at all as vague—such edges are clearly seen in motion with respect to one another. The matching that is obtained between pairs even as dissimilar as those in Figure 2–34 is quite clear and definite, and not at all rivalrous, as it is in stereopsis.

Another argument, which I find quite compelling, comes from a report by Ramachandran, Madhusudhan, and Vidyasagar (1973) that apparent motion can be established between subjective contours and even between disparity edges in a random-dot stereogram. This is almost a paradox from our narrow point of view, because if disparity edges have already been obtained, then we already have the three-dimensional structure, so why initiate this whole structure-from-motion process in order to obtain it?

Our narrow point of view must, I think, be inadequate—one simply cannot understand the motion correspondence process in so confined a way. How, then, is this process essentially different from the stereo correspondence process?

The crucial difference is that one is in space, and the other is in time. For rigid bodies the processes are equivalent, but for pliable surfaces they are not. The shape of an object from the left eye is always the same as its shape from the right eye at the same instant, but its shape a moment later may be different. This is not an uncommon phenomenon at all. A distant bird, for example, changes its shape and appearance very rapidly, both because it is not rigid and also perhaps because the sun catches its beating wings at one particular angle. The bird's image may be quite small and difficult to decompose into roughly rigid components. Nevertheless, although its motions may yield little or no direct clues about its structure, there is no doubt that the changing appearances are all related to one bird.

In other words, time introduces an important new factor, which is rather independent of the precise details of an object's three-dimensional structure. This factor is the *consistency of an object's identity through time,* and it is a different problem entirely. To see this difference, simply consider Ullman's (1977) example of the frog changing into the prince. This is not part of the structure-from-motion problem, because the structure changes, but it *is* part of the object identity problem.

My argument is that the theory should consider the two problems separately, because they have somewhat different computational requirements. The idea of matching disparity edges is inexplicable in the first approach but entirely explicable and almost obviously desirable in the second. For example, consider the patterns of light formed by the surface of a river playing upon a riverbed. The only constants here pertain to the geometry of the riverbed, and therefore we clearly need to be sensitive to just this, independent of its surface radiance. This type of situation may well be the real-life equivalent of Bela Julesz's random-dot "moviegrams," and this type of situation makes it quite comprehensible that we should be able to perceive them. If a fish should happen to glide leisurely by, transiently mottled by the changing patterns of light and dark falling upon it, it may be defined only by its disparity boundaries. These boundaries are moving, but it is the same fish all the time. That is a problem in object constancy.

Separate systems for structure and object constancy

Thus the problems introduced by time yield at least two rather distinct tasks for the correspondence process in apparent motion, and these are themselves distinct from the first of Braddick's two categories, which we discussed in Section 3.4. The first task is the first half of the structure-from-motion problem, and, in an environment of rigid, moving bodies, it is essentially equivalent to the matching problem in stereopsis. The only difference between the two is that a small rotation of one image is added in the motion situation, but this poses no important new problems. The aim, as in stereopsis, is to achieve a very detailed correspondence between accurately localizable items in the image, so that measurements of their position changes may be made to the (second-order) precision necessary for the structure-from-motion computations. In order to achieve this precision, one would expect the primitives used here to be rather low ones, like those in the raw primal sketch or perhaps even zero-crossings.

The goals of the second task are different, and they arise precisely because an object can change between two temporal viewpoints in a way that it cannot between two spatial viewpoints—it can change its shape and configuration (and even reflectance). Precision is not its goal; rough identity is—and this is the key to the difference between visual motion and stereopsis. There is no point to an approximate stereo correspondence by itself; it only has a point if it is a prelude to an exact match. Hence, approximate matches appear as indistinct and rivalrous perceptions. There is, however, a great deal of point to an approximate correspondence in time, since it offers a way of establishing object continuity.

My suggestion, therefore, is that two theories may be needed here, one for when the object is changing *and* moving and one for when it is only moving. The first should use everything it possibly can, including high-level primitives with catholic matching rules and any three-dimensional information that is already available. The phenomena of subjectively smooth motions may even be more concerned with the first system than with the second, since smoothness goes perceptually hand in hand with object constancy, and we know from Attneave's work that smoothness involves three-dimensional perceptual distances. The second system is at a lower level, computationally equivalent to stereopsis, and although it may not be implemented in the same way, zero-crossings may be worth looking at in this regard.

