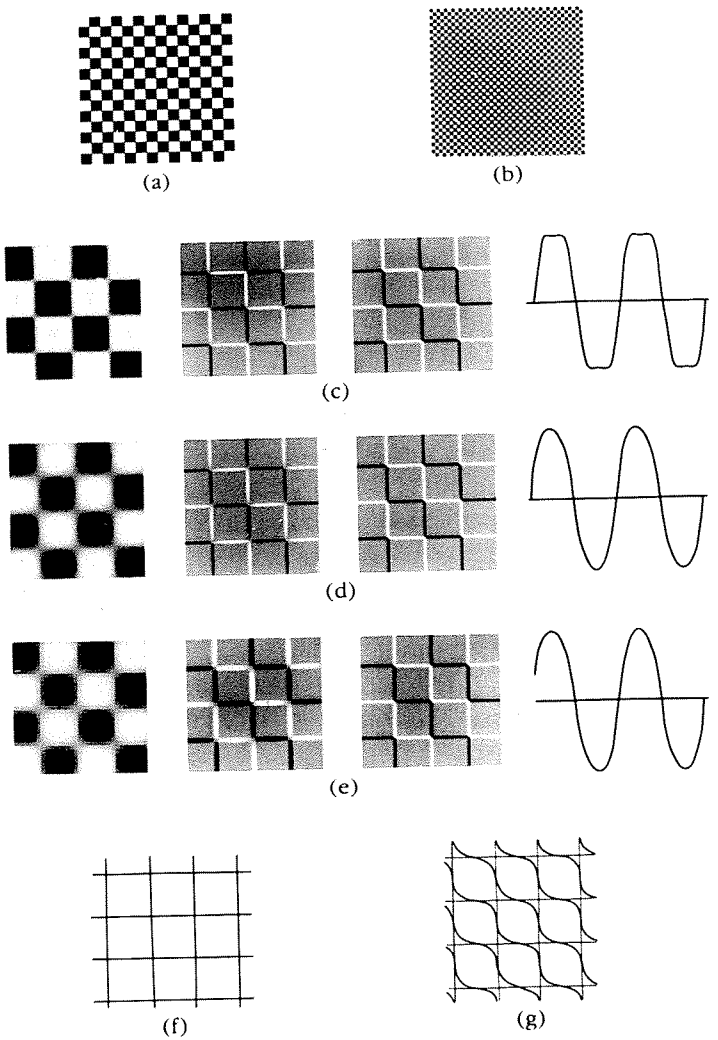
of the physical world does not allow us to deduce, for example, that a low-pass-filtered image contains the important information about how the world is physically and spatially arranged at that scale. We can see how this could be so from the chessboard of Figure 2–24. One important aspect of the organization of this image is that the black and white squares line up horizontally and vertically as well as diagonally. To be sure, the approach of low-pass spectral filtering can tell us about the diagonal organization but not about the horizontal and vertical, and mechanisms for detecting the horizontal and vertical arrangements (making tokens for the squares and noticing how they group) will also find the diagonal organization. So the filtering approach is both unnecessary and deficient.

Another example is provided by the herringbone pattern of Figure 2–2. The vertical organization of the stripes is a clear example of an important spatial organization, yet it cannot be detected by Fourier methods because there is no power in the vertical orientation. However, such organization is easily detectable by methods that take a spatial, physical approach, starting with a representation of the basic intensity changes and then using grouping procedures based on similarity, spatial proximity, and arrangement to work up from there (Marr, 1976). Mayhew and Frisby (1978b) were among the first to appreciate the importance of this point, and they adduced further evidence in its support in experiments that explored our ability to perform texture discrimination tasks. I shall return to their work later on.

Finally, let us consider some evidence that terminations are made explicit at this stage and that they are important. I feel that it is a good thing

---

Figure 2–24. (opposite) The Fourier spectrum of a chessboard pattern (of infinite size) has all its power in the diagonal directions, and none in the horizontal or vertical. Yet in (a) we can see that the vertical, horizontal, and diagonal spatial organizations are all equally visible while in (b) the diagonal organizations are slightly more prominent. (c), (d), and (e) show the analyses of zero-crossings from $\nabla^2 G$ operators of sizes $w_{2-D} = 12$, 24, and 48 pixels, respectively, on a pattern whose block size is 24 pixels, thus giving a range of w values from half to twice the size of the squares. In the first column are the convolution outputs. The second column shows the zero-crossings, with slope displayed as intensity (light and dark intensities representing positive and negative contrasts). In the third column, all the zero-crossings are displayed at uniform intensity; finally, the fourth column provides a cross-section of the convolution output near the zero-crossing contours. (f) and (g) illustrate symbolically the description obtained by channels much smaller and much larger, respectively, than the block size and should be compared with the perceptions one obtains from the chessboards in (a) and (b)—notice, for example, the roughly diagonal organization we see in looking at (b).



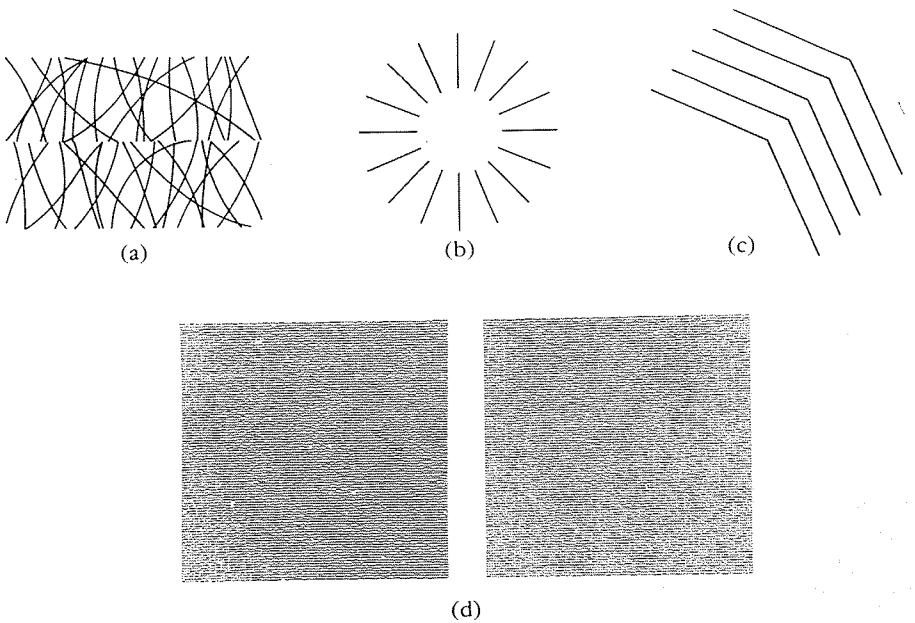


Figure 2–25. Examples of terminations being made explicit. In (a) and (b) subjective contours are constructed by joining termination points. In (c), points of discontinuity in orientation are seen to have a linear arrangement. In the stereogram (d), terminations or discontinuities in the small horizontal lines are probably being matched between the images to yield a square in depth. (Figs. (a), (b) reprinted by permission from D. Marr, "Early processing of visual information," *Phil. Trans. R. Soc. Lond. B 275,* 1976, figs. 9(a)(d). Fig. (d) reprinted by permission from B. Julesz, *Foundations of cyclopean perception,* University of Chicago Press, 1971, fig. 3.6-3.)

to give this information here because although edges, bars, and blobs are rather obvious things, terminations are much more symbolic and abstract. The reader may therefore need some additional persuasion that these things are indeed created and at a rather low level.

Figure 2–25 provides some examples on this point. We have defined a termination as a discontinuity in the zero-crossing orientation or as the termination point of a bar. Figures 2–25(a)–(c) show clear examples where such terminations line up and where it is difficult to think of methods for detecting this fact that do not make the actual positions of the discontinuities explicit. Figure 2–25(d), from Julesz (1971, fig. 3.6–3), is even more interesting, because the things that are being matched in this stereo pair are probably the small discontinuities in the horizontal lines,

and these images can be seen in stereo even when the discontinuities are tiny—less than 20 seconds of arc. Thus not only are such terminations used by stereopsis (as well as our being subjectively aware of them), but they are apparently used quite routinely even when the discontinuities are in the range of hyperacuity (smaller than a retinal receptor). The human visual system is an amazing machine!

2.3 SPATIAL ARRANGEMENT OF AN IMAGE

We come now to the question of representing spatial relations. Up to now, I have been content to assume that each item—each zero-crossing or each descriptive element of the raw primal sketch—has a coordinate in the image that determines its position there. This is reflected in our computer implementation by our use of a bit map of the image to represent basic positional information. That is, as in Figure 2–21(a), whenever there is a descriptive element, a two-dimensional array the size of the image has a 1 at the corresponding position. This 1 is also associated with a pointer to the element's actual description, which has the form shown in the legend to Figure 2–21. Like others before me, I have found that this rather literal representation, which is reminiscent of the topographically organized projections found in the early visual pathways, provides the most convenient starting point for examining geometrical relations in the image.

The reason for this is that there is quite a wide range of spatial relationships that needs to be made explicit in order to get at the useful information in an image. Once again we have the general point that these spatial relationships—things like density, collinearity, and local parallelism—are all implicit in the positions of each item, just as the binary decomposition of thirty-seven is implicit in its representation as XXXVII. But if that number's binary coefficients are necessary for some purpose, they must be made explicit at some point, so it would be advantageous to use the representation 100101.

