CHAPTER 2

Representing the Image

2.1 PHYSICAL BACKGROUND OF EARLY VISION

We cannot develop a rigorous theory of early vision—the first stages of the vision process—unless we know what the theory is for. We have already seen that, in general terms, the aim is to develop useful canonical descriptions of the shapes and surfaces that form the image. It is now time to state the goals more boldly (Marr 1976, 1978).

There are four main factors responsible for the intensity values in an image. They are (1) the geometry and (2) the reflectances of the visible surfaces, (3) the illumination of the scene, and (4) the viewpoint. In an image, all these factors are muddled up, some intensity changes being due to one cause, others to another, and some to a combination. The purpose of early visual processing is to sort out which changes are due to what factors and hence to create representations in which the four factors are separated.

Roughly speaking, it is proposed that this goal is reached in two stages. First, suitable representations are obtained of the changes and structures in the image. This involves things like the detection of intensity changes, the representation and analysis of local geometrical structure, and the detection of illumination effects like light sources, highlights, and transparency. The result of this first stage is a representation called the *primal sketch*. Second, a number of processes operate on the primal sketch to derive a representation—still retinocentric—of the geometry of the visible surfaces. This second representation, that of the visible surfaces, is called the *2½-dimensional (2½-D) sketch*. Both the primal sketch and the 2½-D sketch are constructed in a viewer-centered coordinate frame, and this is the aspect of their structures denoted by the term *sketch*.

The necessity for representing spatial relations, with its attendant complexities of how much should be made explicit and how much can safely be left implicit, raises problems that are typical of and rather special to vision. For example, the reader, especially if from a nonmathematical background, should not be put off by the notion of a coordinate frame, because it is probably a much more general notion than the reader thinks. To say that early visual representations are retinocentric does not literally imply that a Cartesian coordinate system, marked out in minutes of arc, is somehow laid out across the striate cortex, and that whenever some line or edge is noticed it is somehow associated with its particular x- and y-coordinates, whose values are somehow carried around by the neural machinery. This process would be one way of making the representations, to be sure, but no one would seriously propose it for human vision. There are many other ways in which this scheme can be realized in humans—for example, an (implicit) anatomical mapping that roughly preserves the spatial organization of the retina together with a representation that makes local relations explicit (point A is 5′ from point B in direction 35°) would seem plausible.

The important point about a retinocentric frame is that the spatial relations represented refer to two-dimensional relations on the viewer's retina, not three-dimensional relations relative to the viewer in the world around him, nor two-dimensional relations on another viewer's retina, nor three-dimensional relations relative to an external reference point like the top of a mountain. To say that image point A is below image point B is a remark in a retinocentric frame. To say one's hand is to the left of and below one's chest is a remark in one's own three-dimensional, viewer-centered frame. To say that the tip of a certain cat's tail is above and to the left of its body is a remark in a coordinate frame that is centered on the cat. They are all perfectly good ways of specifying rough spatial relationships, yet none uses sets of numbers. One can speak of each of these frames in terms of numbers—as if one was using (x, y, z), for example—but that

does not mean that they have to be implemented this way, and it is important to bear this in mind.

Although it helps a great deal to formulate the purpose of early vision in the rather straightforward terms of separating out the four factors of geometry, reflectance, illumination, and viewpoint, it is important to be aware of the simplifications that are involved in doing so. Perhaps the most important simplification is the rather rigid distinction between surface reflectance and surface geometry. In fact, these two notions are linked, and the distinction between them can be rather imprecise, so that one must be a little cautious when using them. A field of ripening wheat provides a convenient illustration of some of the difficulties. When seen from close by, the individual wheat stems form the reflecting surfaces, and the situation is relatively straightforward. When viewed from afar, however, image resolution is insufficient to distinguish the stems; the field as a whole forms the visible surface, and its reflectance function may now be very complex, since it incorporates considerable variation that should more properly be viewed as spatial (see, for example, Bouguer, 1957; Trowbridge and Reitz, 1975). Thinking of a distant wheat field or the coat of a cat as a surface is probably not too unrealistic an approximation for the theory of perception. We do see surfaces smoothed out. Tyler (1973), for example, found that we cannot see surface corrugations in stereograms if their spatial frequency is higher than about 4 cycles per degree.

In addition to these complexities, the illumination of a scene can only rarely be described in simple terms: Diffuse illumination, reflections, multiple light sources (only some of which are visible), and illumination between surfaces often conspire to create very complex illumination conditions, which will probably never be solved analytically. Nevertheless, our crude division into four categories has its uses. Provided that the variation in depth from the viewer of the surface from which light is reflected is small compared with the viewing distance, I shall assume that what is viewed can be regarded as a reflecting surface, and that the relation between its incident and reflected light may be described by a reflectance function ρ that, for a given illumination and viewpoint, may have a complex spatial structure.

Finally, a general point about the exposition. The purpose of these representations is to provide useful descriptions of aspects of the real world. The structure of the real world therefore plays an important role in determining both the nature of the representations that are used and the nature of the processes that derive and maintain them. An important part of the theoretical analysis is to make explicit the physical constraints and assumptions that have been used in the design of the representations and processes, and I shall be quite careful to do this.

Representing the Image

From an information-processing point of view, our primary purpose now is to define a representation of the image of reflectance changes on a surface that is suitable for detecting changes in the image's geometrical organization that are due to changes in the reflectance of the surface itself or to changes in the surface's orientation or distance from the viewer. If one thinks for a minute about a smooth surface, then changes in orientation and perhaps also in distance are likely to give rise to a change in image intensity. If the surface is textured, then quantities like the orientation or size of tiny elements on the surface—perhaps rough length and width—and measures taken over a small area reflecting the density and spacing of these elements yield the important clues in an image.

Hence we can see in a general way what our representation should contain. It should include some type of "tokens" that can be derived reliably and repeatedly from images and to which can be assigned values of attributes like orientation, brightness, size (length and width), and position (for density and spacing measurements). It is of critical importance that the tokens one obtains correspond to real physical changes on the viewed surface; the blobs, lines, edges, groups, and so forth that we shall use must not be artifacts of the imaging process, or else inferences made from their structure backwards to the structure of the surface will be meaningless. Let us therefore take a look at the general nature of surface reflectance functions, for this will give us important clues as to how we should structure our early representations.

Underlying Physical Assumptions

Existence of surfaces

Our first assumption is that it is proper to speak of surfaces at all, and it refers to the discussion that we had earlier about wheat fields and cats' coats. Stated precisely, it is *that the visible world can be regarded as being composed of smooth surfaces having reflectance functions whose spatial structure may be elaborate.*

Hierarchical organization

The second assumption has to do with the organization of this spatial structure, and it may help to introduce the topic with some examples. As

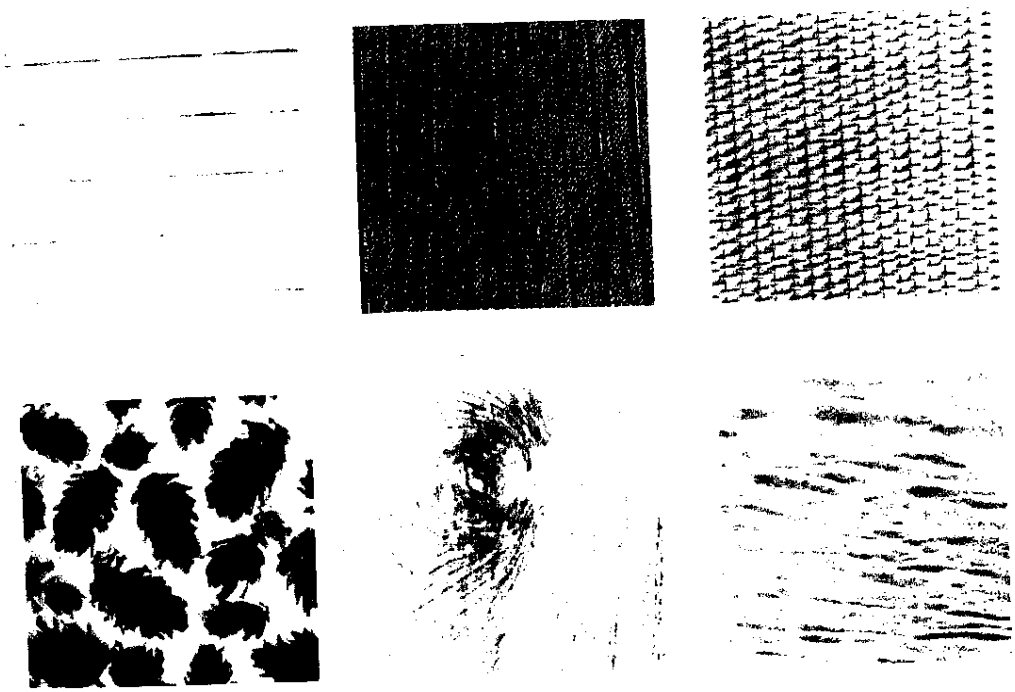


Figure 2–1. Some images of surfaces. Notice how different types of spatial organization occur almost independently at different scales. An important aspect of early vision is concerned with capturing these different organizations. (Reprinted by permission from Phil Brodatz, *Textures: A Photographic Album for Artists and Designers,* Dover, 1966, pl. D11.)

we have already seen, the coat of a cat is composed at the finest level of single hairs, each of which has its own reflectance function. At the next level up, these are organized into a surface by being placed close and parallel to one another. Then, over the coat so formed is the still higher-level organization of surface markings and coloration. The surface of a river has an analogous organization. At the basic level there is the flat water, randomly perturbed by protrusions like rocks or prominences. Superimposed on this surface are ripples oriented by gusts of wind and patches of weed and vegetation oriented by the flow of the river. There are analogous levels of structure in many surfaces—a hedgerow, a fabric, a rush weave, the bark of a tree, the grain of wood, a rock face, and so on (examine for a moment the surfaces illustrated in Figure 2–1).

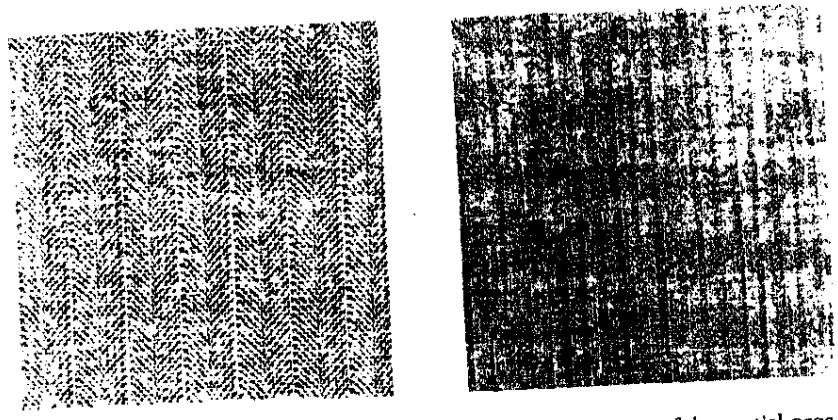


Figure 2–2. In a herringbone pattern such as this, a clear part of the spatial organization consists of the vertical stripes. These cannot be recovered by Fourier techniques such as band-pass filtering the images, but yield easily to grouping processes. (Reprinted by permission from Phil Brodatz, *Textures: A Photographic Album for Artists and Designers,* Dover, 1966, pl. 16, 17.)

From these examples, we see that the attributes carrying the valuable information may emerge at any of a range of scales in the real world, and hence even more so in images because of the additional transformations introduced by the imaging process. Whatever tokens are, we must therefore expect them to be capable of making image features explicit over a wide range of sizes. Furthermore, it is important to realize that these different levels of organization do not correspond simply to what would be seen through medium band-pass spatial-frequency filters* centered on different frequencies. Although several types of organization can be detected in this way, many cannot—for example, the vertical stripes in the pattern of Figure 2–2.

We can therefore formulate our second physical assumption: *The spatial organization of a surface's reflectance function is often generated by a number of different processes, each operating at a different scale.* Consequently, a representation that uses changes in the image of such surfaces to find changes in depth and surface orientation must be capable of capturing changes in attribute values applied to tokens that span a wide range of sizes in the image. In other words, the primitives of our representation must work at a number of different scales.

*Such filters eliminate all spatial frequency components in the image outside a fixed range of frequencies.

Similarity

Our third assumption is of a rather different kind. Suppose that we already had a representation containing primitives of various sizes. It seems intuitively obvious that they should be kept separate in some way—that a given large-scale descriptor should be compared with other large-scale descriptors much more readily than with small-scale ones. And perhaps it also seems obvious that tokens or descriptors having other extreme dissimilarities—very different or even opposite-signed contrasts, for example—should somehow be kept rather separate.

We can, in fact, find a physical basis for why this should be so, and it is apparent in our earlier examples. Recall that among the various levels of organization present in an animal's coat, on the surface of a river, on the bark of a tree, in woven fabric, and so forth, the processes that operated to generate the reflectance function are relatively independent at each scale, but the items for which each process is responsible are visually much more similar to one another than to other things on the same surface. For example, a given hair in a cat's coat is much more similar to neighboring hairs than to the stripes formed by the arrangement of thousands of hairs. Similarity here may be measured in several ways, but a straightforward measure based on local contrast, size (length and width), orientation, and color would suffice (compare Jardine and Sibson, 1971, for a general discussion of dissimilarity measures).

