studying their wiring and interactions, but in order to understand *why* the receptive fields are as they are—why they are circularly symmetrical and why their excitatory and inhibitory regions have characteristic shapes and distributions—we have to know a little of the theory of differential operators, band-pass channels, and the mathematics of the uncertainty principle (see Chapter 2).

Perhaps it is not surprising that the very specialized empirical disciplines of the neurosciences failed to appreciate fully the absence of computational theory; but it is surprising that this level of approach did not play a more forceful role in the early development of artificial intelligence. For far too long, a heuristic program for carrying out some task was held to be a theory of that task, and the distinction between what a program did and how it did it was not taken seriously. As a result, (1) a style of explanation evolved that invoked the use of special mechanisms to solve particular problems, (2) particular data structures, such as the lists of attribute value pairs called property lists in the LISP programing language, were held to amount to theories of the representation of knowledge, and (3) there was frequently no way to determine whether a program would deal with a particular case other than by running the program.

Failure to recognize this theoretical distinction between *what* and *how* also greatly hampered communication between the fields of artificial intelligence and linguistics. Chomsky's (1965) theory of transformational grammar is a true computational theory in the sense defined earlier. It is concerned solely with specifying what the syntactic decomposition of an English sentence should be, and not at all with how that decomposition should be achieved. Chomsky himself was very clear about this—it is roughly his distinction between competence and performance, though his idea of performance did include other factors, like stopping in midutterance—but the fact that his theory was defined by transformations, which look like computations, seems to have confused many people. Winograd (1972), for example, felt able to criticize Chomsky's theory on the grounds that it cannot be inverted and so cannot be made to run on a computer; I had heard reflections of the same argument made by Chomsky's colleagues in linguistics as they turn their attention to how grammatical structure might actually be computed from a real English sentence.

The explanation is simply that finding algorithms by which Chomsky's theory may be implemented is a completely different endeavor from formulating the theory itself. In our terms, it is a study at a different level, and both tasks have to be done. This point was appreciated by Marcus (1980), who was concerned precisely with how Chomsky's theory can be realized and with the kinds of constraints on the power of the human grammatical processor that might give rise to the structural constraints in syntax that Chomsky found. It even appears that the emerging "trace" theory of grammar (Chomsky and Lasnik, 1977) may provide a way of synthesizing the two approaches—showing that, for example, some of the rather ad hoc restrictions that form part of the computational theory may be consequences of weaknesses in the computational power that is available for implementing syntactical decoding.

The Approach of J. J. Gibson

In perception, perhaps the nearest anyone came to the level of computational theory was Gibson (1966). However, although some aspects of his thinking were on the right lines, he did not understand properly what information processing was, which led him to seriously underestimate the complexity of the information-processing problems involved in vision and the consequent subtlety that is necessary in approaching them.

Gibson's important contribution was to take the debate away from the philosophical considerations of sense-data and the affective qualities of sensation and to note instead that the important thing about the senses is that they are channels for perception of the real world outside or, in the case of vision, of the visible surfaces. He therefore asked the critically important question, How does one obtain constant perceptions in everyday life on the basis of continually changing sensations? This is exactly the right question, showing that Gibson correctly regarded the problem of perception as that of recovering from sensory information "valid" properties of the external world. His problem was that he had a much oversimplified view of how this should be done. His approach led him to consider higher-order variables—stimulus energy, ratios, proportions, and so on—as "invariants" of the movement of an observer and of changes in stimulation intensity.

"These invariants," he wrote, "correspond to permanent properties of the environment. They constitute, therefore, information about the permanent environment." This led him to a view in which the function of the brain was to "detect invariants" despite changes in "sensations" of light, pressure, or loudness of sound. Thus, he says that the "function of the brain, when looped with its perceptual organs, is not to decode signals, nor to interpret messages, nor to accept images, nor to *organize* the sensory input or to *process* the data, in modern terminology. It is to seek and extract information about the environment from the flowing array of ambient energy," and he thought of the nervous system as in some way "resonating" to these invariants. He then embarked on a broad study of animals in their environments, looking for invariants to which they might resonate. This was the basic idea behind the notion of ecological optics (Gibson, 1966, 1979).