A bit map is a good representation from which to start because it makes it relatively easy to limit the search of, for example, the raw primal sketch to just those elements in the local neighborhood of interest. Thus if we wish to know the density of certain elements in a circular neighborhood, we simply search that neighborhood in the bit map. When looking for collinear arrangements, we take a pair and search outward in the bit map along the two directions at roughly the specified orientation. The important point is that the bit map saves the trouble of searching through the whole list of primal sketch descriptors checking each coordinate to see

whether it falls within the specified neighborhood. The underlying reason why using a literal bit map representation of an image is more efficient is that most of the spatial relationships that must be examined early on are rather local. If we had to examine arbitrary, scattered, pepper-and-salt-like configurations, then a bit map would probably be no more efficient than a list.

It is not too hard to see the consequences of the bit map representation in terms of nerve cells. If a neuron is to measure the density of a particular type of token in a neighborhood of some fixed size, then provided that the neurons representing the tokens are roughly topographically organized, all our density neuron has to do is count how many of the token neurons are active. Similarly, if a neuron is to measure how much local activity is present at a particular orientation, then provided that the neural representation has a roughly topographical organization, the "oriented-activity neuron" need only count how many neurons tuned to approximately the orientation in question are active within a particular physical neighborhood of the cortex. Of course, if this physical neighborhood is circular, then the neighborhood in image coordinates will not be exactly circular, but it will be roughly so, which is usually good enough.

The reason for laboring this point is that many people have difficulty relating the idea of an x,y-coordinate system of the type that might be used in a computer program to the sort of thinking that must be employed for neurons. I suggested earlier that relating this idea need not be too much of a problem, and I hope it is now clear that at least for certain aspects of local geometry, notions based on rough topographical representation and locally connected receptive fields can provide machinery of adequate power. The other half of the game, the rather precise representation of particular local geometrical relations, is something we turn to now.

The critical question is, What spatial relations are important to make explicit now, and why? The answer to this, of course, depends on the purpose for which the representation is to be used. For us, the purpose is to infer the geometry of the underlying surfaces, and we can use the physical assumptions formulated in Section 2.1, together with the natural consequences for an image of changes in depth and surface orientation. This leads us to the following list of image properties, whose detection will aid the task of decoding surface geometry:

1. Average local *intensity,* from the first physical assumption (changes in average intensity can be caused by changes in illumination, perhaps due to changes in depth, and by changes in surface orientation or surface reflectance).

2. Average *size* of items on a surface that are similar to one another, in the sense of the second and third physical assumptions (the term *size* includes the concepts of length and width).

3. Local *density* of the items defined in image property 2.

4. Local *orientation,* if such exists, of the items defined in image property 2.

5. Local *distances* associated with the spatial arrangement of similar items (the third and fourth physical assumptions), that is, the distance between neighboring pairs of similar items.

6. Local *orientation* associated with the spatial arrangement of similar items (the third, fourth, and fifth physical assumptions), that is, the orientation of the line joining neighboring pairs of similar items.

From a representational point of view, the three broad ideas that we need here are (1) tokens to represent items, and we have already seen that they form one of the pillars of the primal sketch; (2) the notion of similarity between these tokens, and this we have also already encountered (in Figure 2–3 for instance); and (3) spatial arrangement. This last idea has two parts. The one that we have encountered already has to do with density measures of various kinds, and these can be made by counting things in neighborhoods; this gives us image properties 3 and 4 above. But image properties 5 and 6 require a new idea, a new representational primitive on which we can base the analysis of the local configurations of tokens. The information that needs to be made explicit here is the distance between and relative orientation of two similar tokens. To do this, I propose a primitive called the *virtual line,* which is constructed between neighboring similar tokens and has the properties of orientation and length. It also indicates somewhat the way in which the two tokens it joins are similar, so that virtual lines joining two pairs of dissimilar tokens are treated as dissimilar (in the sense of the third physical assumption).

Perceptually, virtual lines are not meant to correspond to subjective contours, although they may be their precursors. Subjective contours, in this theory, are a later construct. They are made in the 2½-D sketch, part of whose business it is to make explicit discontinuities in the distance of visible surfaces from the viewer. Virtual lines, on the other hand, are concerned with representing the organization of images, not surfaces. They are what enables us to see the flow in the Glass patterns (see Figure 2–3) or to see the different rivalrous organizations of Figure 2–5.

The notion of a virtual line is very attractive from a computational point of view, and Stevens (1978) undertook his study of Glass patterns to try to acquire some evidence for the psychophysical existence of such lines

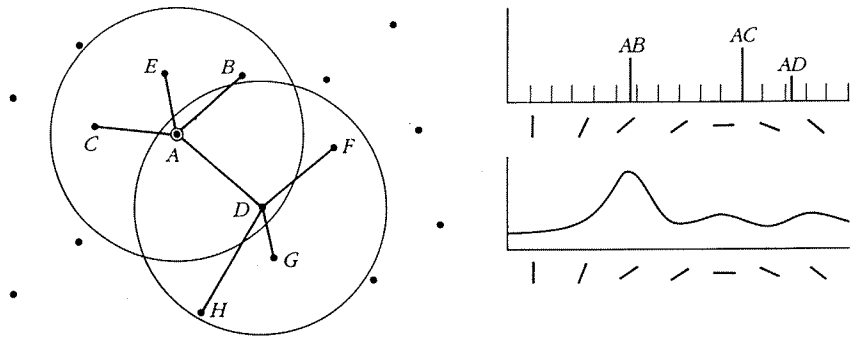
and also to explore the idea of tokens in the images—the supposed entities that virtual lines were thought to connect.

Stevens' study was extremely interesting, for in the space of one short experimental investigation he was able to make seven fascinating points, several of them quite unexpected:

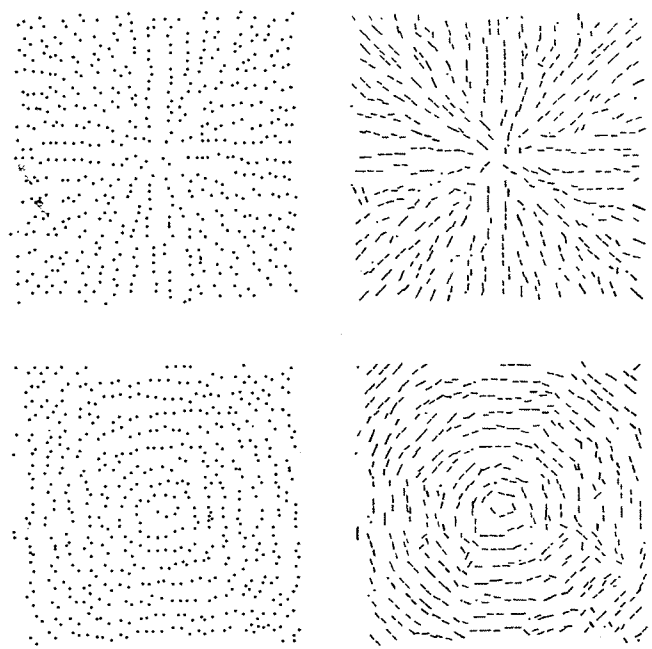
1. The local orientation organization in a Glass pattern can be recovered by a purely local algorithm, illustrated in Figure 2–26. The basic idea is to connect neighboring points with virtual lines and then to search locally among these virtual lines for the predominant orientation. By splitting patterns into several portions, each having a different transformation (see Figure 2–27), Stevens showed that perception of the global gestalt, contrary to Glass' (1969) suggestion, is not necessary for recovery of the local orientation.

2. If our perceptual analysis depends, like Stevens' algorithm, on the analysis of the distribution of orientations of virtual lines joining together dots in the pattern, the virtual lines are created between only nearby dots. The reasons for this are twofold; first and more obvious, the predominant local orientation changes as one moves globally over the pattern; second and not quite so obvious, the more virtual lines one creates from each dot, the more random the orientation distribution becomes locally and the finer must be the buckets in the histogram of the local orientation distribution that is being used to discover the predominant local orientation. If

Figure 2–26. (opposite) (a) Stevens' algorithm for recovering the local orientation organization in a Glass pattern has three fundamental steps. Place tokens that are defined in the image are the input to the algorithm, which is applied in parallel to each token. Since, in the case of the Glass dot patterns, each dot contributes a place token, the first step is to construct a virtual line from a given dot to each neighboring dot (within some neighborhood centered on the dot). A virtual line represents the position, separation, and orientation between a pair of neighboring dots. To favor relatively nearer neighbors, relatively short virtual lines are emphasized by means of a simple weighting function. The second step is to make a histogram of the orientations of the virtual lines that were constructed for each of the neighbors. For example, the neighbor *D* would contribute orientations *AD, DF, DG,* and *DH* to the histogram. The final step (after smoothing the histogram) is to determine the orientation at which the histogram peaks and to select the virtual line (*AB*) closest to that orientation as the solution. (b) The results (on the right) of applying the algorithm to the patterns on the left. (Reprinted by permission from K. A. Stevens, "Computation of locally parallel structure," *Biol. Cybernetics 29* (1978), 19–28, figs. 4, 5.)