This observation gives us the means for selecting items from an image during the assignment of primitives in its representation. It is important, and may be formulated as our third physical assumption that *the items generated on a given surface by a reflectance-generating process acting at a given scale tend to be more similar to one another in their size, local contrast, color, and spatial organization than to other items on that surface.*

The importance of this type of similarity is illustrated by Figure 2–3. Following Glass (1969), these patterns are created by superimposing on a set of random dots the same set of dots but rotated or expanded a little (Figure 2–3a). The effect works for tokens made of squares (Figure 2–3b) or for pairs of tokens made in quite different ways (Figure 2–3c). If the tokens are too different (Figure 2–3d), however, no pattern is seen. Glass and Switkes (1976) showed that the effect fails if the dots have opposite contrast or opponent colors. Stevens (1978, fig 51a) showed that if three sets of dots are superimposed—the original, a rotated, and an expanded set—no organization is visible. If, say, the rotated set is made much brighter than the other two, then one sees the organization present in the dimmer pairs. This proves that the effect is based on a symbolic comparison of the

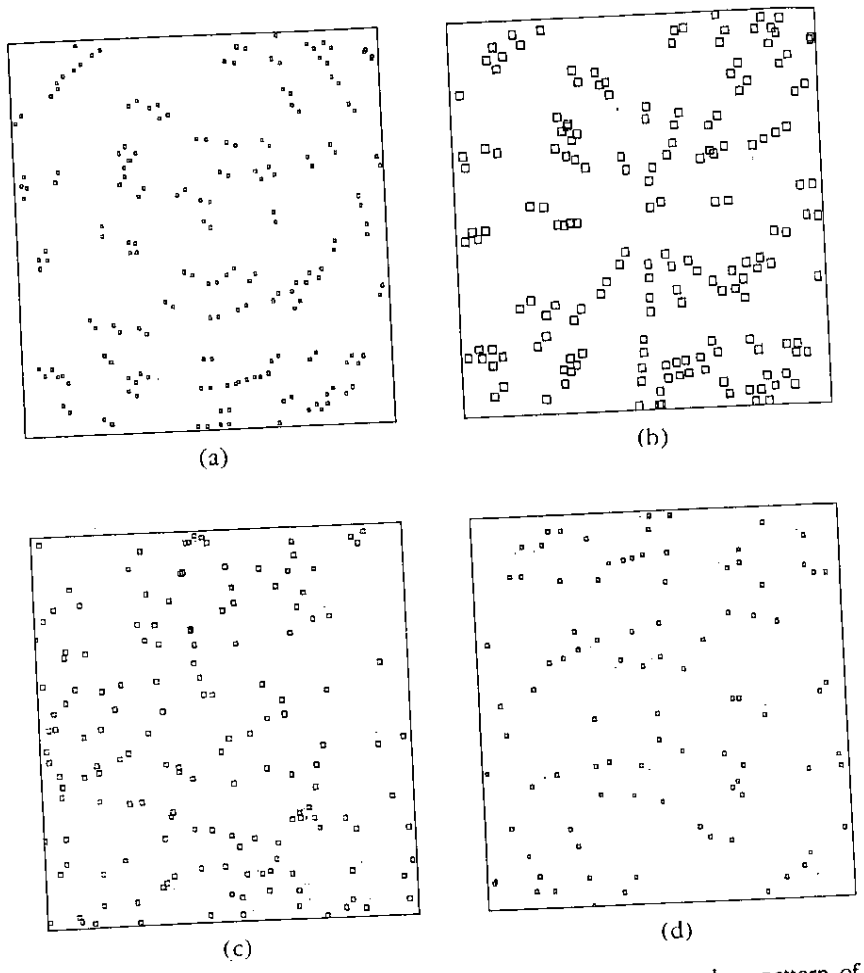


Figure 2–3. These displays are made by superimposing a random pattern of tokens on a slightly rotated or expanded copy of the same pattern. The tokens can be points or small squares (a) or larger squares (b). They do not have to be the same—in (c) one set consists of squares and the other set of four dots—but they do have to be similar. In (d), one set consists of quite large squares, and the other of small dots. These are apparently too dissimilar for us to discern the expanding structure there.

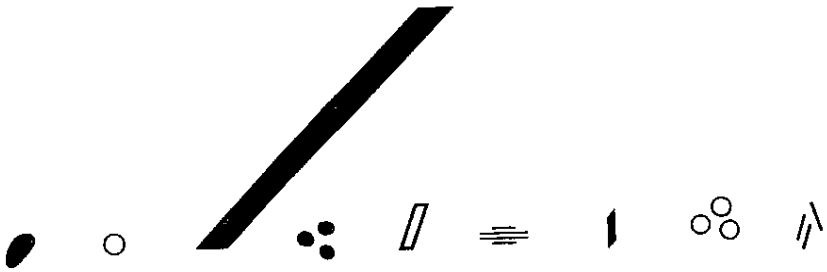


Figure 2–4. More evidence for place tokens. In this diagram every subgroup is defined differently, yet the collinearity of all of them is immediately apparent. This suggests that each group causes a place token to be created, whose collinearity is detected almost independently of the way the token is defined, provided that the tokens represent sufficiently similar items (compare Fig. 2–3d). (Reprinted by permission from D. Marr "Early processing of visual information," *Phil. Trans. R. Soc. Lond. B 275* 1976, fig. 10.)

properties of the local tokens and not, for example, on Hubel and Wiesel simple-cell-like measurements acting directly on the images.

Spatial continuity

In addition to their intrinsic similarity, *markings generated on a surface by a single process are often spatially organized—they are arranged in curves or lines and possibly create more complex patterns.* The basic feature is that markings often form smooth contours on a surface, and hence tokens will do so in an image. We are ourselves very sensitive to spatial continuity. We immediately see the items in Figure 2–4 (from Marr, 1976, fig. 10) as being collinear, despite the fact that every item along the line is defined in a different way: One is a blob, one is a small group of dots, one is the end of a bar, and so forth. They are, however, all about the same size. Figure 2–5 (from Marroquin, 1976, fig. 7) provides another fascinating example. There are very many continuous organizations buried in this pattern, and each one seems to be trying to jump out and dominate the others.

Continuity of discontinuities

One consequence of the cohesiveness of matter is that objects exist in the world and have boundaries. These give rise to the discontinuities in depth or surface orientation with whose detection we are concerned, and an important feature of such boundaries is that they often progress smoothly

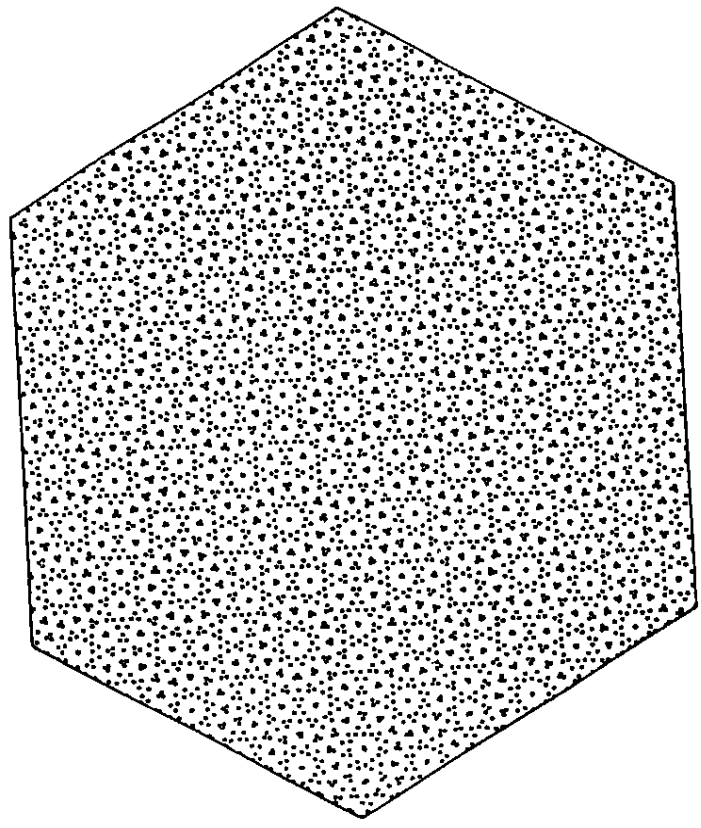


Figure 2–5. Evidence for the existence of active grouping processes. This pattern apparently seethes with activity as the rival organizations seem to compete with one another. (Reprinted by permission from J. L. Marroquin, "Human visual perception of structure," Master's thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1976.)

across an image. We can assume, in fact, that *the loci of discontinuities in depth or in surface orientation are smooth almost everywhere.* This is probably the physical constraint that makes the mechanism of smooth subjective contours a useful one (see Figure 2–6 and Section 4.8).

Continuity of flow

Finally, we must not forget that motion is extremely important for vision—it is ubiquitous. Motion of the viewer or of a physical object can cause

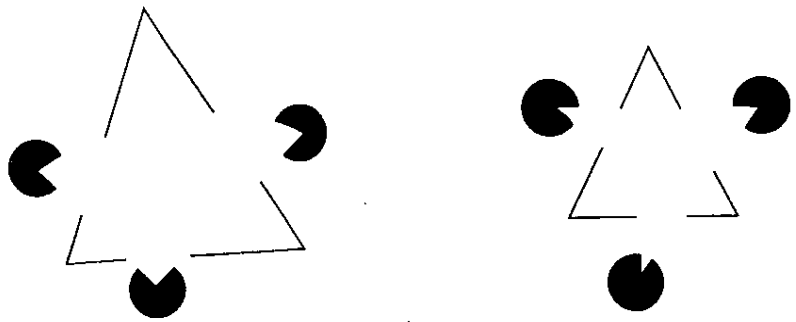


Figure 2–6. Subjective contours. The visual system apparently regards changes in depth as so important that they must be made explicit everywhere, including places where there is no direct visual evidence for them.

movements in the images of that object. If the object is rigid, the motions of the images of nearby portions of the object's surface are similar. Hence, the motions of portions of the object that are close to one another in the image are usually similar. In particular, the velocity field of motion in the image varies continuously almost everywhere, and if it is ever discontinuous at more than an isolated point, then a failure of rigidity (like an object boundary) is present in the outside world. In particular, *if direction of motion is ever discontinuous at more than one point—along a line, for example,—then an object boundary is present.*

General Nature of the Representation

The important message of these physical constraints is that although the basic elements in our image are the intensity changes, the physical world imposes on these raw intensity changes a wide variety of spatial organizations, roughly independently at different scales. This organization is reflected in the structure of images, and since it yields important clues about the structure of the visible surfaces, it needs to be captured by the early representations of the image. Specifically, I propose doing this by a set of "place tokens" that roughly correspond to oriented *edge* or *boundary* segments or to points of *discontinuity* in their orientations, to *bars* (roughly parallel edge pairs) or to their *terminations*; or to *blobs*—roughly, doubly terminated bars. These primitives can be defined in very concrete ways—from pure discontinuities in intensity—or in rather abstract ways. A blob

can be defined from a cloud of dots, for example, or a boundary from certain (but not all) kinds of texture change or from the lining up of a set of tokens that are themselves defined in quite complex ways, as in the example of Figure 2–4.

A rough illustration of the general idea appears in Figure 2–7; this representational scheme is called the *primal sketch* (Marr, 1976). The critical ideas behind it are the following:

1. The primal sketch consists of primitives of the same general kind at different scales—a blob has a rough position, length, width, and orientation at whatever scale it is defined—but the primitives can be defined from an image in a variety of ways, from the very concrete (a black ink mark) to the very abstract (a cloud of dots).

2. These primitives are built up in stages in a constructive way, first by analyzing and representing the intensity changes and forming tokens directly from them, then by adding representations of the local geometrical structure of their arrangement, and then by operating on these things with active selection and grouping processes to form larger-scale tokens that reflect larger-scale structures in the image, and so forth.

3. On the whole, the primitives that are obtained, the parameters associated with them, and the accuracy with which they are measured are designed to capture and to match the structure in an image so as to facilitate the recovery of information about the underlying geometry of the visible surfaces. This gives rise to a complex trade-off between the accuracy of the discriminations that can be made and the value of making them. For example, projected orientations in the image do change if the surface orientation changes, but on the whole by only a rather small amount and probably usually less than the typical variation in orientation to be found in the objective distribution of markings on a surface. This means that except in special situations, it is not worth having a very powerful apparatus for making subtle orientation discriminations. On the other hand, because only a very small relative movement is compelling evidence that two surfaces are separate, it is worth being very sensitive to relative movement.

The three main stages in the processes that derive the primal sketch are (1) the detection of zero-crossings (Marr and Poggio, 1979; Marr, Poggio, and Ullman, 1979; Marr and Hildreth, 1980); (2) the formation of the raw primal sketch (Marr, 1976; Marr and Hildreth, 1980; Hildreth 1980); and (3) the creation of the full primal sketch (Marr, 1976).

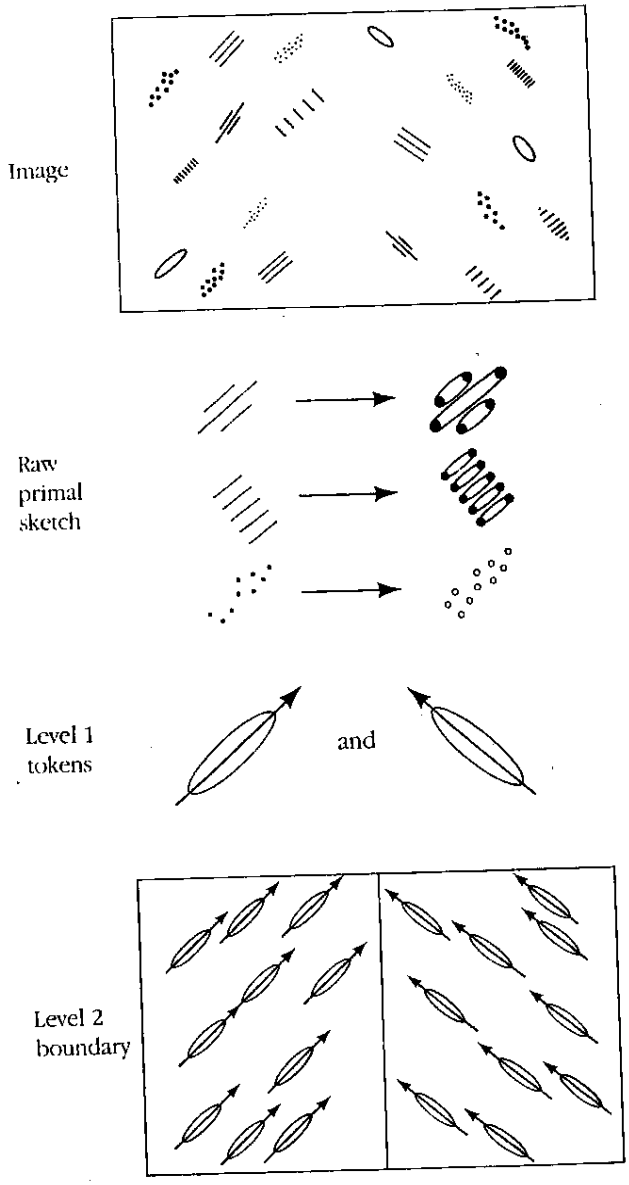


Figure 2–7. A diagrammatic representation of the descriptions of an image at different scales which together constitute the primal sketch. At the lowest level, the raw primal sketch faithfully follows the intensity changes and also represents terminations, denoted here by filled circles. At the next level, oriented tokens are formed for the groups in the image. At the next level, the difference in orientations of the groups in the two halves of the image causes a boundary to be constructed between them. The complexity of the primal sketch depends upon the degree to which the image is organized at the different scales.