Although one can criticize certain shortcomings in the quality of Gibson's analysis, its major and, in my view, fatal shortcoming lies at a deeper level and results from a failure to realize two things. First, the detection of physical invariants, like image surfaces, is exactly and precisely an information-processing problem, in modern terminology. And second, he vastly underrated the sheer difficulty of such detection. In discussing the recovery of three-dimensional information from the movement of an observer, he says that "in motion, perspective information alone can be used" (Gibson, 1966, p. 202). And perhaps the key to Gibson is the following:

> The detection of non-change when an object moves in the world is not as difficult as it might appear. It is only made to seem difficult when we assume that the perception of constant dimensions of the object must depend on the correcting of sensations of inconstant form and size. The information for the constant dimension of an object is normally carried by invariant relations in an optic array. Rigidity is *specified.* (emphasis added)

Yes, to be sure, but *how?* Detecting physical invariants is just as difficult as Gibson feared, but nevertheless we can do it. And the only way to understand how is to treat it as an information-processing problem.

The underlying point is that visual information processing is actually very complicated, and Gibson was not the only thinker who was misled by the apparent simplicity of the act of seeing. The whole tradition of philosophical inquiry into the nature of perception seems not to have taken seriously enough the complexity of the information processing involved. For example, Austin's (1962) *Sense and Sensibilia* entertainingly demolishes the argument, apparently favored by earlier philosophers, that since we are sometimes deluded by illusions (for example, a straight stick appears bent if it is partly submerged in water), we see sense-data rather than material things. The answer is simply that usually our perceptual processing does run correctly (it delivers a true description of what is there), but although evolution has seen to it that our processing allows for many changes (like inconstant illumination), the perturbation due to the refraction of light by water is not one of them. And incidentally, although the example of the bent stick has been discussed since Aristotle, I have seen no philosphical inquiry into the nature of the perceptions of, for instance, a heron, which is a bird that feeds by pecking up fish first seen from above the water surface. For such birds the visual correction might be present.

Anyway, my main point here is another one. Austin (1962) spends much time on the idea that perception tells one about real properties of the external world, and one thing he considers is "real shape," (p. 66), a notion which had cropped up earlier in his discussion of a coin that "looked elliptical" from some points of view. Even so,

> it had a real shape which remained unchanged. But coins in fact are rather special cases. For one thing their outlines are well defined and very highly stable, and for another they have a known and a nameable shape. But there are plenty of things of which this is not true. What is the real shape of a cloud? . . . or of a cat? Does its real shape change whenever it moves? If not, in what posture *is* its real shape on display? Furthermore, is its real shape such as to be fairly smooth outlines, or must it be finely enough serrated to take account of each hair? *It is pretty obvious that there is no answer to these questions—no rules according to which, no procedure by which, answers are to be determined.* (emphasis added), (p. 67)

But there *are* answers to these questions. There are ways of describing the shape of a cat to an arbitrary level of precision (see Chapter 5), and there are rules and procedures for arriving at such descriptions. That is exactly what vision is about, and precisely what makes it complicated.

1.3 A REPRESENTATIONAL FRAMEWORK FOR VISION

Vision is a process that produces from images of the external world a description that is useful to the viewer and not cluttered with irrelevant information (Marr, 1976; Marr and Nishihara, 1978). We have already seen that a process may be thought of as a mapping from one representation to another, and in the case of human vision, the initial representation is in no doubt—it consists of arrays of image intensity values as detected by the photoreceptors in the retina.

It is quite proper to think of an image as a representation; the items that are made explicit are the image intensity values at each point in the array, which we can conveniently denote by $I\ (x,y)$ at coordinate (x,y). In order to simplify our discussion, we shall neglect for the moment the fact that there are several different types of receptor, and imagine instead that there is just one, so that the image is black-and-white. Each value of $I\ (x,y)$ thus specifies a particular level of gray; we shall refer to each detector as a picture element or *pixel* and to the whole array I as an image.

But what of the output of the process of vision? We have already agreed that it must consist of a useful description of the world, but that requirement is rather nebulous. Can we not do better? Well, it is perfectly true that, unlike the input, the result of vision is much harder to discern, let and an important aspect of this new approach is that it makes quite concrete proposals about what that end is. But before we begin that discussion, let us step back a little and spend a little time formulating the more general issues that are raised by these questions.