(a)



(b)

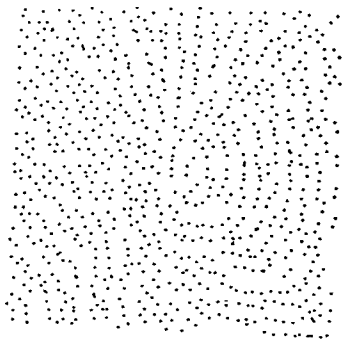


Figure 2–27. The algorithm used by our visual systems for detecting the local orientation structure is also a local one, as one can see from this pattern. Different portions of this pattern have different local orientation structures, and this can easily be discerned. (Reprinted by permission from K. A. Stevens, "Computation of locally parallel structure," *Biol. Cybernetics 29* (1978), 19–28.)

the orientation is analyzed to an accuracy of 10°–15°, then not more than about four virtual lines can be made, on the average, from each dot. Stevens also established that more than one virtual line is made. In a personal communication, he showed that only two have to be made.

3. The phenomenon scales linearly over a range of densities covering two orders of magnitude.

4. The idea that virtual lines join abstract tokens which can be defined in several ways is supported by examples like Figure 2–28, in which one of the sets of dots is replaced by small lines having randomly chosen orientations.

5. The tokens do, however, have to be reasonably similar in order for the analysis to succeed—in our terms, in order for the virtual lines to be inserted (Figure 2–3; Glass and Switkes, 1976). Stevens' own example of this, which I described in Section 2.1, consisted of three superimposed dot patterns, two dim and one bright. We see only the organization inherent in the dim dots. This is evidence both for the idea of tokens and for the notion of similarity. It proves that even at this early stage (Glass patterns can be seen in under 80 ms even with random-dot presentations immediately before and after), the analysis of the image is being carried out in quite abstract terms.

6. Interestingly, if the short lines at the random orientations shown in Figure 2–28 are replaced by short lines having a common orientation, as in Figure 2–29, rivalry appears between the overall orientations due to the short lines and due to the structure of the Glass pattern—in our terms, between the orientations of the real and the virtual lines. This bears upon how more global analysis of the image is implemented and controlled.

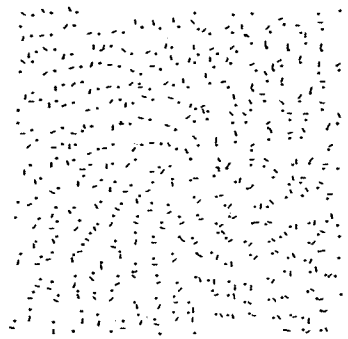


Figure 2–28. As we saw in Figure 2–3, the tokens in the two patterns do not have to be identical in order for their spatial organization to be apparent. They do, however, have to be similar. (Reprinted by permission from K. A. Stevens, "Computation of locally parallel structure," *Biol. Cybernetics 29,* 1978, 19–28).

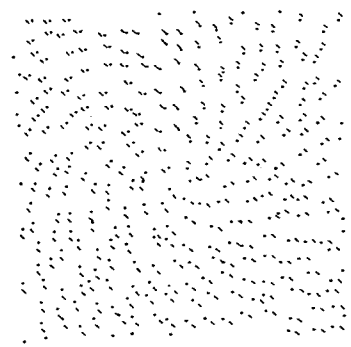


Figure 2–29. Here the superimposed pattern consists of small lines all having the same orientation. Interestingly, we perceive a kind of rivalry between this orientation and the orientation due to the spatial organization of the pattern. (Reprinted by permission from K. A. Stevens, "Computation of locally parallel structure," *Biol. Cybernetics 29,* 1978, 19–28.)

7. Finally, Stevens showed that there is little or no hysteresis in our perception of these patterns. The point at which the organization seems to disappear as the dot patterns are separated is very nearly the point at which the organization reappears as the patterns are brought together again. We were surprised by this. The reason we looked for it was Fender and Julesz's (1967) demonstration of a strong hysteresis effect in stereopsis. This had led Poggio and me to formulate a cooperative algorithm for the stereo matching problem, and the idea of cooperative processes as a way of writing an algorithm directly from constraints was an exciting one that was just emerging then (see also Zucker, 1976). The Glass pattern problem looked very well suited to a cooperative approach based on the constraints of the uniqueness and continuity of local orientation. Stevens' finding, however, showed that our perceptual systems probably do not employ a cooperative algorithm for this problem. Quite soon afterwards, we also realized that

our cooperative stereo algorithm was not the one used by our own visual systems and that matching was probably achieved by an algorithm involving very little cooperativity. Thus the opinion gradually formed that our visual systems do not use cooperative or purely iterative algorithms if it is possible to avoid them. I shall discuss some possible reasons for this later on.

Stevens' study left us somewhat more confident both about the questions we were asking and about some of the details of the primal sketch. At about that time Schatz (1977) argued that the raw primal sketch and virtual lines were by themselves sufficient to explain texture discrimination. The argument did not succeed, however, and to see why, we need to turn our attention to the more complicated levels of image representation that we call the full primal sketch.

2.4 LIGHT SOURCES AND TRANSPARENCY

Although the main stream of our account is concerned with spatial aspects of the image and visible surfaces, it is important not to forget that we are sensitive to other useful physical qualities of the visual world as well. One of these has to do with the detection of light sources—the subjective quality of fluorescence.

An important contribution to the visual detection of light sources was made by Ullman (1976b) in an article of characteristic elegance. He discussed six methods that the visual system might possibly use to help it detect light sources and then explored them empirically using achromatic "Mondrian" stimuli of the type introduced by Land and McCann (1971) in their study of lightness. These stimuli, named after the painter Piet Mondrian, consist of an array of rectangular shapes of black, gray, or white (as in Figure 2–30). In Ullman's display, one of these rectangles was sometimes a light source.

Ullman discussed light-source-detection methods based on the highest intensity in a field, high absolute intensity, high intensity compared with the average in the field, high contrast, and some other parameters. He found that none of these factors defined necessary conditions for the perception of a light source, though a contrast ratio of about 30:1 does provide a sufficient condition. High contrast is not, however, necessary; for example, a light source was perceived in a Mondrian where the ratio of intensities in no place exceeded 3:1.

Ullman then proposed a method based on the idea illustrated in Figure 2–31. In this figure, the x-axis represents distance along a surface illuminated from the right and which consists of three regions, A, B, and C. In A, the surface has reflectance r_1, and in B and C it has reflectance $r_2 < r_1$; in C there is also a source present underneath the surface. A camera looks down at the surface and records the intensity I at different points in the image, and the values of I have been plotted in the figure.

The idea behind Ullman's method is this: At the border between A and B, the intensity I changes and so does the intensity gradient ∇I, but they both change by the same amount so that the ratio $\nabla I/I$ remains constant.

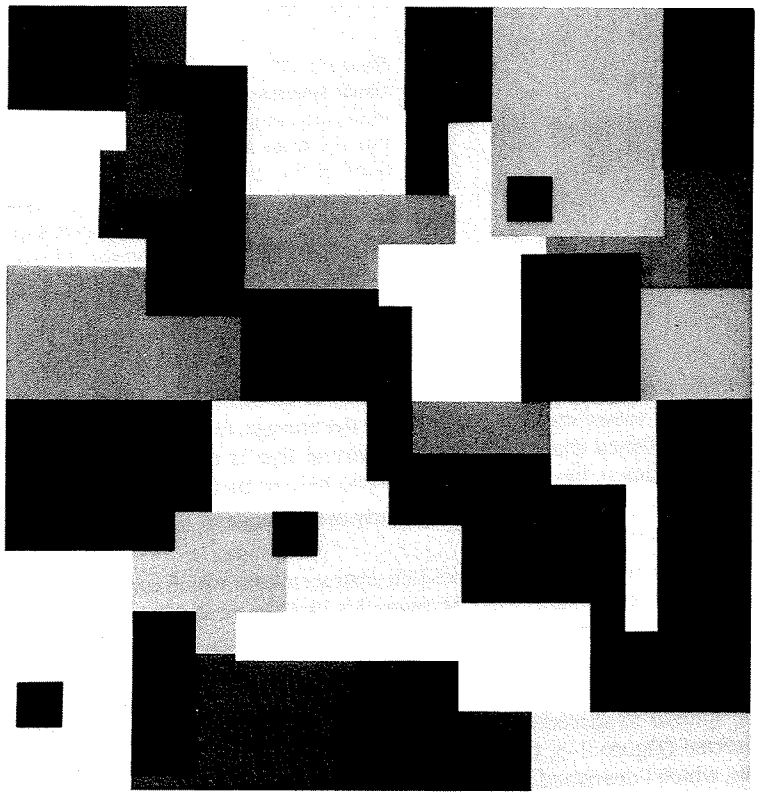


Figure 2–30. A Mondrian stimulus of the sort introduced by Land and McCann and used by Ullman in his study of fluorescence.

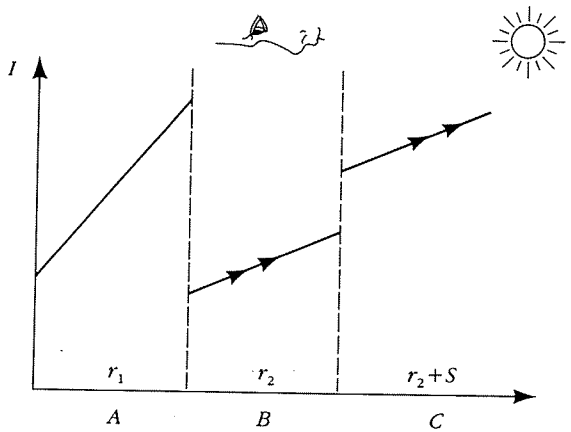


Figure 2–31. The idea behind the visual detection of light sources. Regions *A* and *B* have reflectances r_1 and r_2, and give rise to intensities *I* as shown. The value of *I* and of its gradient ∇I change together between *A* and *B*, so that $\nabla I/I$ remains constant. At *C*, however, a source *S* is added. This changes *I* but not ∇I, as shown. Hence the value of $\nabla I/I$ changes at a source boundary. This fact can be used to detect light sources in Mondrian images.