2.2 ZERO-CROSSINGS AND THE RAW PRIMAL SKETCH

Zero-Crossings

The first of the three stages described above concerns the detection of intensity changes. The two ideas underlying their detection are (1) that intensity changes occur at different scales in an image, and so their optimal detection requires the use of operators of different sizes; and (2) that a sudden intensity change will give rise to a peak or trough in the first derivative or, equivalently, to a *zero-crossing* in the second derivative, as illustrated in Figure 2–8. (A zero-crossing is a place where the value of a function passes from positive to negative).

These ideas suggest that in order to detect intensity changes efficiently, one should search for a filter that has two salient characteristics. First and foremost, it should be a differential operator, taking either a first or second spatial derivative of the image. Second, it should be capable of being tuned to act at any desired scale, so that large filters can be used to detect blurry shadow edges, and small ones to detect sharply focused fine detail in the image.

Marr and Hildreth (1980) argued that the most satisfactory operator fulfilling these conditions is the filter $\nabla^2 G$, where ∇^2 is the Laplacian operator $(\partial^2/\partial x^2 + \partial^2/\partial y^2)$ and G stands for the two-dimensional Gaussian distribution

$$G(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}}$$

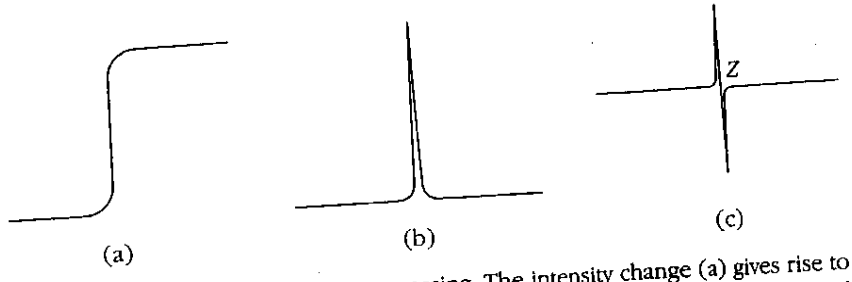


Figure 2–8. The notion of a zero-crossing. The intensity change (a) gives rise to a peak (b) in its first derivative and to a (steep) zero-crossing Z (c) in its second derivative.

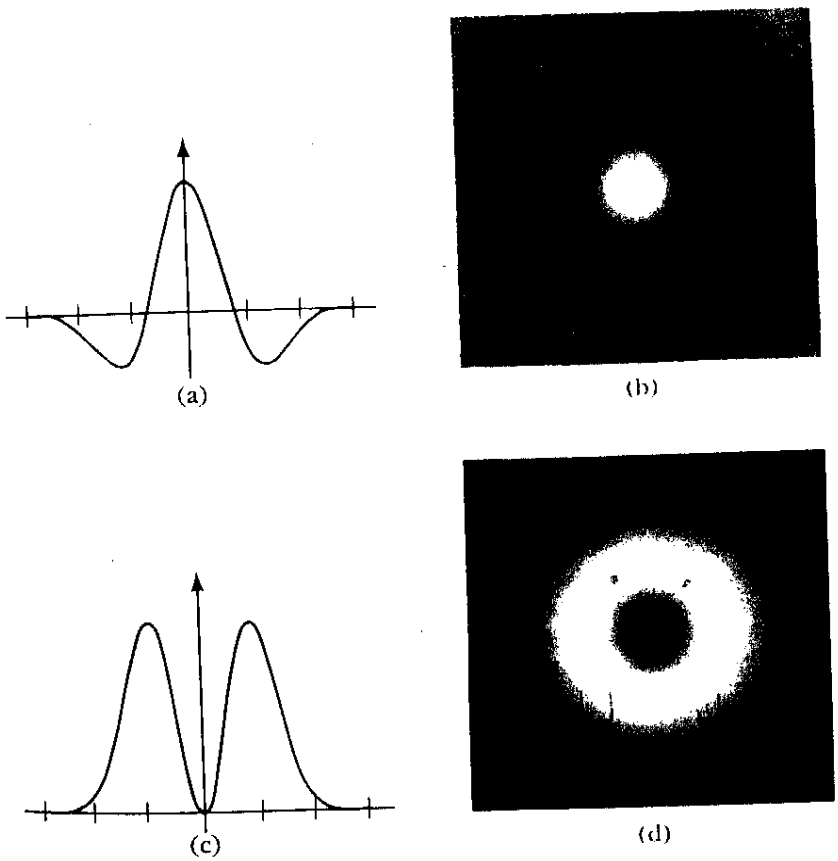


Figure 2–9. $\nabla^2 G$ is shown as a one-dimensional function (a) and in two-dimensions (b) using intensity to indicate the value of the function at each point. (c) and (d) show the Fourier transforms for the one- and two-dimensional cases respectively. (Reprinted by permission from D. Marr and E. Hildreth, "Theory of edge detection," *Proc. R Soc. Lond. B 204,* pp. 301–328.)

which has standard deviation σ. $\nabla^2 G$ is a circularly symmetric Mexican-hat-shaped operator whose distribution in two dimensions may be expressed in terms of the radial distance r from the origin by the formula

$$\nabla^2 G(r) = \frac{-1}{\pi\sigma^4}\left(1 - \frac{r^2}{2\sigma^2}\right) e^{\frac{-r^2}{2\sigma^2}}$$

Figure 2–9 illustrates the one- and two-dimensional forms of this operator, as well as their Fourier transforms.

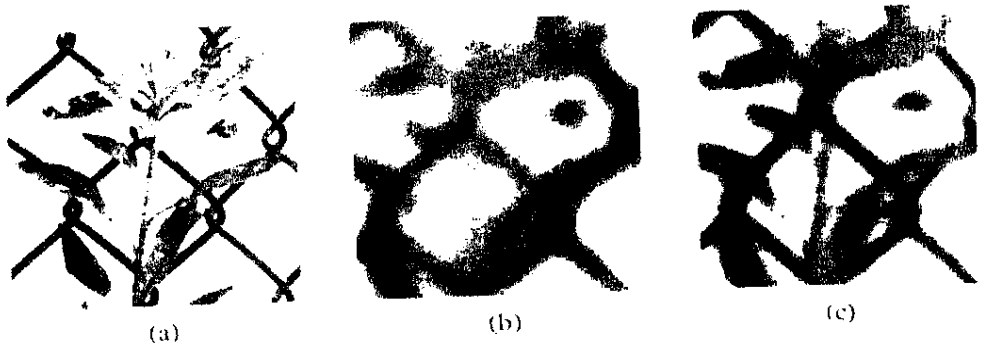


Figure 2–10. Blurring images is the first step in detecting intensity changes in them. (a) In the original image, intensity changes can take place over a wide range of scales, and no single operator will be very efficient at detecting all of them. The problem is much simplified in an image that has been blurred with a Gaussian filter, because there is, in effect, an upper limit to the rate at which changes can take place. The first part of the edge detection process can be thought of as decomposing the original image into a set of copies, each filtered with a different-sized Gaussian, and then detecting the intensity changes separately in each. (b) The image filtered with a Gaussian having $\sigma = 8$ pixels; in (c), $\sigma = 4$. The image is 320 by 320 elements. (Reprinted by permission from D. Marr and E. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond. B 204,* pp. 301–328.)

There are two basic ideas behind the choice of the filter $\nabla^2 G$. The first is that the Gaussian part of it, G, blurs the image, effectively wiping out all structure at scales much smaller than the space constant σ of the Gaussian. To illustrate this, Figure 2–10 shows an image that has been convolved with two different-sized Gaussians whose space constants σ were 8 pixels (Figure 2–10b) and 4 pixels (Figure 2–10c). The reason why one chooses the Gaussian for this purpose, rather than blurring with a cylindrical pillbox function (for instance), is that the Gaussian distribution has the desirable characteristic of being smooth and localized in both the spatial and frequency domains and, in a strict sense, being the unique distribution that is simultaneously optimally localized in both domains. And the reason, in turn, why this should be a desirable property of our blurring function is that if the blurring is as smooth as possible, both spatially and in the frequency domain, it is least likely to introduce any changes that were not present in the original image.

The second idea concerns the derivative part of the filter, ∇^2. The great advantage of using it is economy of computation. First-order directional derivatives, like $\partial/\partial x$ or $\partial/\partial y$, could be used, in which case one would subsequently have to search for their peaks or troughs at each orientation (as illustrated in Figure 2–8b); or, second-order directional derivatives, like $\partial^2/\partial x^2$ or $\partial^2/\partial y^2$, could be used, in which case intensity changes would

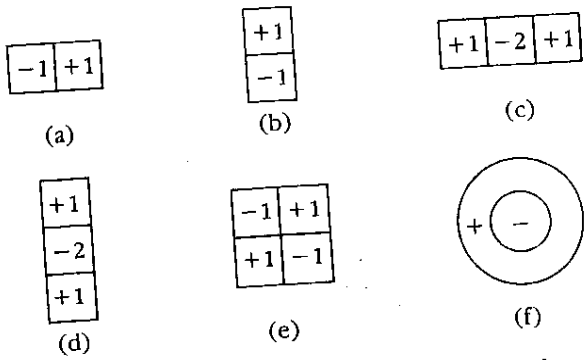
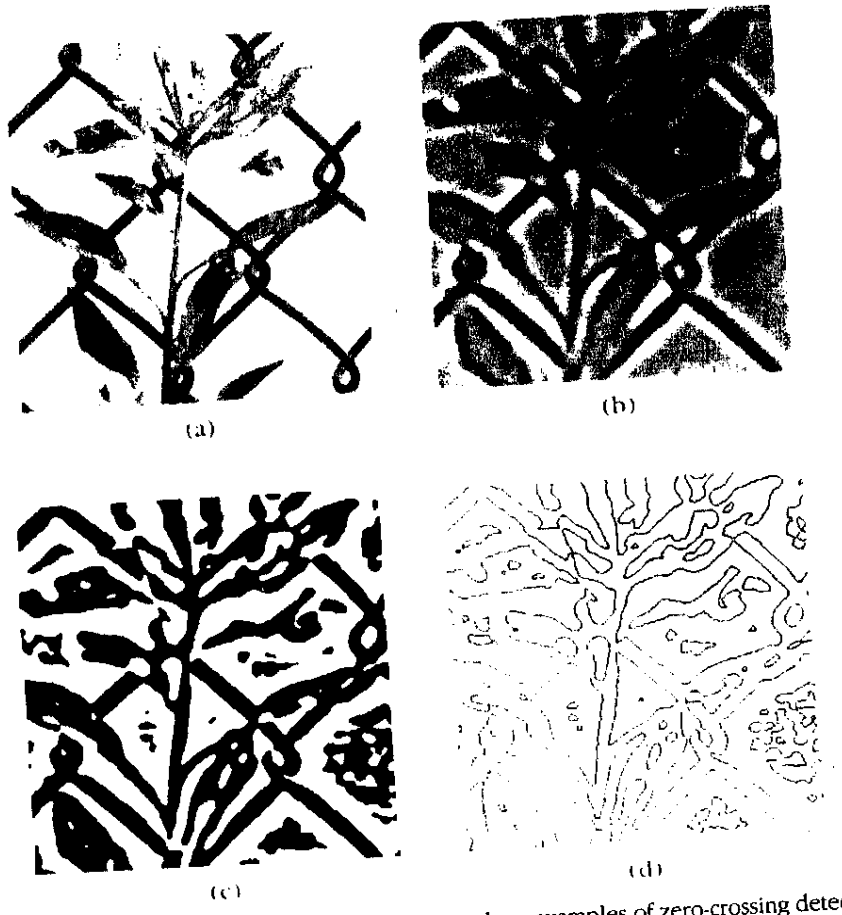


Figure 2–11. The spatial configuration of low-order differential operators. Operators like $\partial/\partial x$ can be roughly realized by filters with the receptive fields illustrated in the figure. (a) $\partial/\partial x$ can be thought of as measuring the difference between the values at two neighboring points along the x-axis. Similarly, (b) shows $\partial/\partial y$. The operator $\partial^2/\partial x^2$ can be thought of as the difference between two neighboring values of $\partial/\partial x$, and so it takes the form shown in (c). The other two second-order operators, $\partial^2/\partial y^2$ and $\partial^2/\partial x \partial y$, appear in (d) and (e), respectively. Finally, the lowest-order isotropic operator, the Laplacian $(\partial^2/\partial x^2 + \partial^2/\partial y^2)$, which we denote by ∇^2, has the circularly symmetric form shown in (f).

correspond to their zero-crossings (see Figure 2–8c). However, the disadvantage of all these operators is that they are directional; they all involve an orientation (see Figure 2–11, which illustrates the spatial organizations, or "receptive fields," in neurophysiological terms of the various first- and second-order differential operators). In order to use the first derivatives, for example, both $\partial I/\partial x$ and $\partial I/\partial y$ have to be measured, and the peaks and troughs in the overall amplitude have to be found. This means that the signed quantity $[(\partial I/\partial x)^2 + (\partial I/\partial y)^2]^{-\frac{1}{2}}$ must also be computed.

Using second-order directional derivative operators involves problems that are even worse than the ones involved in using first-order derivatives. The only way of avoiding these extra computational burdens is to try to choose an orientation-independent operator. The lowest-order isotropic differential operator is the Laplacian ∇^2, and fortunately it so happens that this operator can be used to detect intensity changes provided the blurred image satisfies some quite weak requirements (Marr and Hildreth, 1980).* Images on the whole do satisfy these requirements locally,

*The mathematical notation for blurring an image intensity function $I(x, y)$ with a Gaussian function G is $G * I$ which is read G convolved with I. The Laplacian of this is denoted by $\nabla^2 (G * I)$ and a mathematical identity allows us to move the ∇^2 operator inside the convolution giving $\nabla^2 (G * I) = (\nabla^2 G) * I$.