The Purpose of Vision

The usefulness of a representation depends upon how well suited it is to the purpose for which it is used. A pigeon uses vision to help it navigate, fly, and seek out food. Many types of jumping spider use vision to tell the difference between a potential meal and a potential mate. One type, for example, has a curious retina formed of two diagonal strips arranged in a V. If it detects a red V on the back of an object lying in front of it, the spider has found a mate. Otherwise, maybe a meal. The frog, as we have seen, detects bugs with its retina; and the rabbit retina is full of special gadgets, including what is apparently a hawk detector, since it responds well to the pattern made by a preying hawk hovering overhead. Human vision, on the other hand, seems to be very much more general, although it clearly contains a variety of special-purpose mechanisms that can, for example, direct the eye toward an unexpected movement in the visual field or cause one to blink or otherwise avoid something that approaches one's head too quickly.

Vision, in short, is used in such a bewildering variety of ways that the visual systems of different animals must differ significantly from one another. Can the type of formulation that I have been advocating, in terms of representations and processes, possibly prove adequate for them all? I think so. The general point here is that because vision is used by different animals for such a wide variety of purposes, it is inconceivable that all seeing animals use the same representations; each can confidently be expected to use one or more representations that are nicely tailored to the owner's purposes.

As an example, let us consider briefly a primitive but highly efficient visual system that has the added virtue of being well understood. Werner Reichardt's group in Tübingen has spent the last 14 years patiently unraveling the visual flight-control system of the housefly, and in a famous collaboration, Reichardt and Tomaso Poggio have gone far toward solving the problem (Reichardt and Poggio, 1976, 1979; Poggio and Reichardt, 1976). Roughly speaking, the fly's visual apparatus controls its flight through a collection of about five independent, rigidly inflexible, very fast responding systems (the time from visual stimulus to change of torque is only 21 ms). For example, one of these systems is the landing system; if the visual field "explodes" fast enough (because a surface looms nearby), the fly automatically "lands" toward its center. If this center is above the fly, the fly automatically inverts to land upside down. When the feet touch, power to the wings is cut off. Conversely, to take off, the fly jumps; when the feet no longer touch the ground, power is restored to the wings, and the insect flies again.

In-flight control is achieved by independent systems controlling the fly's vertical velocity (through control of the lift generated by the wings) and horizontal direction (determined by the torque produced by the asymmetry of the horizontal thrust from the left and right wings). The visual input to the horizontal control system, for example, is completely described by the two terms

$$r(\psi)\dot{\psi} + D(\psi)$$

where r and D have the form illustrated in Figure 1–6. This input describes how the fly tracks an object that is present at angle ψ in the visual field and has angular velocity $\dot{\psi}$. This system is triggered to track objects of a certain angular dimension in the visual field, and the motor strategy is such that if the visible object was another fly a few inches away, then it would be

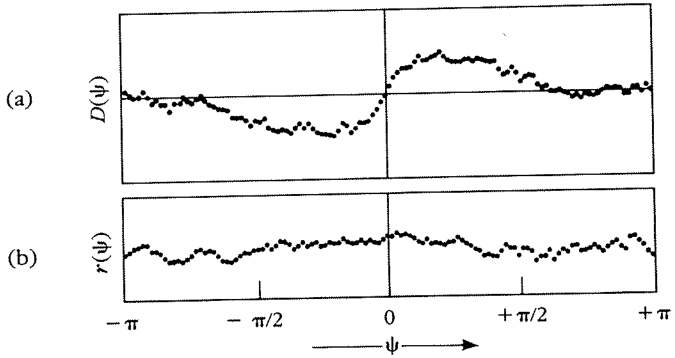


Figure 1–6. The horizontal component of the visual input R to the fly's flight system is described by the formula $R = D(\psi) - r(\psi)\,\dot{\psi}$, where ψ is the direction of the stimulus and $\dot{\psi}$ is its angular velocity in the fly's visual field. $D(\psi)$ is an odd function, as shown in (a), which has the effect of keeping the target centered in the fly's visual field; $r(\psi)$ is essentially constant as shown in (b).

intercepted successfully. If the target was an elephant 100 yd away, interception would fail because the fly's built-in parameters are for another fly nearby, not an elephant far away.