This is not so at the boundary between *B* and *C*, however, because here all that happens is that the constant-source value *S* is added to *I*. So *I* changes, ∇I does not, and hence $\nabla I/I$ does. So the ratio $\nabla I/I$ changes across a light-source boundary but not across a reflectance boundary.

This idea can be turned into a method for detecting light sources in the simplified Mondrian world, and Ullman satisfied himself that some such algorithm accounted for the perception of light sources in this environment.

Other Light-Source Effects

Forbus (1977) suggested that the operator $\nabla I/I$ could be applied to other illumination effects, including the detection of shadows and the various effects of surface wetness, luster, and glossiness that had so intrigued Beck (1972) and Evans (1974). For example, shadow boundaries behave like light-source boundaries with respect to the measure $\nabla I/I$. In addition, they are often, but not always, somewhat fuzzier than surface or reflectance boundaries, since the intensity change at a shadow is rarely sharp. This can be detected by comparing the slopes of the corresponding zero-crossings from the different-sized $\nabla^2 G$ filters, and a measure of the spatial extent of an intensity change is in fact incorporated into the raw primal sketch as the width parameter associated with an edge.

Glossiness is due to the specular or mirrorlike component of a surface reflectance function, so that one can treat the detection of gloss as essentially the detection of light sources that appear reflected in a surface (see Beck, 1972), and this depends ultimately on the ability to detect light sources. Forbus divided the problem into three categories: (1) the specularity is too small to allow gradient measurements; (2) both intensity and gradient measurements are available, but the specularity is local (as it is for a curved surface or a point source); and (3) the surface is planar and the source is extended. He derived diagnostic criteria for each case.

This topic, like the detection of shadows and light sources themselves, needs further study. The reason is that changes in surface orientation alone can also cause changes in $\nabla I/I$, although the orientation must usually change substantially in order to produce noticeable changes in $\nabla I/I$. This means that $\nabla I/I$ cannot be used as a pure diagnostic for illumination effects without taking changes in surface orientation into account. In preliminary studies we found that although in natural images one can find measurable changes in $\nabla I/I$ that are due to changes in surface orientation alone, most of these changes are small. And if one constructs an artificial image in which $\nabla I/I$ changes by a small amount across a boundary, one does not see it as a change in orientation. In fact, one sees nothing special until the change is quite large, at which point one begins to see one region as a light source.

Transparency

Another interesting phenomenon is transparency, which has attracted considerable popular attention. An example is the *Scientific American* article by Metelli (1974), in which he showed that one has the perception of transparency when a variety of inequalities hold in image intensities.

As one might expect, Metelli's inequalities might be deduced from the physics of the situation. Suppose a surface's reflectance changes from r_1 to r_2 along a boundary and that a sheet is overlaid in the manner shown in Figure 2–32. The effective illumination without the sheet is L_2, and with it (after being attenuated twice) L_1. Plainly, if the intensities in each quadrant are i_{11}, i_{12}, i_{21}, and i_{22}, as shown, we have

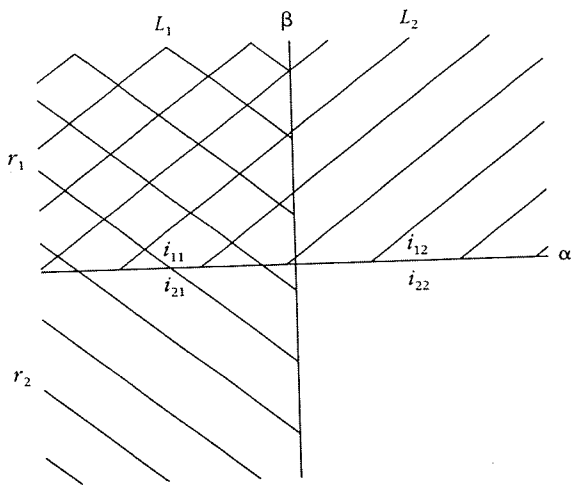


Figure 2–32. Boundary α represents a reflectance boundary, and β a transparency boundary. The quantities r_i represent reflectances; L_i, luminances; and i_{ij} are measured intensity values (for $i,j=1,2$).

$$\frac{i_{11}}{i_{21}} = \frac{i_{12}}{i_{22}} = \frac{r_1}{r_2}$$

and

$$\frac{i_{11}}{i_{12}} = \frac{i_{21}}{i_{22}} = \frac{L_1}{L_2}$$

These relations between the intensity values hold at transparency boundaries and at shadow boundaries; they do not hold at general four-way reflectance changes. Unlike shadow boundaries, however, transparency boundaries are almost always sharp (having a "width" of zero), and they do not cause a change in $\nabla I/I$.

Conclusions

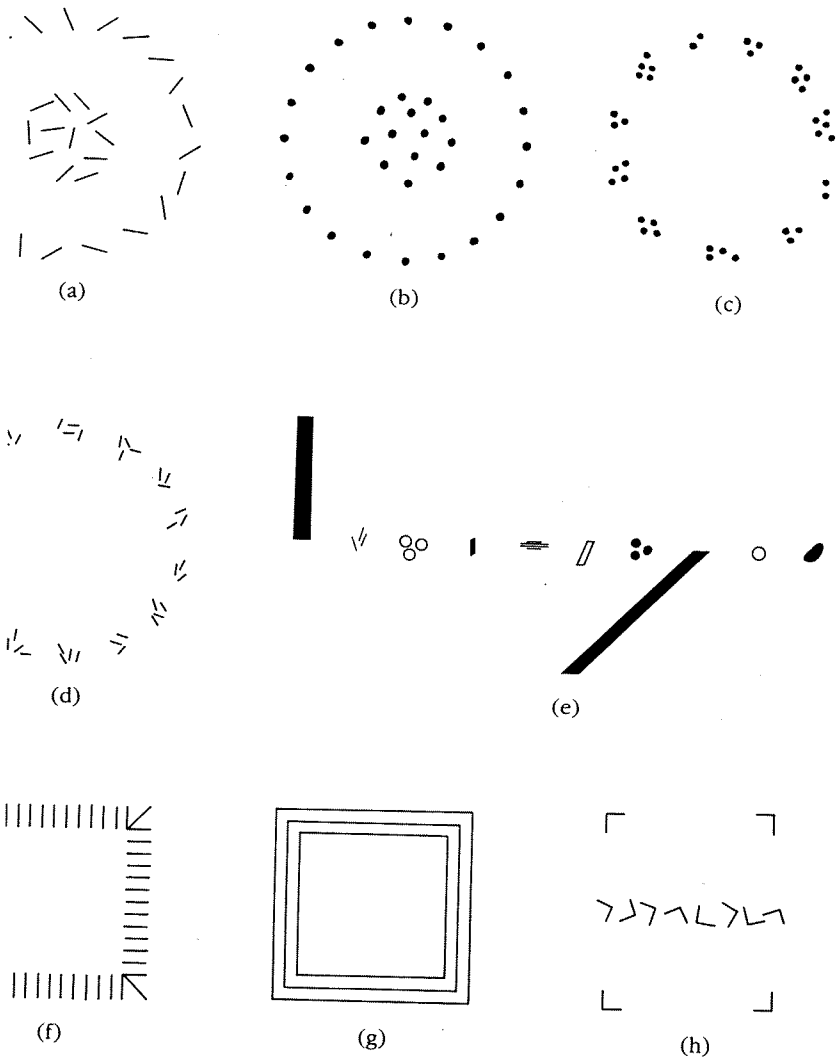
Although these studies are incomplete, they suggest that even quite abstract qualities of the physical world, like fluorescence and transparency, can be detected by early autonomous processes. From a representational point of view, this means that one can hope to include these qualities at an early stage, such as in the primal sketch boundaries. Additional primitives will be necessary to represent them, but this poses no great problem. It will be interesting to see what other qualities of the visual world can be detected at the same rather early level of processing.

2.5 GROUPING PROCESSES AND THE FULL PRIMAL SKETCH

Let us now resume our analysis of the spatial organization of images. There are two main goals to the analysis now; (1) to construct tokens that capture the larger scale structure of the surface reflectance function and (2) to detect various types of change in the measured parameters associated with these tokens that could be of help in detecting changes in the orientation and distance from the viewer of the visible surfaces. Roughly speaking, the goals are to make tokens and to find boundaries. Both tasks require selection processes whose function it is to forbid the combination of very dissimilar types of token, and both tasks require grouping and discrimination processes whose function is to combine roughly similar types of tokens into larger tokens or to construct boundaries between sets of tokens that differ in certain ways.

In general terms, then, the approach is to build up descriptive primitives in almost a recursive manner. The raw material from which everything starts is the primitive description obtained from the image that we called the raw primal sketch. One initially selects roughly similar elements from it and groups and clusters them together, forming lines, curves, larger blobs, groups, and small patches to the extent allowed by the inherent structure of the image. By doing this again and again, one builds up tokens or primitives at each scale that capture the spatial structure at that scale. Thus if the image was a close-up view of a cat, the raw primal sketch might yield descriptions mostly at the scale of the cat's hairs. At the next level the markings on its coat may appear—which may also be detected directly by intensity changes—and at a yet higher level there is the parallel-stripe structure of these markings. The whole description would then be organized somewhat as shown in Figure 2–7. At each step the primitives used are qualitatively similar symbols—edges, bars, blobs, and terminations or discontinuities—but they refer to increasingly abstract properties of the image.