Figures 2–12, 13, 14. These three figures show examples of zero-crossing detection using $\nabla^2 G$. In each figure, (a) shows the image (320 × 320 pixels); (b) shows the image's convolution with $\nabla^2 G$, with $w_{2-D} = 8$ (zero is represented by gray); (c) shows the positive values in white and the negative in black; (d) shows only the zero-crossings.

so in practice one can use the Laplacian. Hence, in practice, the most satisfactory way of finding the intensity changes at a given scale in an image is first to filter it with the operator $\nabla^2 G$, where the space constant of G is chosen to reflect the scale at which the changes are to be detected, and then to locate the zero-crossings in the filtered image.

Figures 2–12 to 2–14 show what an image looks like when processed in this way. The numerical values in the $\nabla^2 G$-filtered image are both positive

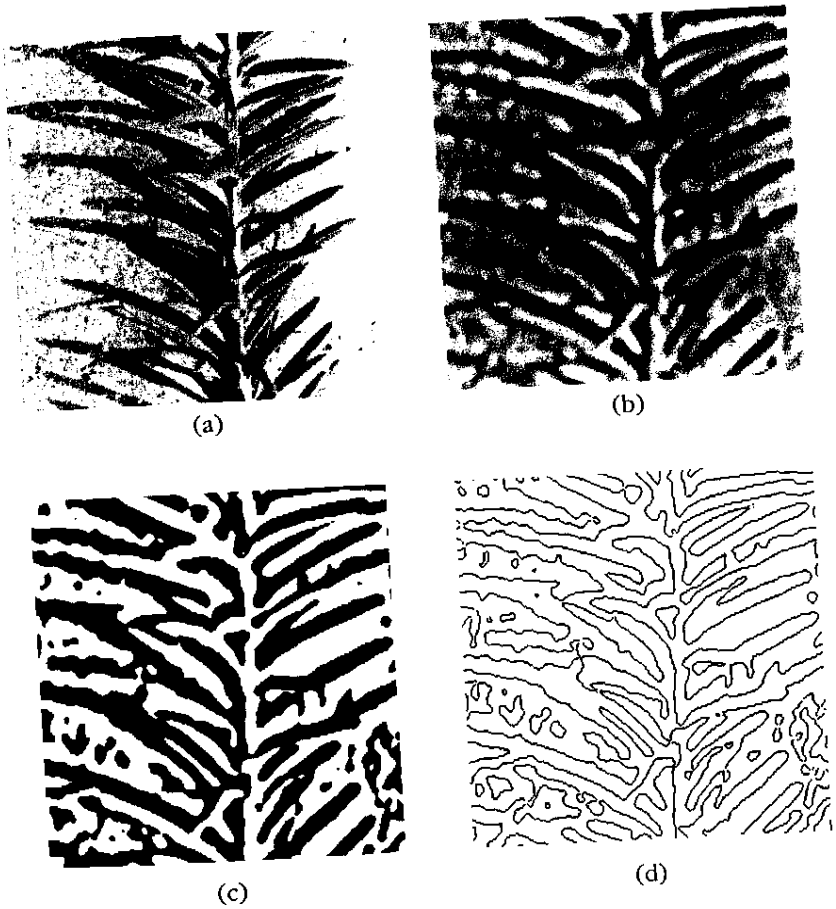


Figure 2–13.

and negative, the overall average being zero. Positive values are represented here by whites, negative by blacks, and the value zero by an intermediate gray. As we have seen, the critical fact about the operator $\nabla^2 G$ is that its zero-crossings mark the intensity changes, as seen at the Gaussian's particular scale. The figures show this well. In Figure 2–12(c), for instance, the filtered image has been "binarized"—that is, positive values were all set to $+1$ and negative values to -1, and in Figure 2–12(d) the zero-crossings alone are shown. The advantage of the binarized representation is that it also shows the sign of the zero-crossing—which side in the image is the darker.

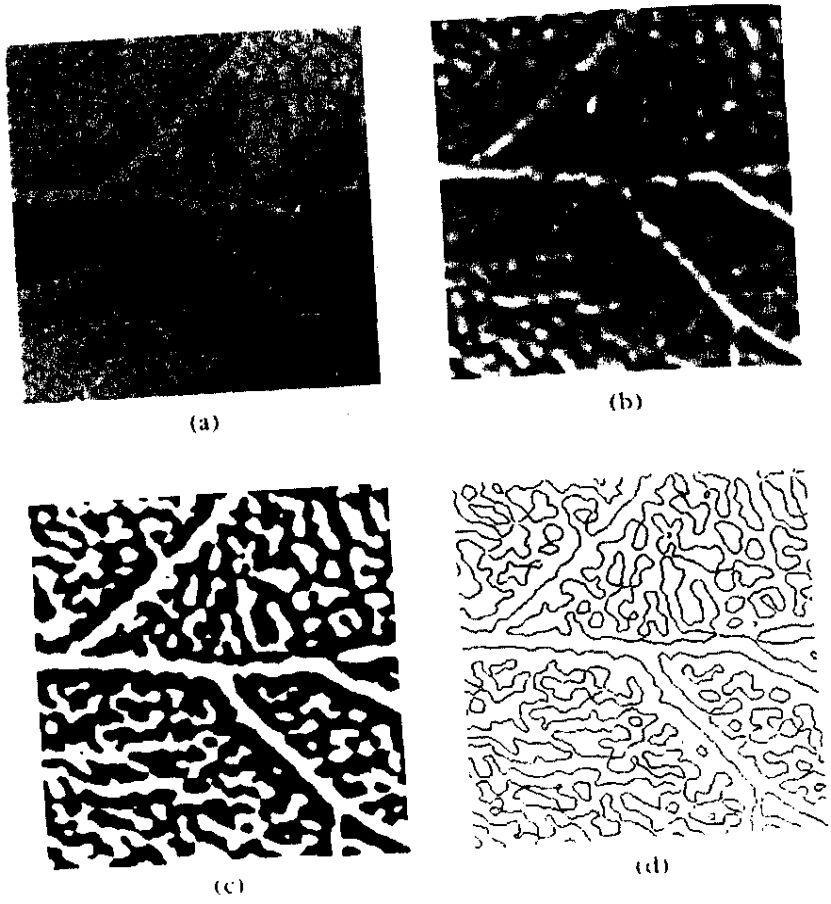


Figure 2–14.

In addition, the slope of the zero-crossing depends on the contrast of the intensity change, though not in a very straightforward way. This is illustrated by Figure 2–15, which shows an original image together with zero-crossings that have been marked with curves of varying intensity. The more contrasty the curve, the greater the slope of the zero-crossing at that point, measured perpendicularly to its local orientation.

Zero-crossings like those of Figures 2–12 to 2–15 can be represented symbolically in various ways. I choose to represent them by a set of oriented primitives called *zero-crossing segments*, each describing a piece of the contour whose intensity slope (rate at which the convolution changes across the segment) and local orientation are roughly uniform. Because of their eventual physical significance, it is also important to make explicit

(a) (b)

Figure 2–15. Another example of zero-crossings; here, the intensity of the lines has been made to vary with the slope of the zero-crossing, so that it is easier to see which lines correspond to the greater contrast. (Courtesy BBC Horizon.)

those places at which the orientation of a zero-crossing changes "discontinuously." The quotation marks are necessary because one can in fact prove that the zero-crossings of $\nabla^2 G * I$ can never change orientation discontinuously, but one can nevertheless construct a practical definition of discontinuity. In addition, small, closed contours are represented as blobs, each also with an associated orientation, average intensity slope, and size defined by its extent along a major and minor axis. Finally, in keeping with the overall plan, several sizes of operator will be needed to cover the range of scales over which intensity changes occur.

Biological Implications

This computational scheme for the very first stages in visual processing leads to an interpretation of many results from the psychophysical and neurophysiological investigations into early vision and to a proposal for the overall strategy behind the design of the first part of the visual pathway.

The psychophysics of early vision

In 1968, Campbell and Robson carried out some adaptation experiments. They found that the sensitivity of subjects to high-contrast gratings was

temporarily reduced after exposure to such gratings and this desensitization was specific to the orientation and spatial frequency of the gratings. The experimenters concluded that the visual pathway included a set of "channels" that are orientation and spatial frequency selective.

This finding provided an explosion of articles investigating various aspects of the detailed structure of these channels, culminating recently in an elegant quantitative model for their structure in humans, constructed on the basis of data from threshold detection studies by Wilson and Giese (1977) and Wilson and Bergen (1979). The model is quite easy to understand. The basic idea is that at each point in the visual field, there are four size-tuned filters or masks analyzing the image. The spatial receptive fields of these filters all have approximately the shape of a DOG, that is, of the difference of two Gaussian distributions, but the smaller two filters exhibit relatively sustained temporal properties, whereas the larger two are relatively transient. The channels are labeled N, S, T, and U, in order of increasing size, and their dimensions scale linearly with increasing eccentricity (angular distance from the fovea). The S channel is the most sensitive under both sustained and transient stimulation; the U channel is the least, having only one-fourth to one-eleventh the sensitivity of the S channel. Wilson himself made no statement about whether the filters were oriented, but he measured their dimensions using light and dark lines. With these one-dimensional stimuli, the widths of the central part of the receptive fields, which I shall denote by the symbol w_{1-D}, had the following values: N channel, 3.1′; S channel, 6.2′; T channel, 11.7′; and U channel, 21′. The receptive field sizes increase linearly with eccentricity, being about doubled at 4° eccentricity. Essentially all of the psychophysical data on the detection of spatial patterns below 16 cycles per degree at contrast threshold can be explained by this model, together with the hypothesis that the detection process is based on a form of spatial probability summation in the channels.

It is the $\nabla^2 G$ filters, I think, that form the basis for these psychophysically determined channels. The $\nabla^2 G$ operator approximates a band-pass filter with a bandwidth at half power of 1.25 octaves. It can be approximated closely by a DOG, the best approximation from an engineering point of view being achieved when the two Gaussians that form the DOG have space constants in the ratio 1:1.6. Figure 2–16 shows how good this approximation is. Wilson's estimate of the ratio for his sustained channels was 1:1.75.

In order to relate the numerical values of w_{1-D} measured by Wilson and Bergen to the values of the diameter w_{2-D} of the central part of the receptive fields of the underlying $\nabla^2 G$ operators, one must remember to multiply their values by $\sqrt{2}$, since all the measurements Wilson made correspond to a linear projection of the circularly symmetric receptive

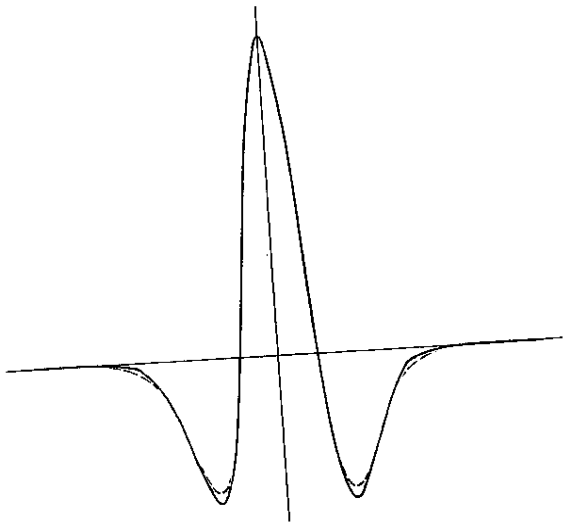


Figure 2–16. The best engineering approximation to $\nabla^2 G$ (shown by the continuous line), obtained by using the difference of two Gaussians (DOG), occurs when the ratio of the inhibitory to excitatory space constraints is about 1:1.6. The DOG is shown here dotted. The two profiles are very similar. (Reprinted by permission from D. Marr and E. Hildreth, "Theory of edge detection, " *Proc. R. Soc. Lond. B 204,* pp. 301–328.)

fields. Hence Wilson's N channel would correspond to a $\nabla^2 G$ filter with $w_{2-D} = 3.1\sqrt{2} = 4.38'$, which corresponds to the diameter of about nine foveal cones. This seems rather large for the smallest channel, and arguments based on a theoretical analysis of acuity and resolution suggest that a smaller one exists. The diameter w_{2-D} of the central part of its receptive field should be about 1′ 20″, and because of diffraction in the eye, it could correspond to the midget ganglion cells, whose receptive field centers are driven by a single cone (see Marr, Poggio, and Hildreth, 1980).

Thus if Wilson's figures are correct, they tell us the sizes that the initial center–surround operators should have in order to produce the observed psychophysical adaptation and other effects. These numbers can then in principle be related to the measurements made by physiologists, in the manner that we shall derive in the next section. The final point to note here is that Campbell also found the adaptation to be orientation specific (and it may also be specific for the direction of movement). This we attribute to the stage at which zero-crossings are detected, which is best explained by looking at the neurophysiology.

The physiological realization of the $\nabla^2 G$ filters

It has been known since Kuffler (1953) that the spatial organization of the receptive fields of the retinal ganglion cells is circularly symmetric, with a central excitatory region and an inhibitory surround. Some cells, called on-center cells, are excited by a small spot of light shone on the center of their receptive fields, and others are inhibited. Rodieck and Stone (1965) suggested that this organization was the result of superimposing a small central excitatory region on a larger inhibitory "dome" that extends over the entire receptive field. Enroth-Cugell and Robson (1966) described the two domes as Gaussians, thus describing the receptive field as a difference of two Gaussians (a DOG). In addition, Enroth-Cugell and Robson divided the larger retinal ganglion cells into two classes, X and Y, on the basis of their temporal response properties. X cells show a fairly sustained response, whereas the Y cells show a relatively transient one—a distinction that is preserved at the lateral geniculate nucleus. Wilson's sustained channels probably correspond to the physiological X cells, and the transient, to the Y cells (Tolhurst, 1975).