Thus, fly vision delivers a representation in which at least these three things are specified: (1) whether the visual field is looming sufficiently fast that the fly should contemplate landing; (2) whether there is a small patch—it could be a black speck or, it turns out, a textured figure in front of a textured ground—having some kind of motion relative to its background; and if there is such a patch, (3) ψ and $\dot{\psi}$ for this patch are delivered to the motor system. And that is probably about 60% of fly vision. In particular, it is extremely unlikely that the fly has any explicit representation of the visual world around him—no true conception of a surface, for example, but just a few triggers and some specifically fly-centered parameters like ψ and $\dot{\psi}$.

It is clear that human vision is much more complex than this, although it may well incorporate subsystems not unlike the fly's to help with specific and rather low-level tasks like the control of pursuit eye movements. Nevertheless, as Poggio and Reichardt have shown, even these simple systems can be understood in the same sort of way, as information-processing tasks. And one of the fascinating aspects of their work is how they have managed not only to formulate the differential equations that accurately describe the visual control system of the fly but also to express these equations, using the Volterra series expansion, in a way that gives direct information about the minimum possible complexity of connections of the underlying neuronal networks.

Advanced Vision

Visual systems like the fly's serve adequately and with speed and precision the needs of their owners, but they are not very complicated; very little objective information about the world is obtained. The information is all very much subjective—the angular size of the stimulus as the fly sees it rather than the objective size of the object out there, the angle that the object has in the fly's visual field rather than its position relative to the fly or to some external reference, and the object's angular velocity, again in the fly's visual field, rather than any assessment of its true velocity relative to the fly or to some stationary reference point.

One reason for this simplicity must be that these facts provide the fly with sufficient information for it to survive. Of course, the information is not optimal and from time to time the fly will fritter away its energy chasing a falling leaf a medium distance away or an elephant a long way away as a direct consequence of the inadequacies of its perceptual system. But this

apparently does not matter very much—the fly has sufficient excess energy for it to be able to absorb these extra costs. Another reason is certainly that translating these rather subjective measurements into more objective qualities involves much more computation. How, then, should one think about more advanced visual systems—human vision, for example. What are the issues? What kind of information is vision really delivering, and what are the representational issues involved?

My approach to these problems was very much influenced by the fascinating accounts of clinical neurology, such as Critchley (1953) and Warrington and Taylor (1973). Particularly important was a lecture that Elizabeth Warrington gave at MIT in October 1973, in which she described the capacities and limitations of patients who had suffered left or right parietal lesions. For me, the most important thing that she did was to draw a distinction between the two classes of patient (see Warrington and Taylor, 1978). For those with lesions on the right side, recognition of a common object was possible *provided* that the patient's view of it was in some sense straightforward. She used the words *conventional* and *unconventional*—a water pail or a clarinet seen from the side gave "conventional" views but seen end-on gave "unconventional" views. If these patients recognized the object at all, they knew its name and its semantics—that is, its use and purpose, how big it was, how much it weighed, what it was made of, and so forth. If their view was unconventional—a pail seen from above, for example—not only would the patients fail to recognize it, but they would vehemently deny that it *could* be a view of a pail. Patients with left parietal lesions behaved completely differently. Often these patients had no language, so they were unable to name the viewed object or state its purpose and semantics. But they could convey that they correctly perceived its geometry—that is, its shape—even from the unconventional view.

Warrington's talk suggested two things. First, the representation of the shape of an object is stored in a different place and is therefore a quite different kind of thing from the representation of its use and purpose. And second, vision alone can deliver an internal description of the shape of a viewed object, even when the object was not recognized in the conventional sense of understanding its use and purpose.

This was an important moment for me for two reasons. The general trend in the computer vision community was to believe that recognition was so difficult that it required every possible kind of information. The results of this point of view duly appeared a few years later in programs like Freuder's (1974) and Tenenbaum and Barrow's (1976). In the latter program, knowledge about offices—for example, that desks have telephones on them and that telephones are black—was used to help "segment" out a black blob halfway up an image and "recognize" it as a telephone. Freuder's program used a similar approach to "segment" and recognize a hammer in a scene. Clearly, we do use such knowledge in real life; I once saw a brown blob quivering amongst the lettuce in my garden and correctly identified it as a rabbit, even though the visual information alone was inadequate. And yet here was this young woman calmly telling us not only that her patients could convey to her that they had grasped the shapes of things that she had shown them, even though they could not name the objects or say how they were used, but also that they could happily continue to do so even if she made the task extremely difficult visually by showing them peculiar views or by illuminating the objects in peculiar ways. It seemed clear that the intuitions of the computer vision people were completely wrong and that even in difficult circumstances shapes could be determined by vision alone.