Structure from Motion

The problem

We have already seen from Ullman's (1979a) counterrotating cylinders experiment, illustrated in Figure 3–52, that both the decomposition of a scene into objects and the recovery of their three-dimensional shapes can be accomplished when the only available information is that afforded by their changing appearances as they move. Each frame in that demonstration consists of an apparently random collection of dots and is by itself uninterpretable. Only when shown as a continuous sequence does the movement of the dots create the perception of two counterrotating cylinders.

We shall therefore consider the simplified problem of how to interpret a sequence of frames, each composed of a set of random dots. In real life, the frames will contain more elaborate primitives than dots, but, just as in the case of stereopsis, the bones of the problem can be expressed in this simple form. Furthermore, we shall assume that correspondences have already been established between successive frames by the correspondence process that I discussed above. In fact, we shall need only the simpler

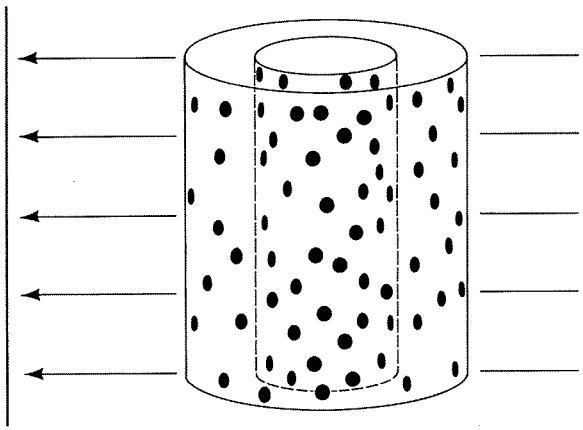


Figure 3–52. Ullman's rotating cylinders demonstration. Dots painted on the two cylinders are projected orthographically onto a screen as indicated by the arrows, giving a sequence of frames like those illustrated in Figure 3–53. Each single frame has the appearance of a set of random dots, yet when seen as a movie, the rotating cylinders are clearly visible.

sort of correspondence process, the one for rigid objects, which we saw was computationally equivalent to the correspondence problem for stereopsis.

Thus the problem posed here is a set of data like that shown in Figure 3–53. Each frame consists of a set of labeled dots (though the labeling is not shown in the figure), where dot A in frame 1 corresponds to dot A in frame 2, and so forth. The question is, How can we make sense of these data? How should we go about a sensible three-dimensional interpretation?

The difficulty here is exactly like the one we met in the stereopsis problem, namely, that the solution is underdetermined. There are an infinite number of three-dimensional configurations that could give rise, through orthographic projection, to the images of Figure 3–53—any number of different, randomly changing snowstorms, for example. But we do not see any of these different possibilities; we see only one, and it is the correct one.

Just as in stereopsis, therefore, we must be bringing additional information to bear on the problem that constrains the solutions one finds. This additional information must at the same time be powerful and true but rather unspecific. Powerful because it forces a solution that is usually unique; true because not only does one perceive only one solution, but

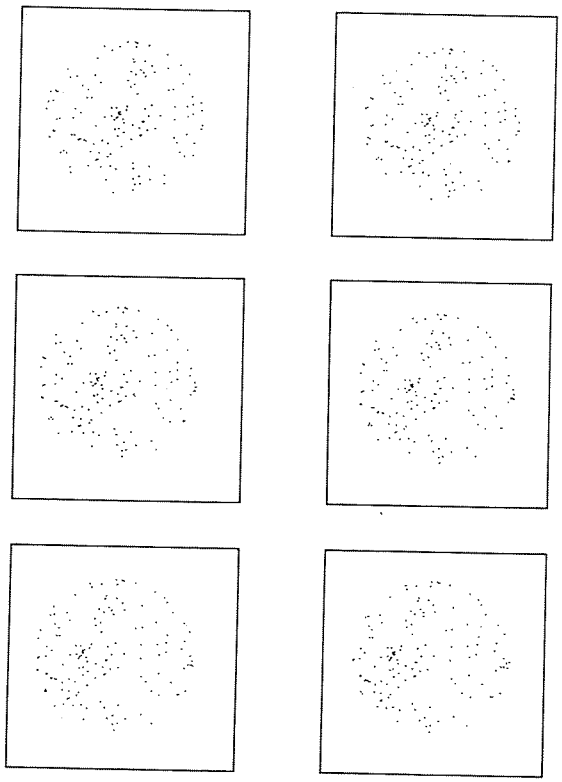


Figure 3–53. The structure-from-motion problem. This set of frames contains three-dimensional information (see Figure 3–52). How are we to recover it?

that solution is also the correct one physically; and unspecific because the system works in unfamiliar situations, without specific a priori knowledge of the shapes to be viewed.