Thus it is not too unreasonable to propose that the $\nabla^2 G$ function is what is carried by the X cells of the retina and lateral geniculate body, positive values being carried by the on-center X cells, and negative values by the off-center X cells. To illustrate the physiological point, Figure 2–17 compares the predicted X-cell responses, using $\nabla^2 G$, against actual published records of retinal and lateral geniculate cells, which we identified as X cells, for three stimuli—an edge, a thin bar, and a wide bar. As we can see, the qualitative agreement is very good. I shall discuss the function of the Y cells in Section 3.4.

The physiological detection of zero-crossings

From a physiological point of view, zero-crossing segments are easy to detect without relying on the detection of zero values, which would be a physiologically implausible idea. The reason is that just to one side of the zero-crossing will lie a peak positive value of the filtered image $\nabla^2 G * I$, and just to the other side, a peak negative value. These peaks will be roughly $w_{2\text{-D}}/\sqrt{2}$ apart, where $w_{2\text{-D}}$ is the width of the receptive field center of the underlying filter $\nabla^2 G$. Hence, just to one side, an on-center X cell will be firing strongly, and just to the other side, an off-center X cell will be firing strongly; the sum of their firings will correspond to the slope of the zero-crossing—a high-contrast intensity change producing stronger firing than a low-contrast change. The existence of a zero-crossing can therefore

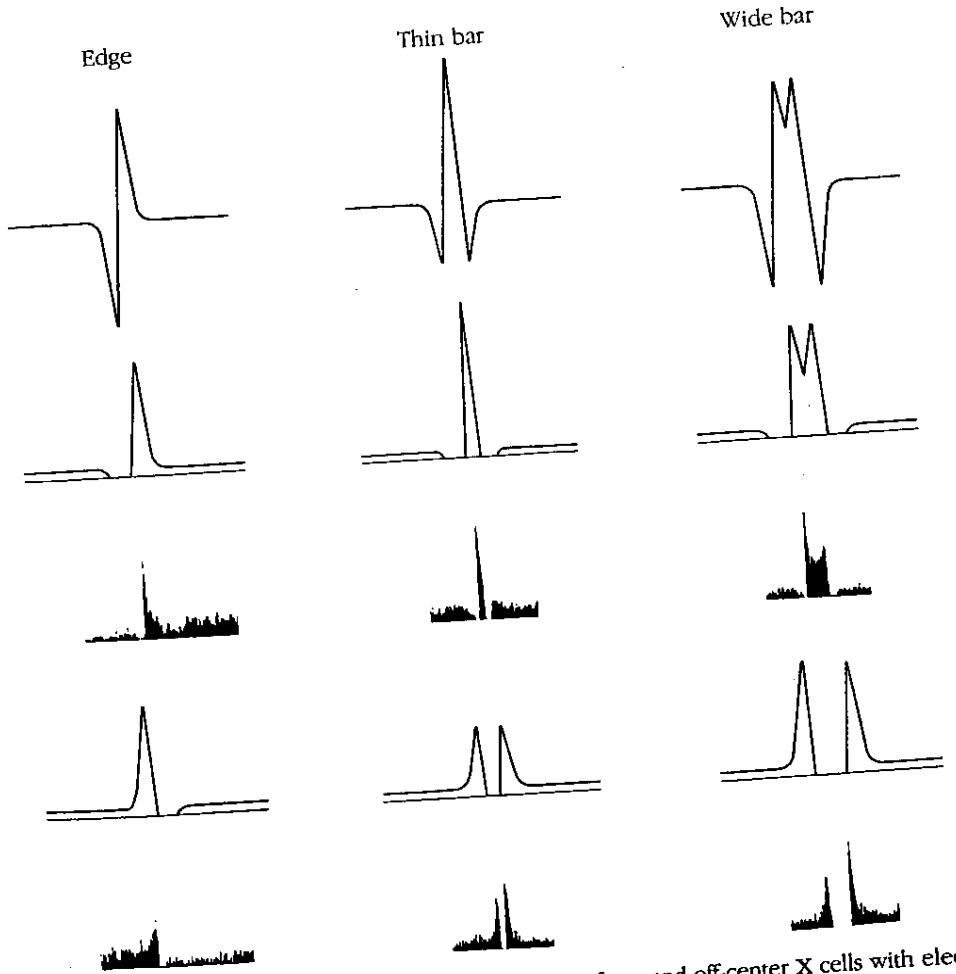


Figure 2–17. Comparison of the predicted responses of on- and off-center X cells with electrophysiological recordings. The first row shows the response to $\nabla^2 G * I$ for an isolated edge, a thin bar (bar width $= 0.5w_{1-D}$, where $_{1-D}$ is the width of the central excitatory region of the receptive field projected onto a line), and a wide bar (bar width $= 2.5w_{1-D}$). The predicted traces are calculated by superimposing the positive (in the second row) or the negative (in the fourth row) parts of $\nabla^2 G * I$ on a small resting or background discharge. The corresponding physiological responses (third and fifth rows) are taken from Dreher and Sanderson (1973, figs. 6d and 6e) for the responses to an edge and from Rodieck and Stone (1965, figs. 1 and 2), using traces from bars 1° and 5° wide. (Reprinted by permission from D. Marr and S. Ullman, "Directional selectivity and its use in early visual processing," *Phil. Trans. R. Soc. B 275*, pp. 483–524.)

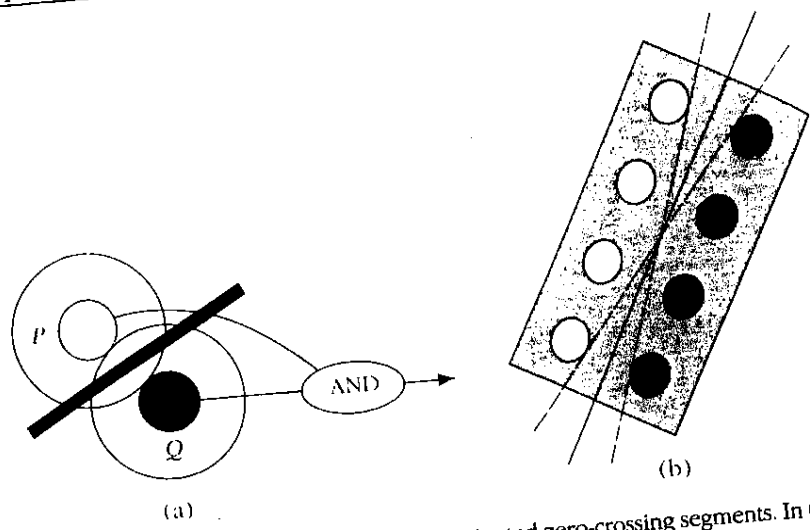


Figure 2–18. A mechanism for detecting oriented zero-crossing segments. In (a), if P represents an on-center geniculate X-cell receptive field, and Q an off-center, then a zero-crossing must pass between them if both are active. Hence, if they are connected to a logical AND gate as shown, the gate will detect the presence of the zero-crossing. If several are arranged in tandem as in (b) and are also connected by logical AND's, the resulting mechanism will detect an oriented zero-crossing segment within the orientation bounds given roughly by the dotted lines. Ideally, we would use gates that responded by signaling their sum only when all their P and Q inputs were active. (Reprinted, by permission, by D. Marr and E. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond. B 204*, pp. 301–328.)

be detected by a mechanism that connects an on-center cell and an off-center cell to an AND gate,* as illustrated in Figure 2–18(a).

It is a simple matter to adapt this idea to create an oriented zero-crossing segment detector: simply arrange on- and off-center X cells into two columns, as illustrated in Figure 2–18(b). If these units are all connected by AND gates or some suitable approximation to them, the result will be a unit that detects a zero-crossing segment whose orientation lies roughly between the dotted lines of Figure 2–18(b). This idea provides the basis for the model of cortical simple cells, which we shall derive in Section 3.4. It is enough to note here that such units would be orientation dependent and spatial-frequency-tuned (as well as directionally selective, after the modifications of Section 3.4). These are the units, I believe, that Campbell and Robson found that they could adapt in their 1968 experiments.

*A simple logical device that produces a positive output only when all of its inputs are positive.

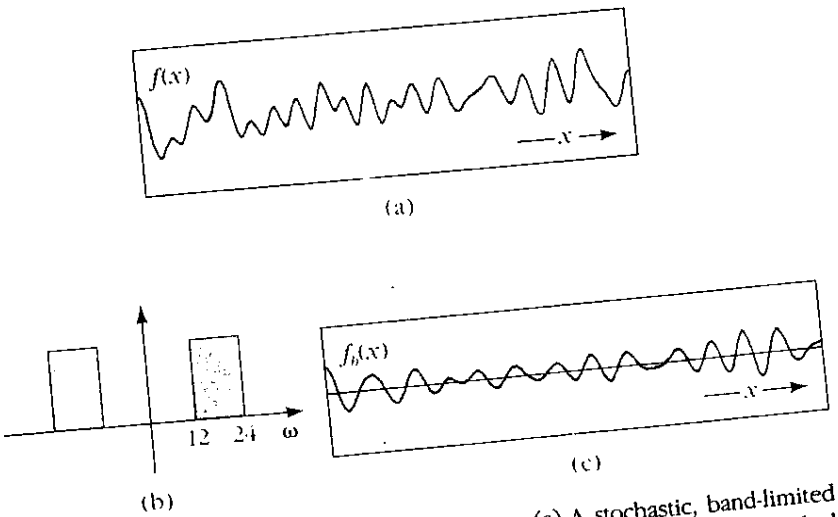


Figure 2–19. The meaning of Logan's theorem. (a) A stochastic, band-limited Gaussian signal $f(x)$. (b) The passband—in the frequency domain—of an ideal one-octave band-pass filter. (c) The result $f_b(x)$ of filtering (a) with the filter described by (b). Provided that (c) has no zeros in common with its Hilbert transform, Logan's theorem tells us that (c) is determined, up to a multiplicative constant, by the positions of its zero-crossings alone. The aspect of Logan's result that is important for early visual processing is that, under the right conditions, the zero-crossings alone are very rich in information. (Reprinted by permission from D. Marr, T. Poggio, and S. Ullman, "Bandpass channels, zero-crossings, and early visual information processing," *J. Opt. Soc. Am. 69*, 1979, fig. 1.)

The first complete symbolic representation of the image

Zero-crossings provide a natural way of moving from an analogue or continuous representation like the two-dimensional image intensity values $I(x,y)$ to a discrete, symbolic representation. A fascinating thing about this transformation is that it probably incurs no loss of information. The arguments supporting this are not yet secure (Marr, Poggio, and Ullman, 1979) and rest on a recent theorem of B. F. Logan (1977). This theorem states that provided certain technical conditions are satisfied, a one-octave band-pass signal can be completely reconstructed (up to an overall multiplicative constant) from its zero-crossings. Figure 2–19 illustrates the idea; the proof of the theorem is difficult, but consists essentially of showing that if the signal is less than an octave in bandwidth, then it must cross the x-axis at least as often as the standard sampling theorem requires.

Unfortunately, Logan's theorem is not quite strong enough for us to be able to make any direct claims about vision from it. The problems are

twofold. First, the zero-crossings in the visual application lie in two dimensions and not one, and it is often difficult to extend sampling arguments from one dimension to two. Second, the operator $\nabla^2 G$ is not a pure one-octave band-pass filter; its bandwidth at half power is 1.25 octaves, and at half sensitivity, 1.8 octaves. On the other hand, we do have extra information, namely, the values of the slopes of the curves as they cross zero, since this corresponds roughly to the contrast of the underlying edge in the image. An analytical approach to the problem seems to be very difficult, but in an empirical investigation, Nishihara (1981) found encouraging evidence supporting the view that a two-dimensional filtered image can be reconstructed from its zero-crossings and their slopes.

Figure 2–20 summarizes pictorially the point we have now reached. It shows the image, of a sculpture by Henry Moore, as seen through three different-sized channels; that is, it shows the zero-crossings of the image after filtering it through $\nabla^2 G$ filters where the Gaussians, G, have three different space constants. The next question is, What should we do with this information?

The Raw Primal Sketch

Up to now I have studiously avoided using the word *edge*, preferring instead to discuss the detection of intensity changes and their representation by using oriented zero-crossing segments. The reason for this is that the term *edge* has a partly physical meaning—it makes us think of a real physical boundary, for example—and all we have discussed so far are the zero values of a set of roughly band-pass second-derivative filters. We have no right to call these edges, or, if we do have a right, then we must say so and why. This kind of distinction is vital to the theory of vision and probably to the theories of other perceptual systems, because the true heart of visual perception is the inference from the structure of an image about the structure of the real world outside. The theory of vision is exactly the theory of how to do this, and its central concern is with the physical constraints and assumptions that make this inference possible.

We meet this for the first time now, as we address the problem posed by Figure 2–20—namely, How do we combine information from the different channels? The $\nabla^2 G$ filters that are actually used by the visual system are an octave or more apart, so there is no priori reason why the zero-crossings obtained from the different-sized filters should be related. There is, however, a physical reason why they often should be. It is a consequence of the first of our physical assumptions of the last chapter, and it is called the *constraint of spatial localization* (Marr and Hildreth, 1980). The things

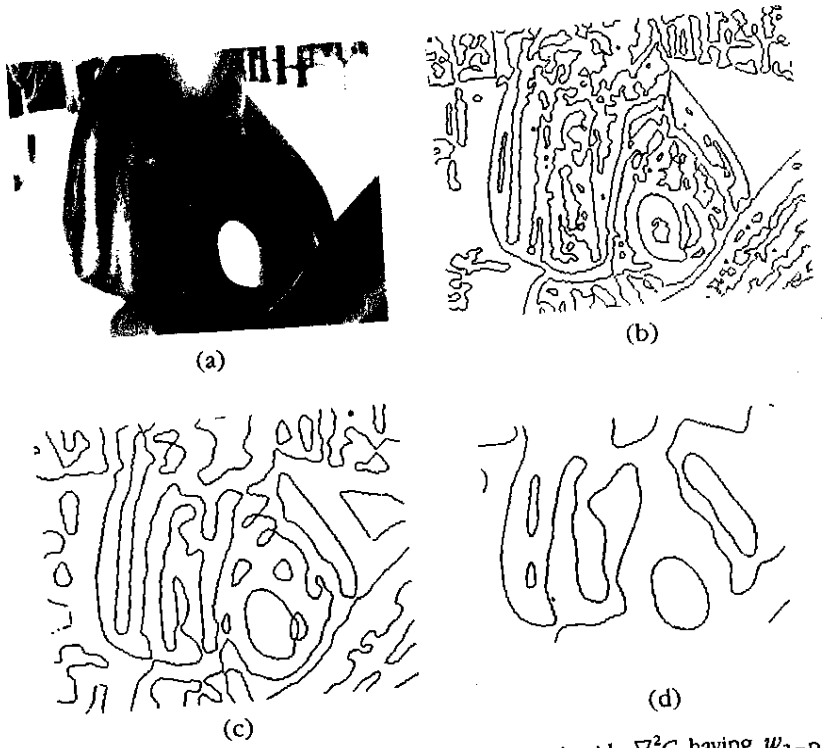


Figure 2–20. The image (a) has been convolved with $\nabla^2 G$ having $w_{2-D} = 2\sqrt{2}\,\sigma = 6, 12$, and 24 pixels. These filters span approximately the range of filters that operate in the human fovea. (b), (c), and (d) show the zero-crossings thus obtained. Notice the fine detail picked up by the smallest. This set of figures neatly poses the next problem—How do we combine all this information into a single description? (Reprinted by permission from D. Marr and E. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond. B 204,* pp. 301–328.)

in the world that give rise to intensity changes in an image are (1) illumination changes, which include shadows, visible light sources, and illumination gradients; (2) changes in the orientation or distance from the viewer of the visible surfaces; and (3) changes in surface reflectance.