The second important thing, I thought, was that Elizabeth Warrington had put her finger on what was somehow the quintessential fact of human vision—that it tells about shape and space and spatial arrangement. Here lay a way to formulate its purpose—building a description of the shapes and positions of things from images. Of course, that is by no means all that vision can do; it also tells about the illumination and about the reflectances of the surfaces that make the shapes—their brightnesses and colors and visual textures—and about their motion. But these things seemed secondary; they could be hung off a theory in which the main job of vision was to derive a representation of shape.

To the Desirable via the Possible

Finally, one has to come to terms with cold reality. Desirable as it may be to have vision deliver a completely invariant shape description from an image (whatever that may mean in detail), it is almost certainly impossible in only one step. We can only do what is possible and proceed from there toward what is desirable. Thus we arrived at the idea of a sequence of representations, starting with descriptions that could be obtained straight from an image but that are carefully designed to facilitate the subsequent recovery of gradually more objective, physical properties about an object's shape. The main stepping stone toward this goal is describing the geometry of the visible surfaces, since the information encoded in images, for example by stereopsis, shading, texture, contours, or visual motion, is due to a shape's local surface properties. The objective of many early visual computations is to extract this information.

However, this description of the visible surfaces turns out to be unsuitable for recognition tasks. There are several reasons why, perhaps the most prominent being that like all early visual processes, it depends critically on the vantage point. The final step therefore consists of transforming the viewer-centered surface description into a representation of the three-dimensional shape and spatial arrangement of an object that does not depend upon the direction from which the object is being viewed. This final description is object centered rather than viewer centered.

The overall framework described here therefore divides the derivation of shape information from images into three representational stages: (Table 1–1): (1) the representation of properties of the two-dimensional image,

Table 1–1. Representational framework for deriving shape information from images.

Name	Purpose	Primitives
Image(s)	Represents intensity.	Intensity value at each point in the image
Primal sketch	Makes explicit important information about the two-dimensional image, primarily the intensity changes there and their geometrical distribution and organization.	Zero-crossings Blobs Terminations and discontinuities Edge segments Virtual lines Groups Curvilinear organization Boundaries
2½-D sketch	Makes explicit the orientation and rough depth of the visible surfaces, and contours of discontinuities in these quantities in a viewer-centered coordinate frame.	Local surface orientation (the "needles" primitives) Distance from viewer Discontinuities in depth Discontinuities in surface orientation
3-D model representation	Describes shapes and their spatial organization in an object-centered coordinate frame, using a modular hierarchical representation that includes volumetric primitives (i.e., primitives that represent the volume of space that a shape occupies) as well as surface primitives.	3-D models arranged hierarchically, each one based on a spatial configuration of a few sticks or axes, to which volumetric or surface shape primitives are attached

such as intensity changes and local two-dimensional geometry; (2) the representation of properties of the visible surfaces in a viewer-centered coordinate system, such as surface orientation, distance from the viewer, and discontinuities in these quantities; surface reflectance; and some coarse description of the prevailing illumination; and (3) an object-centered representation of the three-dimensional structure and of the organization of the viewed shape, together with some description of its surface properties.

This framework is summarized in Table 1–1. Chapters 2 through 5 give a more detailed account.

PART II

Vision

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 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).