A previous approach

Although there have been a number of previous approaches to this problem, only one of them deserves comment. It originated with Helmholtz (1910; Braunstein, 1962; Hershberger and Starzec, 1974) and initiated the idea that motion and stereopsis are analogous: Specifically, recovering structure from motion is analogous to recovering distance from disparity.

different parts of the visual field can be engaged in quite different motions. Now for the correspondence problem this does not matter, since that is essentially a local process. We have already made use of the fact that, for rigid objects and short time intervals, the two correspondence problems are in fact equivalent. We noted, however—without worrying particularly—that two different local motions would induce two different eye-pair positions to produce the equivalent stereo correspondence problem. The reason why this is not at all worrisome is that for correspondence the combination rules do not depend upon the precise position of the eyes. They have only to be close together and so have similar views. Hence, correspondence is unaffected by the fact that different portions of the visual field effectively induce different equivalent eye-pair positions.

Not so for the recovery of depth from disparity, however. As we saw, this depends critically on the effective interocular distance δ, and the induced δ's are in general different for each differently moving rigid object. There is no way of deducing their values a priori, and since they change, there is no way of comparing what is happening in one part of the visual field with what is happening in another. Hence, although this approach is actually valid for the correspondence problems in the two domains (provided one restricts oneself to rigid motions and short time intervals), it is not valid at all for the recovery of three-dimensional structure.

It follows from these arguments that changes in velocity in the visual field (which are the analogues of changes in disparity) should not yield

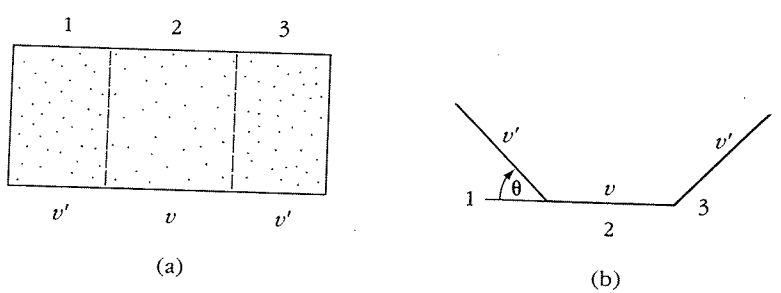


Figure 3–54. The conveyor belt demonstration. The dots in regions 1 and 2 move to the right with speed $v' = v \cos \theta$, and those in region 2 with speed v. However, the observer of (a) does not perceive the geometrical configuration (b). Instead, all of the regions appear in the frontal plane, and the dots appear to move faster in region 2. (Reprinted from Shimon Ullman, *The Interpretation of Visual Motion*, by permission of The MIT Press, Cambridge, Massachusetts, fig. 4.2. Copyright © 1979 by The Massachusetts Institute of Technology.)

direct impressions of depth, nor should common velocities be necessarily very useful for grouping. The Gestaltists, for example, had the notion of "grouping by common fate," which included grouping by common velocity, and Potter (1974) recently revived a form of this idea. However, the counterrotating cylinders demonstration includes points having the same velocity that belong to different cylinders. Evidence against the other half of the conclusion, that changes in velocity should yield changes in the impression of depth, is provided by Ullman's conveyor belt demonstration, illustrated in Figure 3–54. Dots in regions 1 and 3 have velocity v', and in region 2 they have velocity v. One does not perceive the different sections as planes at different depths or even as being arranged in the configuration of Figure 3–54(b). Instead, the dots all appear to be in the same frontal plane; they appear to speed up as they pass from region 1 to region 2 and to slow down again as they pass from region 2 to region 3.

The rigidity constraint

Most of the structures in the visual world are rigid or at least nearly so. This has been noticed by many students of motion perception (for example, Wallach and O'Connell, 1953; Gibson and Gibson, 1957; Green, 1961; Hay, 1966; Johansson, 1964, 1975), who formed the opinion that rigidity plays a special role in the problem. What they failed to realize, and what Ullman pointed out, was that searching for rigid interpretations is not merely a bias of our motion perception machinery; it enables us to solve the structure-from-motion problem unambiguously, without the need for any other constraining influence. This remarkable fact follows from a piece of mathematics that Ullman called the structure-from-motion theorem. It states that given three distinct orthographic views of four non-coplanar points in a rigid configuration, the structures and motions compatible with the three views are uniquely determined, up to a reflection where the closer points become the more distant ones. In other words, three views of four non-coplanar points suffice to determine their three-dimensional structure, provided that the correspondence problem has already been solved. Again, this result is not restricted to apparent motion; in continuous motion, what counts as three views depends solely on the resolution of the underlying systems measuring the position changes in time.