The critical observation here is that, at their own scale, these things can all be thought of as spatially localized. Apart from the occasional diffraction pattern, the visual world is not constructed of ripply, wavelike primitives that extend over an area and that add together over it (compare Marr, 1970, p. 169). By and large, the visual world is made of contours, creases, scratches, marks, shadows, and shading, and these are spatially localized. Hence, it follows that if a discernable zero-crossing is present in

an image filtered through $\nabla^2 G$ at one size, then it should be present at the same location for all larger sizes. If this ceases to be so at some larger size, it will be for one of two reasons: Either two or more local intensity changes are interfering—being averaged together—in the larger channel, or two independent physical phenomena are operating to produce intensity changes in the same region of the image but at different scales. An example of the first situation is a thin bar, whose edges would be accurately located by small channels but not by large ones. Situations of this kind can be recognized by the presence of two nearby zero-crossings in the small channels. An example of the second situation is a shadow superimposed on a sharp reflectance change, which can be recognized if the zero-crossings in the large channels are displaced relative to those in the smaller ones. If the shadow has exactly the correct position and orientation, the locations of the zero-crossings may not contain enough information to separate the two physical phenomena, but in practice this situation will be rare.

Thus, the physical world constrains the geometry of the zero-crossings from the different-sized channels. We can exploit this by using it to formulate the *spatial coincidence assumption:*

If a zero-crossing segment is present in a set of independent $\nabla^2 G$ channels over a contiguous range of sizes, and the segment has the same position and orientation in each channel, then the set of such zero-crossing segments indicates the presence of an intensity change in the image that is due to a single physical phenomenon (a change in reflectance, illumination, depth, or surface orientation).

In other words, provided that the zero-crossings from independent channels of adjacent sizes coincide, they can be taken together. If the zero-crossings do not coincide, they probably arise from distinct surfaces or physical phenomena. It follows (1) that the minimum number of $\nabla^2 G$ channels required to establish physical reality is two and (2) that if there is a range of channel sizes, reasonably well separated in the frequency domain and covering an adequate range of the frequency spectrum, rules can be derived for combining their zero-crossings into a description whose primitives are physically meaningful (Marr and Hildreth, 1980).

The actual details of the rules are quite complicated because a number of special cases have to be taken into account, but the general idea is straightforward. Provided the zero-crossings in the larger channels are "accounted for" by what the smaller channels are seeing, either because they are in one-to-one correspondence with the zero-crossings in the

smaller channels or because they are blurred, averaged copies of them, then all the evidence points to a physical reality that is roughly what the smaller channels are seeing, perhaps modified and smoothed a little by the noise-reducing, averaging effects of the larger ones. In order to determine whether this accountability holds, configurations in which the zero-crossings of the smaller channels lie close to one another have to be detected explicitly, because it is these situations that can "fool" the larger channels. Hence the need for the explicit detection of spatial configurations such as thin bars and blobs.

If the larger channels' zero-crossings cannot be accounted for by what the smaller channels are seeing, then new descriptive elements must be developed, because the larger channels are recording different physical phenomena. This can happen in many ways: There may be a soft shadow, for example, or a wire grid in focus with an out-of-focus landscape behind; or a water beetle scurrying along the ripply surface of a pond with the weeds at the bottom forming a defocused background.

The description of the image to which these ideas lead is called the *raw primal sketch* (Marr and Hildreth 1980; Hildreth, 1980). Its primitives are edges, bars, blobs, and terminations, and these have attributes of orientation, contrast, length, width, and position. An example appears in Figure 2–21. It can be thought of as a binary map (Figure 2–21a) specifying the precise positions of the edge segments, together with the specifications at each point along them of the local orientation and of the type and extent of the intensity change (Figure 2–21d). Blob (Figure 2–21c), bar (Figure 2–21e), and discontinuity (termination) primitives can be made explicit in the same way. The representation of a long straight line, for example, consists of a termination, several segments having the same orientation, followed by a termination at the other end, as shown in Figure 2–22(a). The width, contrast, and orientation are in principle specified all along the way, although in practice it would be enough to provide this information at an adequate sampling interval. If the line is thicker than about the value of w for the smallest available channel, independent edge descriptions for its two sides would also be available. If the line curves, the orientation would gradually change along it (Figure 2–22b). If a discontinuity in orientation exists at some point along the line, then its location will be identified with a termination or discontinuity assertion (Figure 2–22c).

The raw primal sketch is a very rich description of an image, since it contains virtually all the information in the zero-crossings from several channels (two in the example of Figure 2–21). Its importance is that it is the first representation derived from an image whose primitives have a high probability of reflecting physical reality directly.

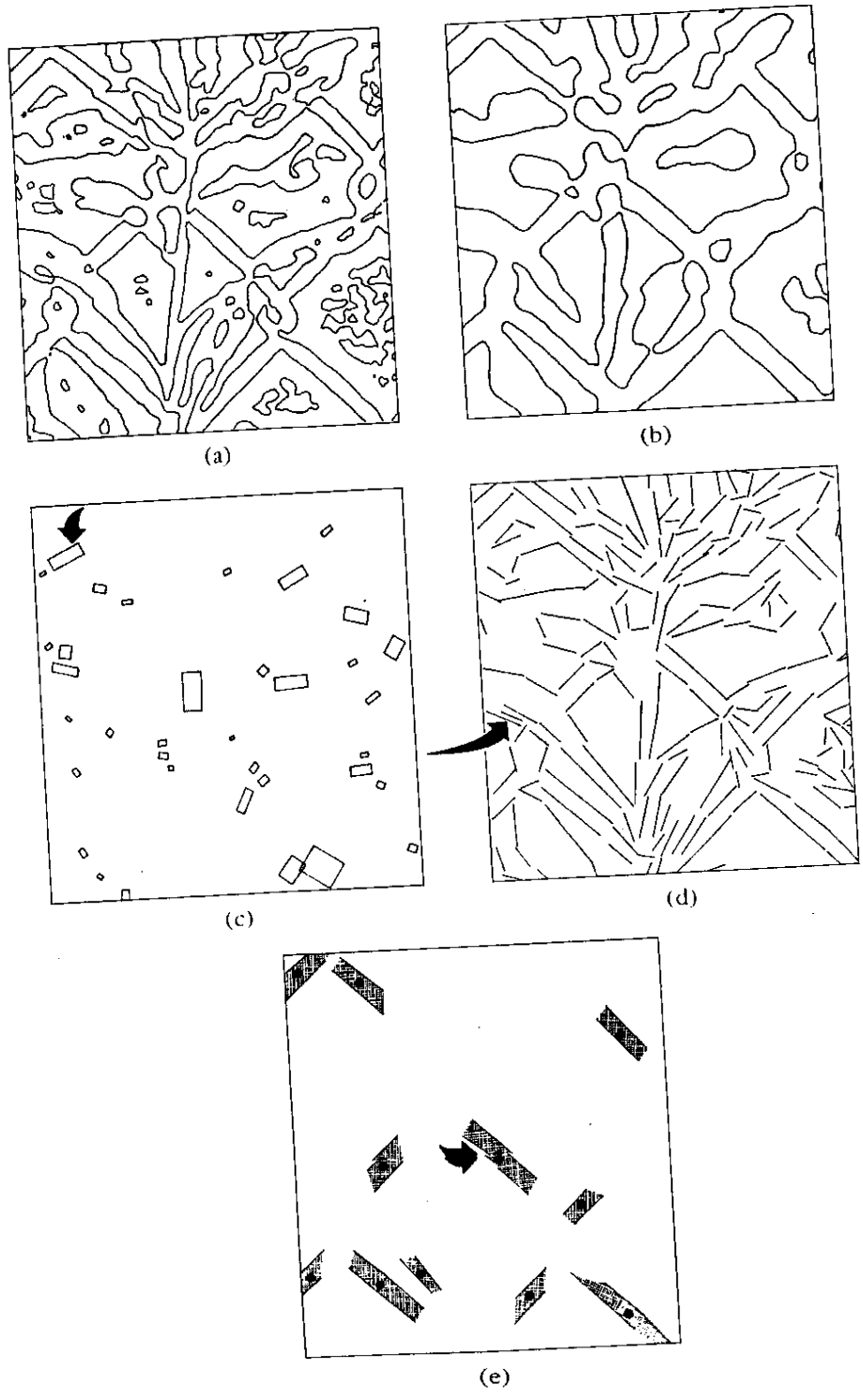

(a)
(b)
(c)
(d)
(e)

Figure 2–21. (opposite) The raw primal sketch as computed from two channels. (a), (b) The zero-crossings obtained from the image of Figure 2–12 by using masks with $w_{2-D} = 9$ and 18 pixels. Because there are no zero-crossings in the larger channel that do not correspond to zero-crossings in the smaller channel, the locations of the edges in the combined description also correspond to (a). (c), (d), and (e) Symbolic representations of the descriptors attached to the zero-crossing locations shown in (a). (c) Blobs. (d) Local orientations assigned to the edge segments. (e) Bars. The diagrams show only the spatial information contained in the descriptors. Typical examples of the full descriptors are:

BLOB	EDGE	BAR
(POSITION 146 21)	(POSITION 184 23)	(POSITION 118 134)
(ORIENTATION 105)	(ORIENTATION 128)	(ORIENTATION 128)
(CONTRAST 76)	(CONTRAST −25)	(CONTRAST −25)
(LENGTH 16)	(LENGTH 25)	(LENGTH 25)
(WIDTH 6)	(WIDTH 4)	(WIDTH 4)

The descriptors to which these correspond are marked with arrows. The resolution of this analysis of the image of Figure 2–12 roughly corresponds to what a human would see when viewing it from a distance of about 6 ft. Reprinted, by permission, from D. Marr and E. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond. B 204,* pp. 301–328.)

Subjectively, you are aware of the raw primal sketch—and of the full primal sketch of Section 2.5—but you are not aware of the zero-crossings from which it is made. In order to see what the larger channels are telling your brain, you have to screw up your eyes or defocus the image somehow. Only by doing so, for example, can you see Abraham Lincoln in L. D. Harman's discretely sampled and quantized picture of him (Figure 2–23), and only by doing so can you see lines running diagonally down a chessboard (Figure 2–24). Although the larger channels are "seeing" these things all the time, as shown in Figure 2–23, what they see is adequately accounted for by the zero-crossings that occur in the smaller channels. If the middle spectral frequencies are removed from the picture of Lincoln, this is no longer the case. The processes that combine zero-crossings now find no relation between what the smaller channels see and what the larger ones see, so they both give rise to primitives in the raw primal sketch. The result, as Harmon and Julesz (1973) found, is that one sees Abraham Lincoln behind a visible graticule. The primal sketch machinery assumes that the two different kinds of information are due to different physical phenomena, so we see both.

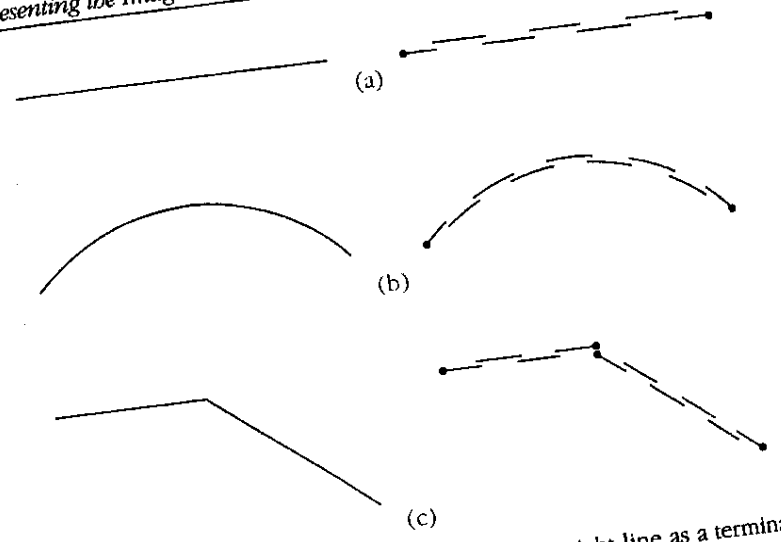


Figure 2–22. The raw primal sketch represents a straight line as a termination, several oriented segments, and a second termination (a). If the line is replaced by a smooth curve, the orientations of the inner segments will gradually change (b). If the line changes its orientation suddenly in the middle (c), its representation will include an explicit pointer to this discontinuity. Thus in this representation, smoothness and continuity are assumed to hold unless explicitly negated by an assertion.