Some examples of these primitives appear in Figure 2–7. Other examples are the bloblike groups in the centers of Figures 2–33(a),(b), the small



33. The essence of the higher primitives in the primal sketch is their ability to capture ıge of image items as a group or token and their ability to be arranged into groups and :s. These diagrams show some examples of the different ways of defining place tokens ɔuping them. In each one a small line, a group of lines, or a group of dots is being and treated as a single unit.

clusters in Figures 2–33(c), (d), the rather heterogeneous collection of items that make up the groups in Figure 2–33(e), the sides of the squares in Figures 2–33(f), (g), and the central line in Figure 2–33(h). Any kind of local cluster or blob or group, the ability to treat it as a single item—these are the fruits of this class of processes, the processes responsible for token formation. The representation of the three-dimensional angles between two lines or the notions of a square or triangle, for example, are not included in the repertoire of the primal sketch, since they concern properties of the real world that form the image, not of the image itself.

Once these primitives have been constructed, they can tell us about the geometry of the visible surfaces—either through the detection of changes in surface reflectance or through the detection of changes that could be due to discontinuities in surface orientation or depth. About the first type of detection, one can say virtually nothing, except to remark that at a change in the surface, the change in the reflectance function is usually so great that almost any measure will detect it. I shall therefore restrict attention here to the second—the detection of boundaries that might be caused by surface discontinuities. There are two rather different ways in which these boundaries can be detected; one is by finding sets of tokens that owe their existence to the physical discontinuity and are therefore organized geometrically along it. An example of this is the lining up of terminations or of discontinuities, as illustrated in Figures 2–25(a), (b). The machinery for finding such things, I think, is also responsible for the circles in Figures 2–33(a) through (d) or the line in Figure 2–33(e).

The second type of clue to surface discontinuity consists of discontinuities in various parameters that describe the spatial organization of an image. In the section before last, we isolated six image properties that are useful to measure, three of them intrinsic to a token—average brightness, size (perhaps length and width), and orientation—and three pertaining to the spatial arrangement of tokens—their local density, distance apart, and the orientation structure, if any, of their spatial arrangement. Changes in any of these will help us to infer the geometry of the visible surfaces, and by our second physical assumption, we shall want to measure such changes at a variety of scales.

Examples of this type of clue appear in Figure 2–34. Figure 2–34(a) shows a boundary that is due to a change in dot density. In Figure 2–34(b) it is due to the change in average size of the squares. In Figure 2–34(c) it is due to a change of 45° in orientation, and in Figure 2–34(d) several of these factors change.

Thus the point of the second type of task is to measure locally (at different scales) the six quantities we defined above and to make explicit, by means of a set of boundary or edge primitives, places where discontin-

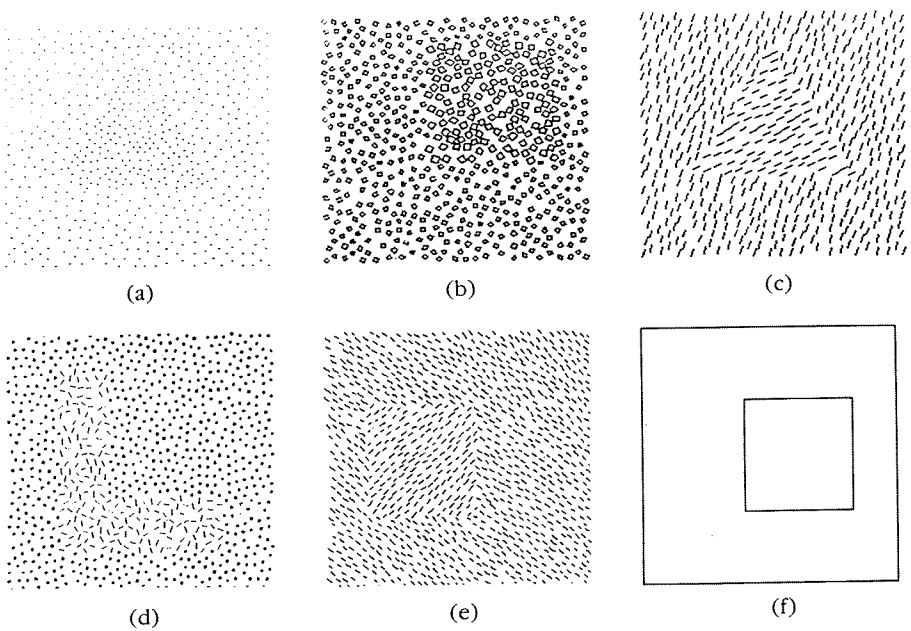


Figure 2–34. Another important aspect of the primal sketch is the construction of boundaries between regions on the basis of cues that could be caused by discontinuities in surface orientation or distance from the viewer. All examples in this figure are due to M. Riley, and they give rise psychophysically to boundaries in the sense defined in the text. The boundaries in (a) to (c) could be of geometric origin, but not in (d). Motion correspondence can be obtained between the boundaries in (e) and (f).

uities occur in these measures. The reason for adding such boundaries to the representation of the image is that they may provide important evidence about the location of surface discontinuities. This point of view has the important consequence that parameter changes likely to have arisen because of discontinuities in the surface ought to be those that give rise to perceptual boundaries, whereas those that probably could not have their origins traced to geometrical causes should be much less likely to produce perceptual boundaries. I call this the *hypothesis of geometrical origin for perceptual texture boundaries.* The principal limitations on its usefulness come from the fact that reflectance functions seldom have a precise geometrical structure. For example, if there is an oriented component to the surface structure, it is usually not very exact. Hence small changes in orientation in an image that may be produced by small changes in surface

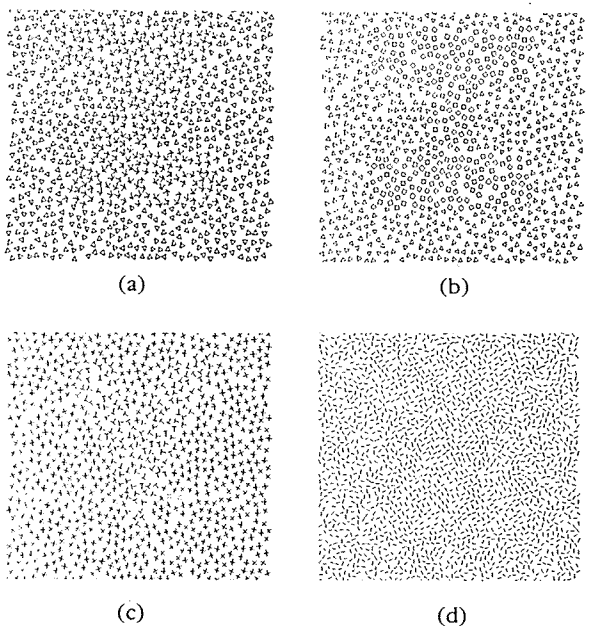


Figure 2–35. These examples, also provided by M. Riley, show texture differences that cc be of purely geometrical origin. They do not give rise psychophysically to boundaries in th defined in the text, even though we are sometimes able to say that one region differs from in some way. In example (d), the inner region contains lines of just two orientations, whe outer region contains lines of all orientations. It is interesting to contrast these examples wi of Figure 2–34.

orientation will not usually produce a clear signal. The same applies to changes in apparent size in an image, although density allows a more sensitive discriminant. Hence, only when an image structure is extremely regular would one expect to find high perceptual acuity for these discriminations. On the whole, we should be pretty bad at them—as indeed we are (see Figure 2–35).

Before summarizing this line of argument, I should perhaps make a final point. Although it is convenient to separate grouping processes into the two categories of token formation and boundary formation, they are not, in fact, quite separate, and the two categories can overlap. In Figure 2–7, for example, some of the dot-density boundaries are boundaries of tokens. The tokens could be constructed either from such boundaries or from the cluster of the cloud of dots there, or, of course, in both ways. In

Figure 2–34(a), the triangle could be made by the linear grouping of nearby dots, by finding a local increase in dot density, or even by a local decrease in average brightness. A single boundary is often defined in many ways, a fact of life that aids its recovery by the visual system but raises difficulties for the experimental psychophysicist.

Main Points in the Argument

The idea, then, is to start with the raw primal sketch and operate on it with processes of selection, grouping, and the discrimination to form tokens, virtual lines, and boundaries at different scales. The approach I have outlined gives the reasons for doing this: It enables us to deduce what types of tokens should be made, what types of selection and grouping should be available, which circumstances should give rise to perceptual boundaries and which should not, and perhaps even how to compare differences in acuity due to different discriminants. For example, when token size is viewed as a discriminant that indicates a change in surface orientation, the resolution of the analysis of token size should be comparable to the resolution of the analysis of token orientation. These arguments provide a physical basis for the suggestion that some types of visual discrimination of texture rest on first-order discriminations acting on the primal sketch (Marr, 1976). We now explore this question in more detail.