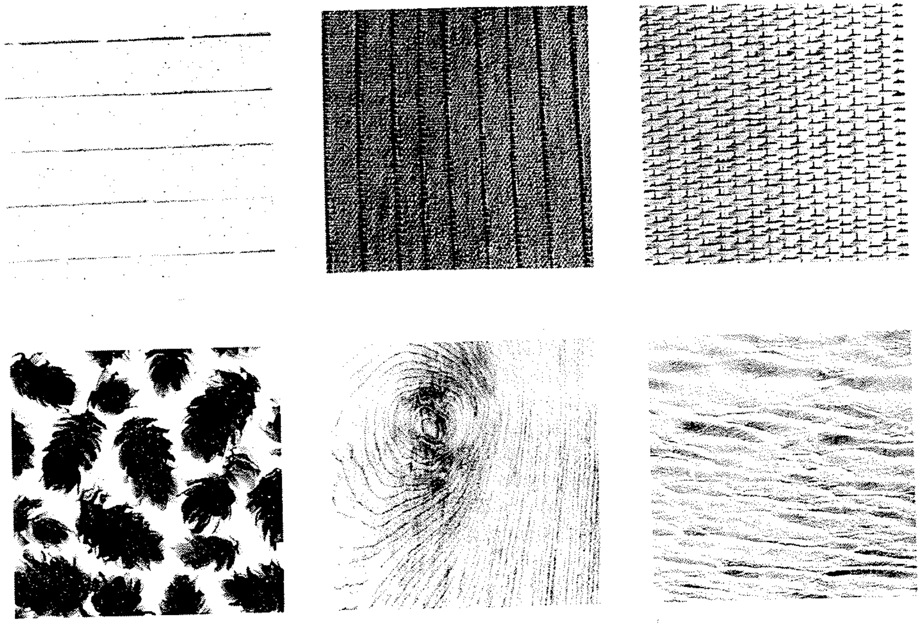


Figure 2–1. Some images of surfaces. Notice how different types of spatial organization occur almost independently at different scales. An important aspect of early vision is concerned with capturing these different organizations. (Reprinted by permission from Phil Brodatz, *Textures: A Photographic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

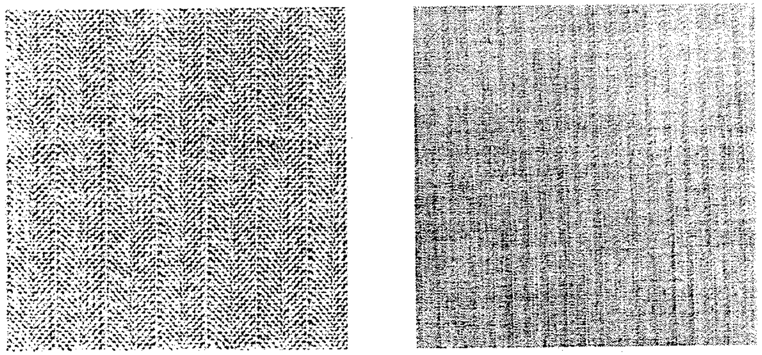


Figure 2–2. In a herringbone pattern such as this, a clear part of the spatial organization consists of the vertical stripes. These cannot be recovered by Fourier techniques such as band-pass filtering the images, but yield easily to grouping processes. (Reprinted by permission from Phil Brodatz, *Textures: 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

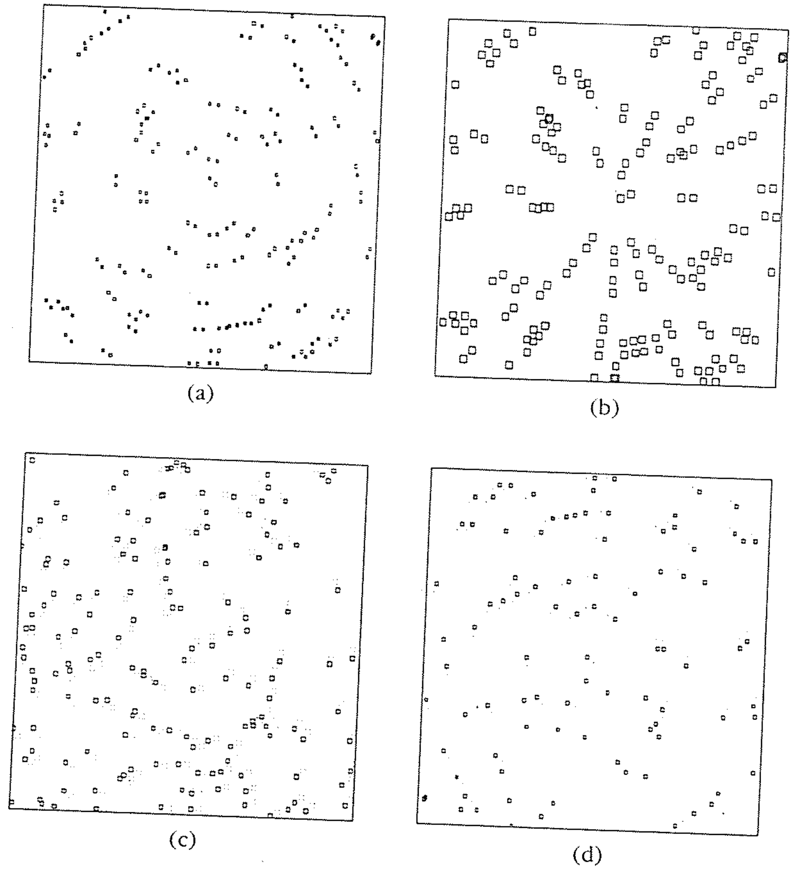


Figure 2–3. These displays are made by superimposing a random pattern of tokens on a slightly rotated or expanded copy of the same pattern. The tokens can be points or small squares (a) or larger squares (b). They do not have to be the same—in (c) one set consists of squares and the other set of four 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.