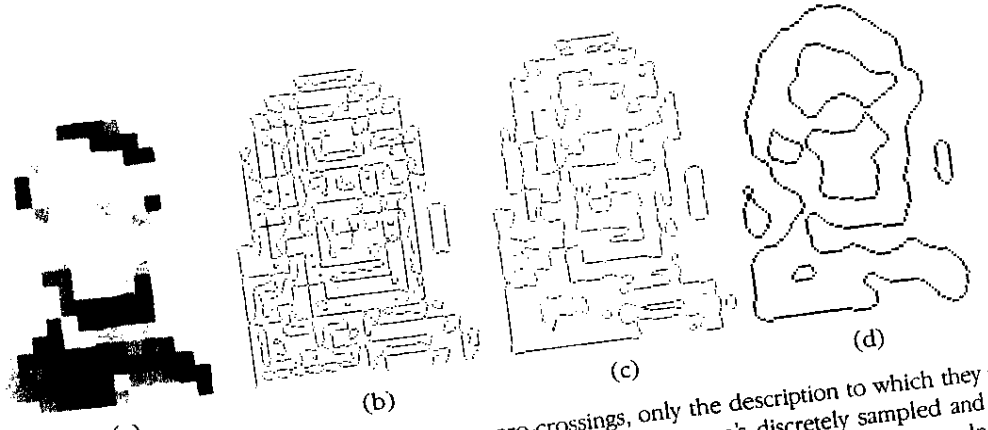


Figure 2–23. We cannot sense the primitive zero-crossings, only the description to which they give rise in the raw primal sketch. This can be seen in L. D. Harmon's discretely sampled and quantized image of Abraham Lincoln (a). No amount of voluntary effort allows us to see Lincoln without defocusing the image or squinting the eyes, despite the fact that the zero-crossings in the larger channels are producing an approximate representation of Lincoln's face. (b), (c), (d) The zero-crossings from the three sizes of the $\nabla^2 G$ operator used in Figure 2–20.

Philosophical Aside

It is interesting that the visual system takes this spatial, physical approach so seriously. It apparently does not allow the perception of a raw zero-crossing just on its own. Additional evidence, like a coincident zero-crossing from another channel seems to be required. Zero-crossings are also thought to form the input for the stereo matching process (see Chapter 3). Here again the input from two channels is combined, but this time from different eyes. Similar arguments hold for analyses based on directional selectivity, which is probably detected at the level of zero-crossings (see Section 3.4). However, once more, additional information is probably required before it is used, in this case an analysis of the coherence of the local motions in the visual field. The conclusion is that zero-crossings alone are insufficient, and this has a deep message for the whole approach, namely, that the visual system tries to deal only with physical things, using rules based on constraints supplied by the physical structure of the world to build up other descriptions that again have physical meanings.

This means that extreme care is required in the formulation of theories because nature seems to have been very careful and exact in evolving our visual systems. In this respect it is a great help to have the framework of the three levels explicitly available. Having to formulate the computational theory of a process introduces a great and useful discipline into the subject. No longer are we allowed to invoke a mechanism that seems to have some features in common with the problem and to assert that the mechanism works *like* the process. Instead, we have to analyze exactly what will work and be prepared to prove it. Stereo matching, for example, is like a lot of other things, but it is not the same as any of them. It is like a correlation, but it is not a correlation, and if it is treated like a correlation, the methods chosen will be unreliable. The job of stereo fusion is to match items that have definite physical correlates, because the laws of physics can guarantee only that items will be matchable if they correspond to things in space that have a well-defined physical location. Gray-level pixel values do not. Hence, gray-level correlation fails.

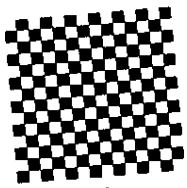
Again, the enterprise of looking for structure at different scales in an image, as illustrated by Figure 2–7 and developed in the next section, is reminiscent of ideas like filtering the image with different band-pass filters. Campbell (1977), for example, explicitly suggested that the fine details of a tank, like its registration number, might be explored using a high-pass filter, whereas the overall outline, which indicates that it is a tank, may be derived from a low-pass-filtered image. The point is once again that, just as for gray-level correlation and stereopsis, these ideas based on Fourier theory are *like* what is wanted, but they are not *what* is wanted; the structure

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.

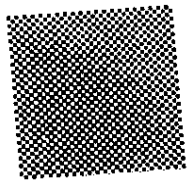
Another example is provided by the herringbone pattern of Figure 2–2. The vertical organization of the stripes is a clear example of an important spatial organization, yet it cannot be detected by Fourier methods because there is no power in the vertical orientation. However, such organization is easily 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

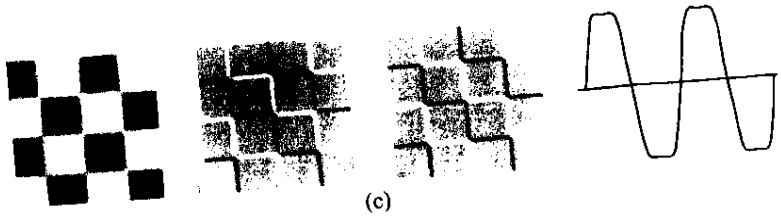
Figure 2–24. *(opposite)* The Fourier spectrum of a chessboard pattern (of infinite size) has all its power in the diagonal directions, and none in the horizontal or vertical. Yet in (a) we can see that the vertical, horizontal, and diagonal spatial organizations are all equally visible while in (b) 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).



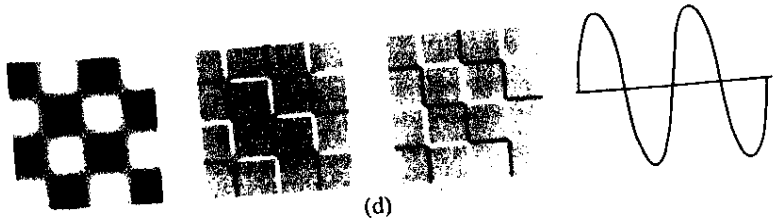
(a)



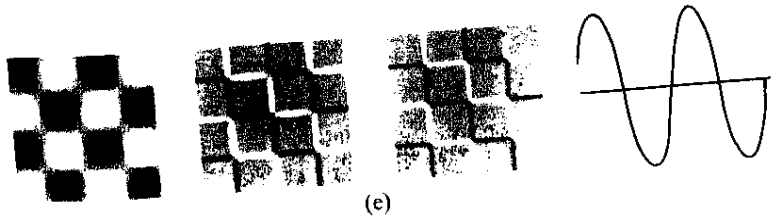
(b)



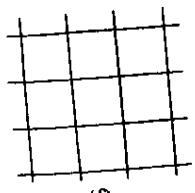
(c)



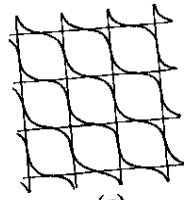
(d)



(e)



(f)



(g)

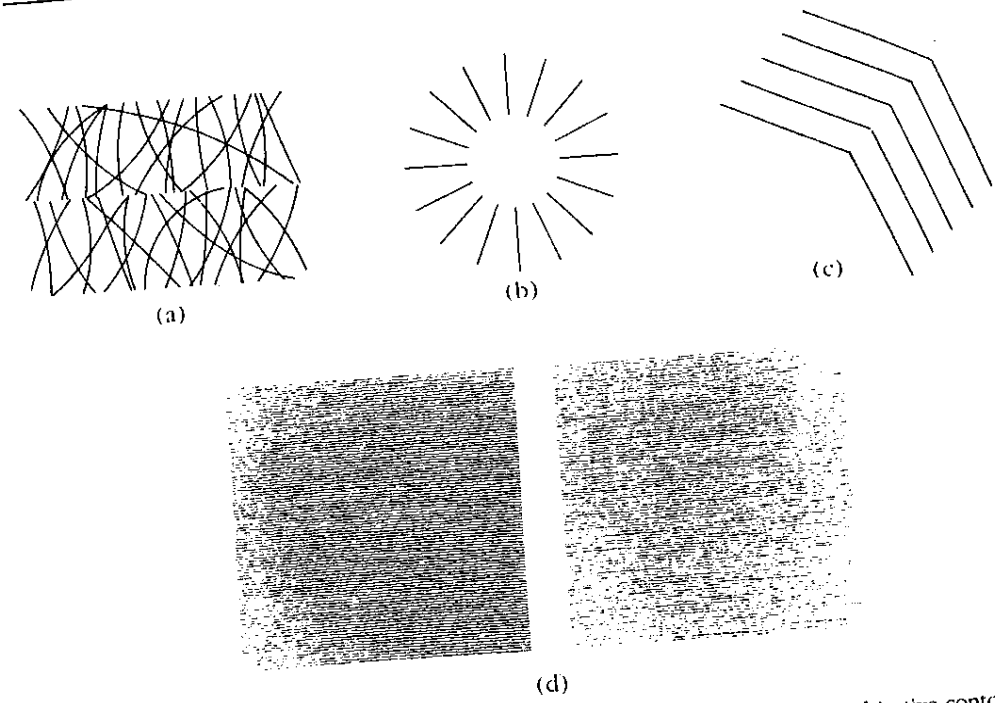


Figure 2–25. Examples of terminations being made explicit. In (a) and (b) subjective contours are constructed by joining termination points. In (c), points of discontinuity in orientation are seen to have a linear arrangement. In the stereogram (d), terminations or discontinuities in the small horizontal 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

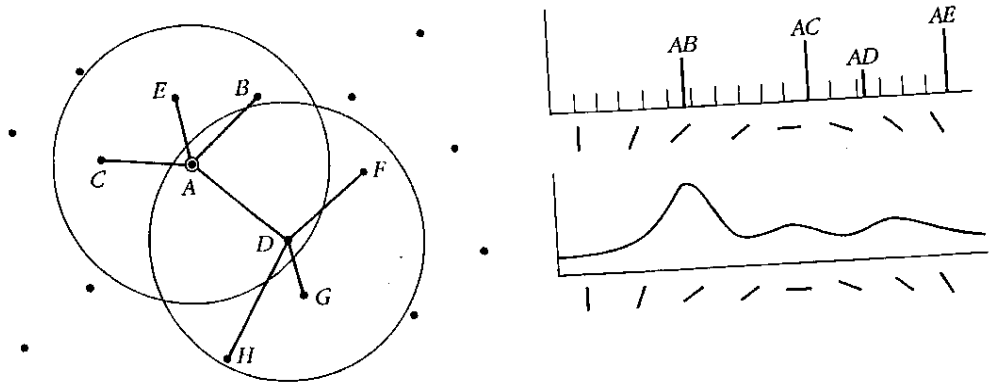
and also to explore the idea of tokens in the images—the supposed entities that virtual lines were thought to connect.

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:

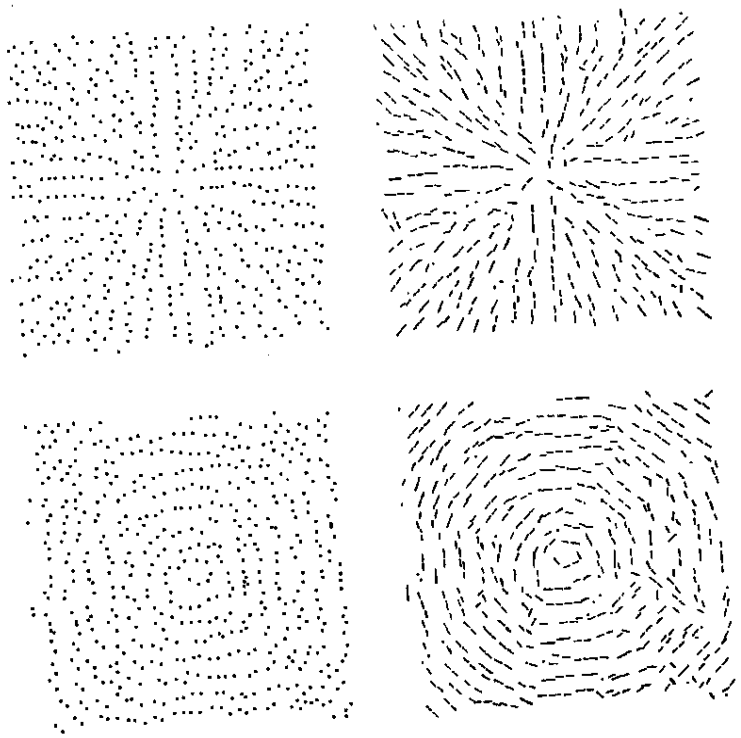
1. The local orientation organization in a Glass pattern can be recovered by a purely local algorithm, illustrated in Figure 2–26. The basic idea is to connect neighboring points with virtual lines and then to search locally among these virtual lines for the predominant orientation. By splitting patterns 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)

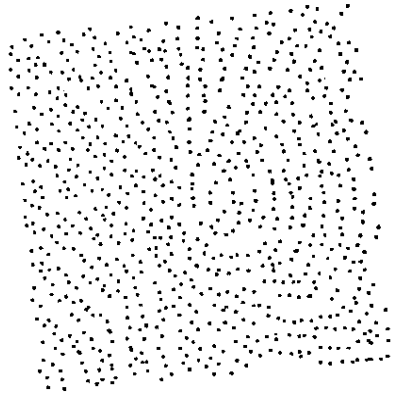


Figure 2–27. The algorithm used by our visual systems for detecting the local orientation structure is also a local one, as one can see from this pattern. Different portions of this pattern have different local orientation structures, and this can easily be discerned. (Reprinted by permission from K. 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.