The Computational Approach and the Psychophysics of Texture Discrimination

From a purely psychophysical point of view, it has been difficult to define exactly what is meant by the phrase *texture discrimination*. In his well-known series of articles on the subject, Bela Julesz (for example, see Julesz, 1975) distinguishes between textures that can be immediately distinguished (so-called preattentive perception) and those that cannot be distinguished without close and often prolonged study (so-called scrutiny). He limited his investigations to discriminations of the first kind, those that can be distinguished in under 200 ms—roughly, those that can be distinguished without eye movements.

I should perhaps point out that the approach I have suggested to the problem is somewhat more restrictive, for it also requires that perceptual boundaries be formed at the borders between the textures. Not all of the textures devised by Julesz have this property. None of the examples in Figure 2–35 do, for instance, whereas all the examples in Figure 2–34 do. Psychophysically, then, our approach requires that the discrimination be made quickly—to be safe, in less than 160 ms—and that a clear psychophysical boundary be present. There are various criteria for this second requirement. One is that, in addition to being able to state that two textures are present in a Julesz display like those in Figure 2–34, one should also be able to give information about the shape of the distinguished region. Schatz (1977), for example, included this condition as one of his experimental criteria.

Another possibility, suggested to me by Shimon Ullman, is to try to obtain apparent motion between texture boundaries that have been generated in different ways in two frames. Frame 1, for example, might consist of Figure 2–34(e), and frame 2, presented after an interstimulus interval of, say, 100 ms, of Figure 2–34(f). If the boundaries appear to move in the obvious way, this is corroborating evidence that they are in fact constructed. If the boundaries obey the same local correspondence rules that are obeyed by intensity boundaries (Ullman, 1979b), this is then very strong evidence that the boundaries are being made explicit. The examples illustrated in Figure 2–34 all pass both the shape and apparent-motion tests.

A third criterion for when a boundary is being constructed perceptually may perhaps be developed from a finding by Kidd, Frisby, and Mayhew (1979). They found, using suitably constructed stereograms, that certain kinds of texture boundary are capable of initiating disjunctive eye movements, which are eye movements that cause the two lines of sight to converge or diverge.

If all these criteria succeed or fail together at the different types of boundary, we shall have a powerful technique for saying when a perceptual boundary is created from a change in visual texture. Similar combined approaches may also help us to determine whether something like the full primal sketch is in fact obtained from the image by telling us what types of tokens are made explicit in preattentive perception.

Finally, it seems to me that psychophysical studies of the relative power of the different discrimination processes can be most convincing if something like Barlow's (1978) absolute measures of efficiency are used. In this study, Barlow asked how sensitively humans could detect targets of greater dot density embedded in backgrounds of random dots. He found that his subjects were able to use about two-thirds of the objective signal-to-noise ratio of the displays, which corresponds to about 50% of the statistical information available. He also suggested an interesting, economical model to explain his results, consisting of "dot-number estimating" elements that are roughly circular and of variable size. They are sufficient in number to

cover the central area of vision with neighborhoods 1°–4° in diameter, and with an average mismatch and overlap of 50%. They integrate temporally for about 0.1 s. I hope that studies like this can be extended to other discrimination tasks.

That ends our discussion of how to represent an image. We now turn to the use of these representations in deriving surface information.

CHAPTER 3

From Images to Surfaces

3.1 MODULAR ORGANIZATION OF THE HUMAN VISUAL PROCESSOR

Our overall goal is to understand vision completely, that is, to understand how descriptions of the world may efficiently and reliably be obtained from images of it. The human system is a working example of a machine that can make such descriptions, and as we have seen, one of our aims is to understand it thoroughly, at all levels: What kind of information does the human visual system represent, what kind of computations does it perform to obtain this information, and why? How does it represent this information, and how are the computations performed and with what algorithms? Once these questions have been answered, we can finally ask, How are these specific representations and algorithms implemented in neural machinery?

The study of working visual systems can help us in this endeavor, and nowhere is this clearer than in the study of visual processes. At the level of computational theory, the investigator's first question is, What computational problems are being solved, and what information is needed to solve them?

As usual, the point is best made with an example. Because of how our eyes are positioned and controlled, our brains usually receive similar images of a scene from two nearby points at the same horizontal level. If two objects are separated in depth from the viewer, the relative positions of their images will differ in the two eyes. You can see that this is so by holding your thumb at various distances from your eyes against a background. Closing first one eye and then the other will then convince you that objects in the world have somewhat different positions in the images cast upon each of your retinas. The relative difference in position is called *disparity*; it is usually measured in minutes of arc, and the disparity between the images of your thumb and the background in your two eyes increases as you move your thumb nearer to you. One minute of disparity roughly corresponds to a depth difference of 1 in. for an object 5 ft away.

The brain is capable of measuring disparity and using it to create the sensation of depth. For purposes of demonstration, a stereoscope from a souvenir shop will do: When individual views are seen with just one eye at a time, they look flat. However, if you have good stereo vision and look with both eyes, the situation is quite different. The view is no longer flat: The landscape jumps sharply into relief, and your perceptions are clearly and vividly three-dimensional.

How does stereo vision work? Unfortunately, we cannot even begin to ask the right questions from just the evidence described above. The reason is that from the experience of everyday life or even from the small experiment with the stereoscope, it is not at all clear how separate stereoscopic processing is from the more familiar, monocular analysis of each image. If stereo processing were an isolated module, so to speak, then one could tackle it on its own. But it may not be isolated—for example, stereo vision could involve a complicated and gradually increasing interaction between the individual processings of each eye and a comparison of the results between the two eyes. This is not as absurd as it seems. It does not take much imagination to see how such a scheme might work. We could start by finding, for example, the images of an oak tree as seen independently by the left and right eyes. Then we could find the trunk in each image and then, perhaps, the lowest branch on the right hand side of the trunk. Pretty soon we would have correspondences between the small details of the left and right images whose disparity could be measured accurately. And because the match has been obtained in this general-to-specific way, there is never any real problem in deciding what should match what.

This type of approach, incidentally, is typical of the so-called top–down school of thought, which was prevalent in machine vision in the 1960s and early 1970s, and our present approach was developed largely in reaction to it. Our general view is that although some top–down information is

Figure 3–1. The interpretation of some images involves more complex factors as well as more straightforward visual skills. This image devised by R. C. James may be one example. Such images are not considered here.

sometimes used and necessary (see Figure 3–1 and Marr, 1976, fig. 14), it is of only secondary importance in early visual processing. The evidence for this comes from psychophysics and for some reason was willfully ignored by the computer vision community. The argument suggested by this evidence is a simple one. If, using the human visual processor, we can experimentally isolate a process and show that it can still work well, then it cannot require complex interactions with other parts of vision and can therefore be understood relatively well on its own.

One way of isolating a visual process is to provide images in which, as much as possible, all kinds of information except one have been removed and then to see whether we can make use of just that one kind. Bela Julesz did this for stereopsis by inventing the computer-generated random-dot stereogram, which we met in Figure 1–1. Both the left and right images shown there are computer-generated assemblies of black and white squares that are identical except for a centrally located, square-shaped region shifted horizontally in one image relative to the other. That

is, it has a different disparity. The stereo pair contains no information whatever about visible surfaces except for this disparity.

When the pair is viewed stereoscopically and fused, one vividly and unmistakably perceives a square floating in space above the plane of the background. This proves two things: (1) Disparity alone can cause the sensation of depth, and (2) if there is any top–down component to the processing (and, in fact, we think that there probably is a little), it must be of a very limited kind, because neither image contains any recognizable large-scale monocular organization.

This observation—which is qualitative rather than quantitative, not at all technical, and, like many of Julesz's demonstrations, absolutely and strikingly convincing to behold—is fundamental to our approach, for it enables us to begin separating the visual process into pieces that can be understood individually. Computer scientists call the separate pieces of a process its *modules,* and the idea that a large computation can be split up and implemented as a collection of parts that are as nearly independent of one another as the overall task allows, is so important that I was moved to elevate it to a principle, the *principle of modular design.* This principle is important because if a process is not designed in this way, a small change in one place has consequences in many other places. As a result, the process as a whole is extremely difficult to debug or to improve, whether by a human designer or in the course of natural evolution, because a small change to improve one part has to be accompanied by many simultaneous, compensatory changes elsewhere. The principle of modular design does not forbid weak interactions between different modules in a task, but it does insist that the overall organization must, to a first approximation, be modular.

From a theoretical point of view, observations like Bela Julesz's are extremely valuable because they enable us to formulate clear computational questions that we know must have answers because the human visual system can carry out the task in question. It was Julesz's findings that allowed us to formulate our theory of human stereopsis (Marr and Poggio, 1979). The analogous findings of Miles (1931) and of Wallach and O'Connell (1953) allowed Ullman (1979b) to develop his theory of structure from motion. Some other experiments by Julesz (1971, chap. 4), together with Braddick's (1974) identification of a short-range, short-term process in apparent motion, contributed to the formulation of our theory of directional selectivity.

The existence of a modular organization in the human visual processor proves that different types of information can be analyzed in relative isolation. As H. K. Nishihara (1978) put it, information about the geometry and reflectance of visible surfaces is encoded in the image in various ways

and can be decoded by processes that are almost independent. When this point was fully appreciated, it led to an explosion of theories about possible decoding processes. This chapter describes the computational theories of those decoding processes that are now quite well understood. These processes are (1) stereopsis, (2) directional selectivity, (3) structure from apparent motion, (4) depth from optical flow, (5) surface orientation from surface contours, (6) surface orientation from surface texture, (7) shape from shading, (8) photometric stereo (the determination of surface orientation and reflectance from scene radiances—the intensity of reflected light—observed by a fixed sensor under varying lighting conditions), and (9) lightness and color as an approximation to reflectance. Of course, other cues are available, like occlusion, but unless I have been able to give a process a reasonably integrated treatment, I have not discussed it here. Not all of the methods described here have biological relevance—photometric stereo certainly has none—but they are all of interest as ways of inferring the geometry and reflectance of visible surfaces from their images.