The four-points–three-views combination of the structure-from-motion theorem is the minimal combination in the following sense. With just two views, any number of points can be constructed that have no unique three-dimensional interpretation (although some combinations fortunately will), so that in general two frames are not enough. With three frames, three points are again in general too few to yield a unique solution; one needs four points.

One can give a rough plausibility argument for four points and three views based on the number of degrees of freedom involved. Suppose that we label the four points *O, A, B,* and *C*, the point *O* always corresponding to the origin (0, 0, 0), and let us label the three views 1, 2, and 3. There are 15 variables to be determined. Nine of them determine the three-dimensional positions in view 1 of *A, B,* and *C* relative to *O* (three points with three coordinates for each one), and the remaining six determine the three-dimensional rotations needed to obtain views 2 and 3 from view 1. (We rule out translations by superimposing the point *O* in each view). It takes three variables to specify a three-dimensional rotation, two to specify the axis, and one to specify the amount.

The amount of information we gain from each view is 6 relations, the two-dimensional coordinates of each of *A, B,* and *C*. (The point *O* is always [0, 0].) Hence, two views give us 12 relations, fewer than the 15 unknowns and so insufficient to determine the structure. Three views give us 18 relations, which exceeds 15 and so will be sufficient provided that there are not too many singularities or internal dependencies. The difficult part of the proof lies in showing that the 18 relations are in fact independent. The fact that there are 18 relations and only 15 unknowns means that there is some information left over, and this is ultimately what allows one to test internally the hypothesis of rigidity.

The rigidity assumption

In our analysis of the use of directional selectivity to infer properties of the visible surfaces, we saw that lines of discontinuity in motion direction cannot arise by accident. They have to mean the presence of a boundary between two incompatibly moving surfaces. In our analysis of the stereopsis problem, we saw that the constraints of uniqueness and continuity guarantee that a solution exists and is unique, and this theorem formed the basis for stereo analysis, since it allowed us to formulate and rely upon the fundamental assumption of stereopsis.

The same is true here. The structure-from-motion theorem, together with the general truth that most things in the world are locally rigid, allows us to formulate the fundamental assumption for the recovery of structure from motion. It was called the *rigidity assumption* by Ullman (1979a), and it states: *Any set of elements undergoing a two-dimensional transformation that has a unique interpretation as a rigid body moving in space is caused by such a body in motion and hence should be interpreted as such.*

The structure-from-motion theorem tells us that if a body is rigid, we can find its three-dimensional structure from three frames (up to a reflection, because we are dealing with the orthographic projection). If it is not rigid, the chances of there being an accidental rigid interpretation are vanishingly small, so in practice, the method will fail. The method is therefore self-verifying, and we know that if we can find a three-dimensional structure that fits the data, it is unique and correct. The proof of the theorem is constructive and enables one to formulate a set of equations whose solution yields the three-dimensional structure if it exists.

It is easy to implement this scheme, because it requires only four points as input data and so can be run in parallel independently throughout the visual field. This makes the scheme a particularly attractive candidate for understanding how human motion perception works. However, the particular algorithms suggested by directly applying the methods used in the proof of the theorem are not biologically plausible. They do not, for example, satisfy all the guidelines that I set out in Section 3.1—in particular, the principle of graceful degradation. Simply setting up the equations and solving them provides an algorithm that is far too rigid. If the data are inaccurate or if the viewed object is not quite rigid, this method will fail and give no help.

What is wanted is an algorithm that degrades gracefully in at least two senses. First, if the data are noisy but more than three views are available, the algorithm should be able to deliver an account of the structure that is at first rather rough but which becomes increasingly accurate as more views and hence more information are presented. And second, if the viewed object is not quite rigid, the algorithm should be able to produce the not-quite-rigid structure, perhaps again at the price of needing more points or more views to work on. Algorithms with this kind of robustness are being developed at our laboratory.

Until a particular algorithm has been developed as a candidate for the one actually used by our visual systems, and until the consequent psychophysical and neurophysiological experiments have been carried out, we shall not know for sure whether this approach to motion perception is appropriate. One thing, however, is certain; we now know what the important experimental questions are. Until Ullman took a computational approach to the problem, we did not know.

A note about the perspective projection

It is thought that algorithms for decoding the perspective, rather than the orthographic projection, are not part of the human visual system. The underlying reason is probably that the changes between frames are usually small already, and the differences between the changes seen by the two projections are usually very small indeed. The psychophysical evidence is that receding motion, which gives rise to changes only in the perspective