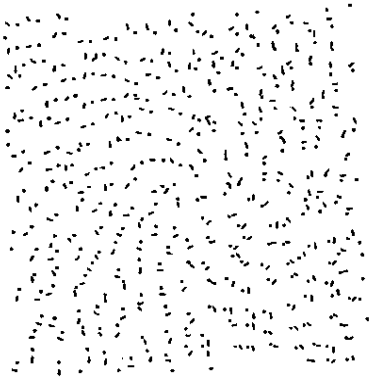


Figure 2–28. As we saw in Figure 2–3, the tokens in the two patterns do not have to be identical in order for their spatial organization to be apparent. They do, however, have to be similar. (Reprinted by permission from K. A. Stevens, "Computation of locally parallel structure," *Biol. Cybernetics 29*,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

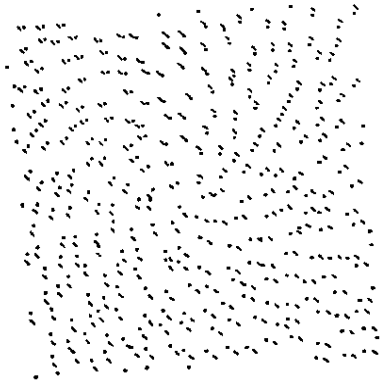


Figure 2–29. Here the superimposed pattern consists of small lines all having the same orientation. Interestingly, we perceive a kind of rivalry between this orientation and the orientation due to the spatial organization of the pattern. (Reprinted by permission from K. A. Stevens, "Computation of locally 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 Fig-

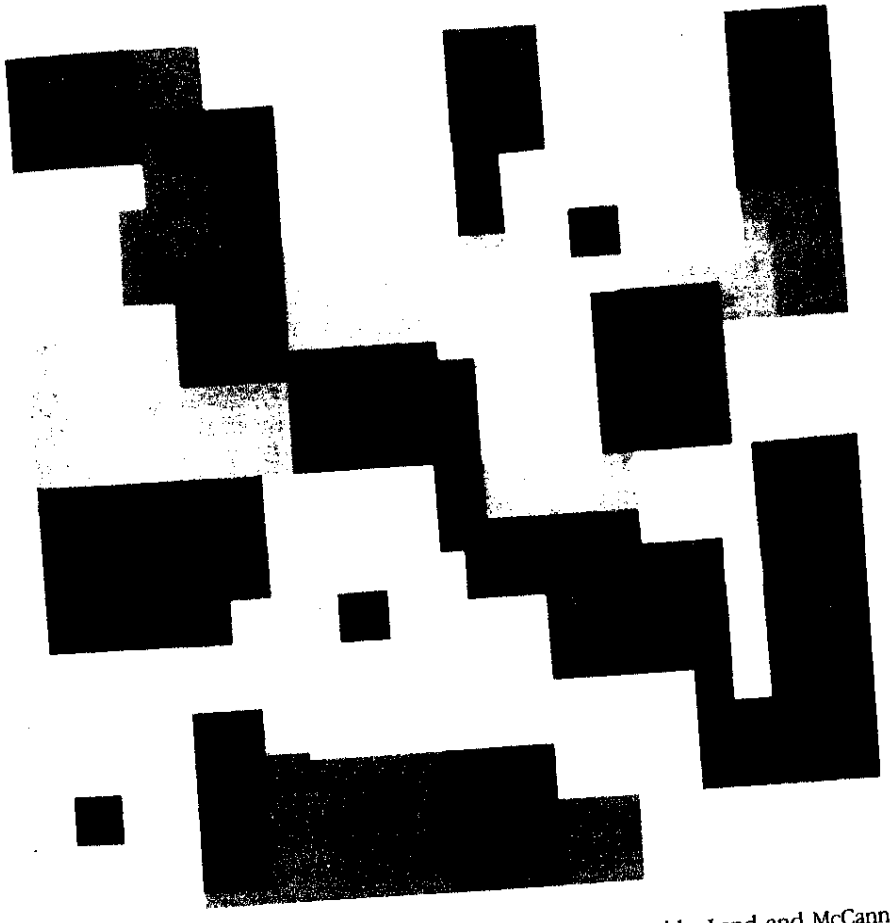


Figure 2–30. A Mondrian stimulus of the sort introduced by Land and McCann and used by Ullman in his study of fluorescence.

ure 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.

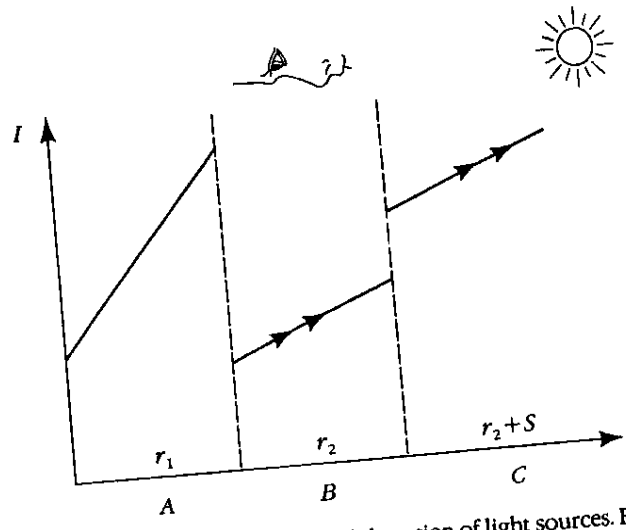


Figure 2–31. The idea behind the visual detection of light sources. Regions A and B have reflectances r_1 and r_2, and give rise to intensities I as shown. The value of I and of its gradient ∇I change together between A and B, so that $\nabla I/I$ remains constant. At C, however, a source S is added. 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

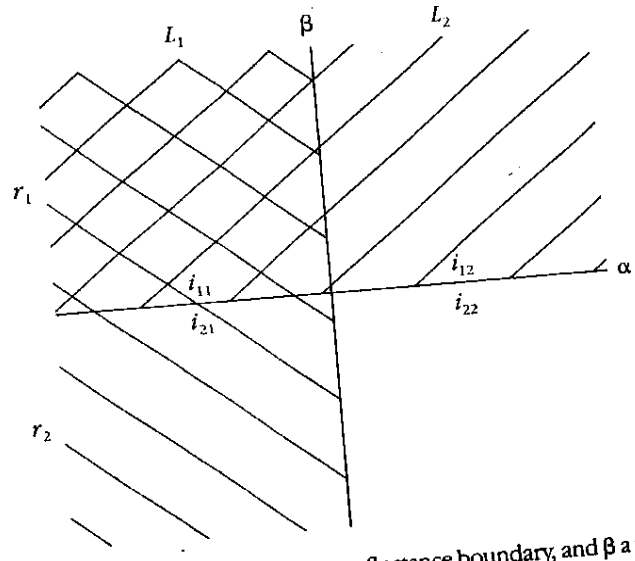


Figure 2–32. Boundary α represents a reflectance boundary, and β a transparency boundary. The quantities r_i represent reflectances; L_i, luminances; and i_{ij} are measured intensity values (for $i,j=1,2$).

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

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

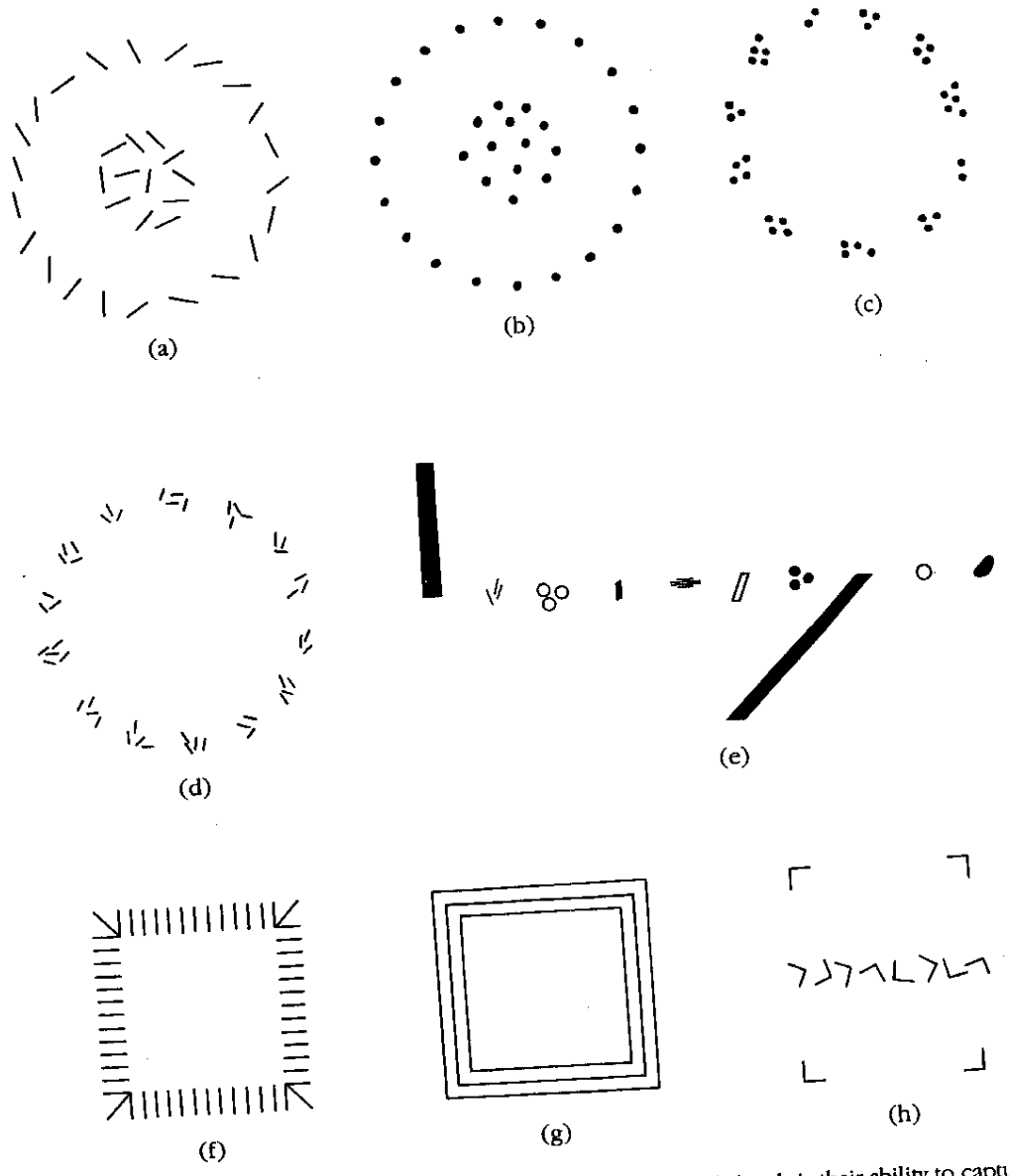


Figure 2–33. The essence of the higher primitives in the primal sketch is their ability to capture a wide range of image items as a group or token and their ability to be arranged into groups and boundaries. These diagrams show some examples of the different ways of defining place tokens and of grouping them. In each one a small line, a group of lines, or a group of dots is being combined 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-

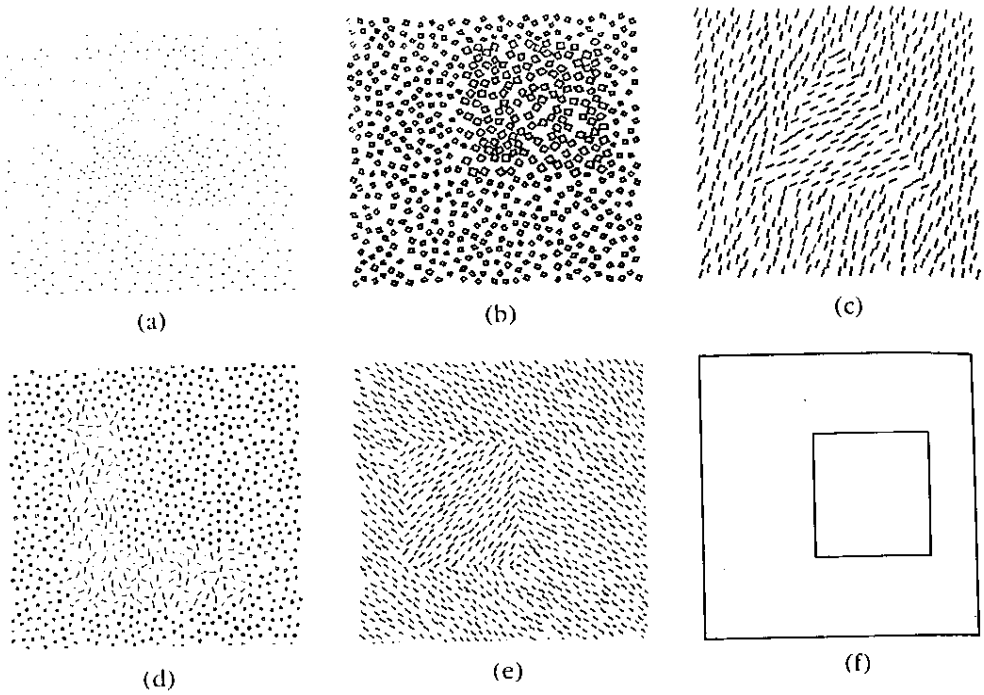


Figure 2–34. Another important aspect of the primal sketch is the construction of boundaries between regions on the basis of cues that could be caused by discontinuities in surface orientation or distance from the viewer. All examples in this figure are due to M. Riley, and they give rise psychophysically 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

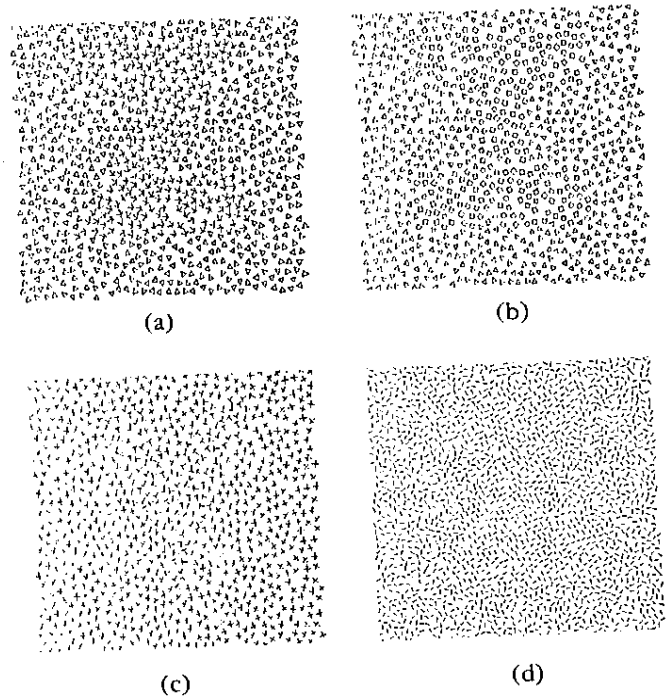


Figure 2–35. These examples, also provided by M. Riley, show texture differences that could not be of purely geometrical origin. They do not give rise psychophysically to boundaries in the sense defined in the text, even though we are sometimes able to say that one region differs from another in some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