3.2 PROCESSES, CONSTRAINTS, AND THE AVAILABLE REPRESENTATIONS OF AN IMAGE

Before embarking on a detailed description of the different theories, I should make some remarks about the general nature of these theories and what the reader should look for in them and expect from them.

The first point is to remind the reader that we expect to analyze processes at three levels (remember Figure 1–4)—the levels of computational theory, of algorithm, and of implementation. Of course, the vision problem has not been completely solved yet, so we cannot analyze at all three levels every process within the human visual system. But we can analyze some processes at all three levels, and many of them at one or two—perhaps even most of the processes that discern surfaces from images.

In every case, we start with the first level—the computational theory—because this book is about the computational approach to vision. And at this level the reader should look out for the physical constraints that allow the process to do what it does. The situation is quite like what happened in Chapter 2. There we were dealing with ways of representing the image, and in order to say what would be useful and what would not, we were continually referring to the interaction between the imaging process and the underlying properties of the physical world that gives rise to structure

images. In this chapter, where we deal with processes instead of representations, the situation is entirely analogous but arises in a slightly different way. We have already met an example of this new situation in the theory of how to combine zero-crossings from different-sized filters in order to make the physically meaningful primitives of the raw primal sketch. The critical point was that, in general, there is no reason why the zero-crossings from two channels that do not overlap in the frequency domain should be related. They are related in early vision because intensity changes are caused by markings on a surface, the edges of objects, and so on, and these happen to have the critical property of spatial localization.

This interaction between the imaging process and the underlying properties of the physical world commonly occurs in the study of visual processes, and we shall meet several examples here. Frequently an apparently insoluble problem arises, such as which dots in the left-hand pattern in Figure 1–1 should match which dots in the right-hand pattern. From the image alone one just cannot tell. The critical step in formulating the computational theory of stereopsis is the discovery of additional constraints on the process that are imposed naturally and that limit the result sufficiently to allow a unique solution. Finding such constraints is a true discovery—the knowledge is of permanent value, it can be accumulated and built upon, and it is in a deep sense what makes this field of investigation into a science (Marr, 1977b).

Once we have isolated where the extra information comes from—in what ways, if you like, the information is constrained by the world—we can incorporate it into the design of a process. For combining zero-crossings, for example, this was done by the spatial coincidence *assumption*—that coincident zero-crossings are adequate evidence of a physical edge. Thus, the constraints are used by turning them into an assumption that may or may not be internally verifiable.

This, then, is one aspect of the top-level computational theory of a process, but there is another, almost as important. We saw in Chapter 1 that a process can be viewed as a transformation from one representation to another. Addition, for example, maps a pair of numbers into a number. All the processes that we shall discuss take as their inputs properties of the image and produce as their outputs properties of the surfaces—indicating to us either something about the geometry or the reflectance of the surfaces.

We shall discuss ways of representing the outputs of these processes in the next chapter, but now we are concerned with their inputs. What should serve as the inputs to these processes? We already have four options—the image itself, zero-crossings, the raw primal sketch, and the full primal sketch. Part of the computational theory must indicate which of these four should be used (or if something else entirely is appropriate) and why, and a portion of the investigation of each process will deal with this question.

Ultimately, of course, psychophysics tells us which input representation is used—if the process is in fact incorporated in the human visual system. There is, however, one useful point to bear in mind (Marr, 1974b): Essentially, since the constraints allow the processes to work, and since the constraints are imposed by the real world, by and large the primitives that the processes operate on should correspond to physical items that have identifiable physical properties and occupy a definite location on a surface in the world. Thus one should not try to carry out stereo matching between gray-level intensity arrays, precisely because a pixel corresponds only implicitly and not explicitly to a location on a visible surface.

This point is important. For example, failure to recognize it held Wallach and O'Connell (1953) up for years by their own admission. They could not understand why the shadow of a bent wire should be different from the shadow of a smooth solid object. If a wire is rotated, its shadow moves, and one instantly perceives the wire's three-dimensional shape; if a solid object is rotated, its shadow moves but one cannot perceive its shape. The reason is that the shadow of the wire produces an outline that is effectively in one-to-one correspondence with fixed points on the wire, each having a definite physical location that changes from frame to frame, admittedly, but that always corresponds to the same piece of wire. For the rotating object this is just not true. From moment to moment, the points on the silhouette correspond to quite different points on the object's surface. The image primitives are no longer effectively tied to a constant physical entity. Hence the shape recovery process fails.

On the other hand, the more complex the derivation of a representation from an image, the longer the derivation is liable to take. In real life, time is often of the essence; especially in the analysis of motion, an answer is required as soon as possible—before the image has become out-of-date or before the mover has eaten the viewer. In general, therefore, evolution is prejudiced toward getting things started as soon as possible.

Hence, although processes that operate on the information in an image could use any of a wide variety of input representations in principle, in practice they are likely to use the earliest representations that they possibly can. The range that we have discussed includes the gray-level image, zero-crossings, the raw primal sketch, and the full primal sketch. The earlier ones are not yet "physical," and so a bit unsafe, which might cause us to make mistakes. But for some purposes this possible error is worth the extra speed, for example, in the control of eye movements in response to a sudden change in an image and perhaps also for looming

detectors in the theory of directional selectivity (see Section 3.4). Furthermore, just because a boundary is physical does not always make it safe to use. The edges of a uniform cylindrical lamppost give rise to perfectly good edges in the images seen by the left and right eyes, but these edges correspond to different lines on the physical surface. This gives the stereopsis process trouble when, having matched the images, it tries to calculate how far away the lamppost is.

So our rule, then, that the inputs to a process should consist of elements with close physical correlates, is only a general one. It is clearly inappropriate for some things, like shape from shading or photometric stereo, but probably rather important for things like the correspondence process in apparent motion (Ullman, 1978) or the analysis of shape from surface contours or texture. The rule has its attendant dangers, though, and for some processes it is obeyed only marginally—for example, I think that both stereopsis and directional selectivity can use zero-crossings directly. However, the important point is that the rule is sufficiently strong and apparently valid and that violations cannot be allowed to go unnoticed. They have to be defended.

So much, then, for the level of computational theory. The second of the three levels of understanding a process is the level of the algorithm. At this level we formulate a particular procedure for implementing a computational theory. There are two principles that guide the design of algorithms, and they probably ought to be satisfied by any serious candidate for an early visual process in the human visual system. One principle says, roughly, that the algorithm has to be robust: the other, that it must behave smoothly. They are as follows (Marr, 1976):

1. *Principle of graceful degradation.* This principle is designed to ensure that, wherever possible, degrading the data will not prevent the delivery of at least some of the answer. It amounts to a condition on the continuity of the relation between different stages in the processing. For example, it should be required that a rough two-dimensional description of the kind that a vision system might compute out of a drawing enable the system to compute a rough three-dimensional description of what the drawing represents.

2. *Principle of least commitment.* This principle requires not doing something that may later have to be undone, and I believe that it applies to all situations in which performance is fluent. It states that algorithms that are constructed according to a hypothesize-and-test strategy should be avoided because there is probably a better method. My experience has been that if the principle of least commitment has to be disobeyed, one is

It would be nice to be able to give general rules about processes at the third level of analysis, the level of neural implementation. Unfortunately, only a few process theories have been developed to the point where specific neural implementations have been proposed, and none of these implementations have been confirmed experimentally in every detail so we are not yet in a position to formulate such rules.

However, one suggestion of a rule can be extracted from our experience with cooperative algorithms for stereopsis and locally parallel organization (Marr and Poggio, 1976; Stevens, 1978). It is only a suggestion, however, and I give it with that caution. It is that, if possible, the nervous system avoids iterative methods—that is, pure iteration in which no new information is introduced at each cycle. Instead, it seems to prefer one-shot methods, like Stevens' (1978) one-shot algorithm for finding the local orientation in Glass patterns. The nervous system also seems to prefer methods that run from the coarse to the fine, doing essentially the same thing at each state but being saved from pure iteration by introducing new information at each cycle. Our stereo algorithm has this form, as we shall see in the next section. And it might be a sound design principle, too, since it effortlessly incorporates the principles of graceful degradation and least commitment.

Yet cooperative methods (a type of nonlinear, iterative algorithm) look very plausible from some points of view. They are very robust, for example, and often have a structure that is readily translatable into the inhibitory and excitatory connections of a plausible neural network. Why, then, are they not used?

One possible explanation may be that cooperative methods take too long and demand too much of the neural hardware to be implemented in any direct way. The problem with iteration is that it demands the circulation of numbers around some kind of loop, which could be carried out by some system of recurrent collaterals or closed loops of neuronal connections. However, unless the numbers involved can be represented quite accurately as they are circulated, errors characteristically tend to build up rather quickly. To use a neuron to represent a quantity with an accuracy of even as low as 1 in 10, it is necessary to use a time interval that is sufficiently long to hold between 1 and 10 spikes in comfort. This means at least 50 ms per iteration for a medium-sized cell, which means 200 ms for four iterations—the minimum time ever required for our cooperative algorithm to solve a stereogram. And this is too slow.

This argument against purely iterative algorithms is not compelling. It is, however, persuasive enough to make me skeptical of them as candidates for processes used by the human visual processor, and it suggests that one should try very hard when designing ways of implementing a process to use algorithms with a more open and flexible structure

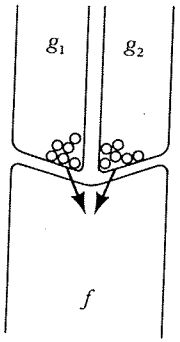


Figure 3–2. The synaptic arrangement considered by Torre and Poggio (1978). Such an arrangement could approximate an AND–NOT gate.

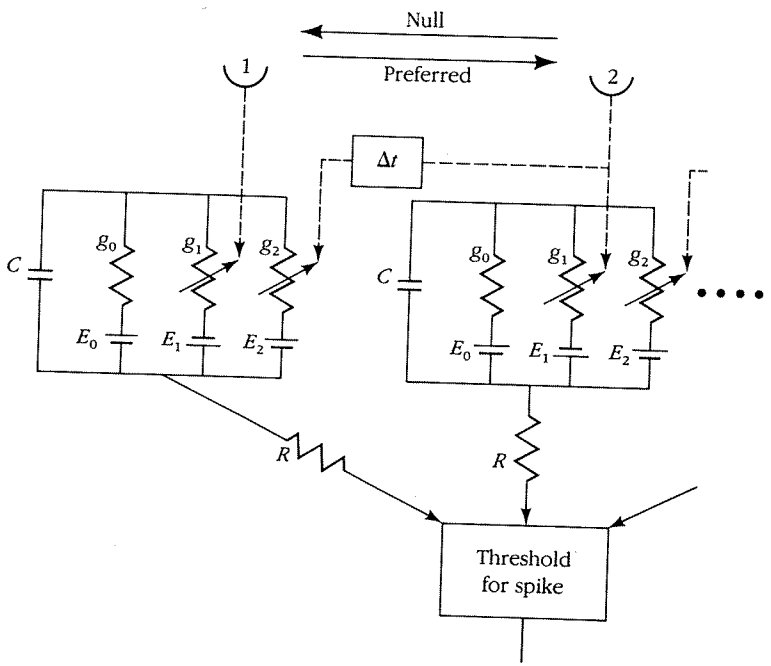


Figure 3–3. The electrical circuit equivalent of the synaptic arrangement shown in Figure 3–2 in the configuration suggested by Torre and Poggio (1978) for implementing directional selectivity. The interaction implemented by the circuit has the form $g_1 - \alpha g_1 g_2$, which approximates a logical AND–NOT gate. A logical AND gate can be implemented by a similar circuit.

One other lesson about neural implementations may perhaps be drawn, this time from the work of Torre and Poggio (1978), who showed how the nonlinear operation AND–NOT could be implemented at the level of synaptic interactions on a dendrite. They showed, using a cable-theoretical analysis, which calculates the time dependent electrical properties of the dendrite from its geometry, that the synaptic arrangement shown in Figure 3–2 has the electrical properties of the circuit shown in Figure 3–3 and the behavior shown in Figure 3–4. It approximately com-

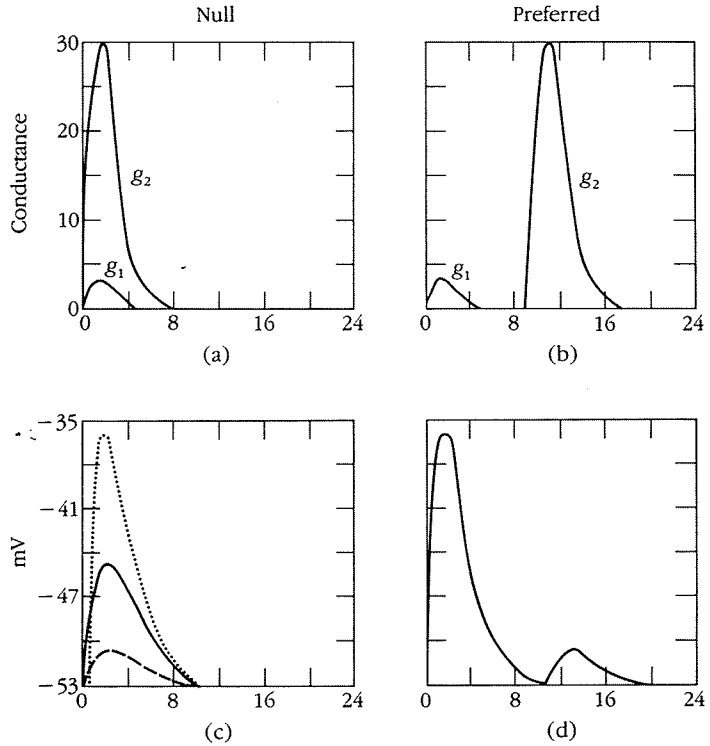


Figure 3–4. The calculated behavior of the circuit in Figure 3–3. For movement in the null direction, the time course of the inputs g_1 and g_2 is shown in (a), and the output of the circuit is the solid line in (c). The dotted and dashed curves show, respectively, the responses with g_1 and g_2 separately. For motion in the opposite direction, the inputs arrive as shown in (b), and the output of the circuit is shown in (d). Notice how attenuated (c) is relative to (d). In this manner, the output of the system can be made directionally selective. The time courses (horizontal axes) are plotted in units of the membrane time constant.

putes $g_1 - \alpha g_1 g_2$, which behaves like AND–NOT, and they suggested that this might be how the ideas of Hassenstein and Reichardt (1956) and of Barlow and Levick (1965) about directional selectivity in the fly and rabbit retinas are implemented (see Section 3.4). Poggio and Torre (1978) extended this idea, showing that a wide range of primitive, nonlinear operations could be implemented using local synaptic mechanisms.

One message of this work is that neurons might do more than we think. Early models, like those of McCulloch and Pitts (1943), tended to see neurons as basically linear devices that could implement nonlinear functions by means of a threshold, which could perhaps be variable if produced by an inhibitory interneuron. This way of thinking led Barlow and Levick to formulate their model of directional selectivity, and I employed it myself when I was interested in the cerebellar cortex (Marr, 1969). We have already seen, however, that local nonlinearities may be important. For example, the scheme for zero-crossing detection in Figure 2–18 is based on the use of many AND gates. The force of Poggio and Torre's work is that such things as AND gates may not require whole cells for their implementation—they can perhaps be executed much more compactly by local synaptic interactions in small pieces of dendrite.

Enough, then, of generalities; let us turn to the processes themselves. I shall start with stereopsis, since it was the first psychological process to be understood and because it led to much of the general knowledge about early vision already incorporated into my account. I have tried not to be too technical in describing the various processes, my aim being to give the reader a general feel for how they all work and to show some examples of them working. For full details, the reader may consult the original articles.

One final point about the organization of the account. Many of these processes divide naturally into two parts, the first concerned with setting up and making a measurement, so to speak, and the second with using the measurement to recover three-dimensional structure. In stereopsis, for example, the first step is the matching process, which establishes the correspondence between the two eyes so that disparities can be measured; the second is the trigonometry that recovers distance and surface orientation from disparity. The first step is the difficult one; the second is easy. In directional selectivity, the first step is to establish the local direction of movement, and the second is to use this sparse local information to help separate figure from ground. Neither step is particularly difficult. In apparent motion, the first step is to establish a correspondence between successive "frames" so that the displacements between frames can be measured; the second step is to use these measurements to recover three-dimensional structure. Here both steps are difficult.

For this reason I have split several of the sections into two parts. Of course, whether a process is indeed implemented by the human visual processor is sometimes unknown, and, even if it were known, whether it is divided as I have described is still an open psychophysical question. In such cases, I have tried to make clear what the current evidence is and what needs to be done to resolve the open questions.

3.3 STEREOPSIS

We saw earlier that the two eyes form slightly different images of the world. The relative difference in the positions of objects in the two images is called disparity, which is caused by the differences in their distance from the viewer. Our brains are capable of measuring this disparity and of using it to estimate the relative distances of the objects from the viewer. I shall use the term *disparity* to mean the angular discrepancy in position of the image of an object in the two eyes; the term *distance* will refer to the objective physical distance from the viewer to the object, usually measured from one of the two eyes; and the term *depth* I shall reserve for the subjective distance to the object as perceived by the viewer.

I shall divide the account into two parts, the first concerned with measuring disparity, and the second with using it. Both parts are separated into the three levels of Figure 1–4. The articles on which this account is based are by Marr (1974b) and Marr and Poggio (1976), which deal with the computational theory; by Marr and Poggio (1979), which deals with the algorithm thought to be used by the human visual system; and by Grimson and Marr (1979) and Grimson (1981), which describe Eric Grimson's computer implementation of the algorithm. Between 1977 and 1979, the additional work done on zero-crossings (Marr, Poggio, and Ullman, 1979; Marr and Hildreth, 1980) allowed certain simplifications in the implementation of the algorithm; most notably, we found from mathematical arguments that we could use circularly symmetric instead of oriented receptive fields for the initial convolutions. This particular detail was arrived at independently on psychophysical grounds by Mayhew and Frisby (1978a).

Measuring Stereo Disparity

Computational theory

Three steps are involved in measuring stereo disparity: (1) A particular location on a surface in the scene must be selected from one image; (2)