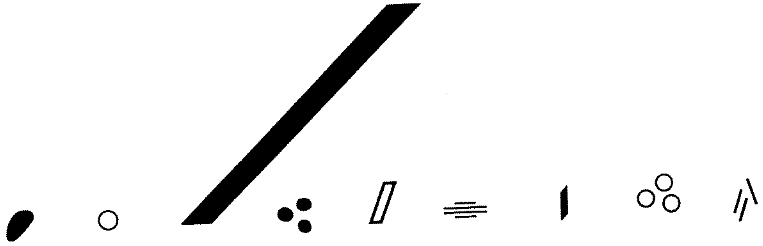


Figure 2–4. More evidence for place tokens. In this diagram every subgroup is defined differently, yet the collinearity of all of them is immediately apparent. This suggests that each group causes a place token to be created, whose collinearity is detected almost independently of the way the token is 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

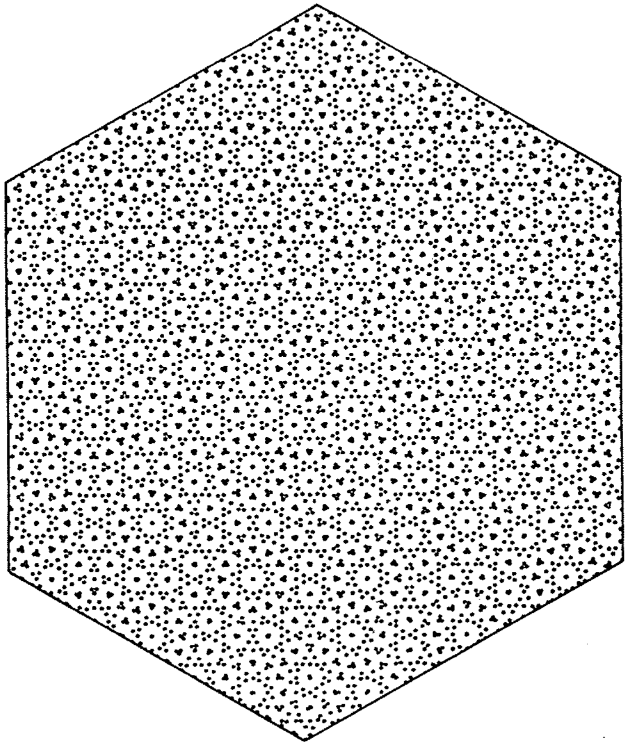


Figure 2–5. Evidence for the existence of active grouping processes. This pattern apparently seethes with activity as the rival organizations seem to compete with one another. (Reprinted by permission from J. L. Marroquin, "Human visual perception of structure," Master's thesis, Department of Electrical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

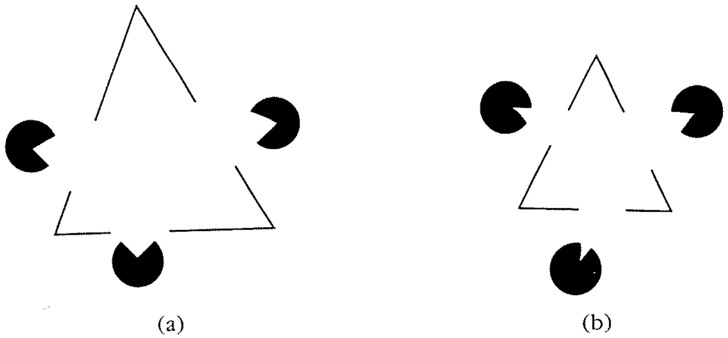


Figure 2–6. Subjective contours. The visual system apparently regards changes in depth as so important that they must be made explicit everywhere, including places where there is no direct visual evidence for them.

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).

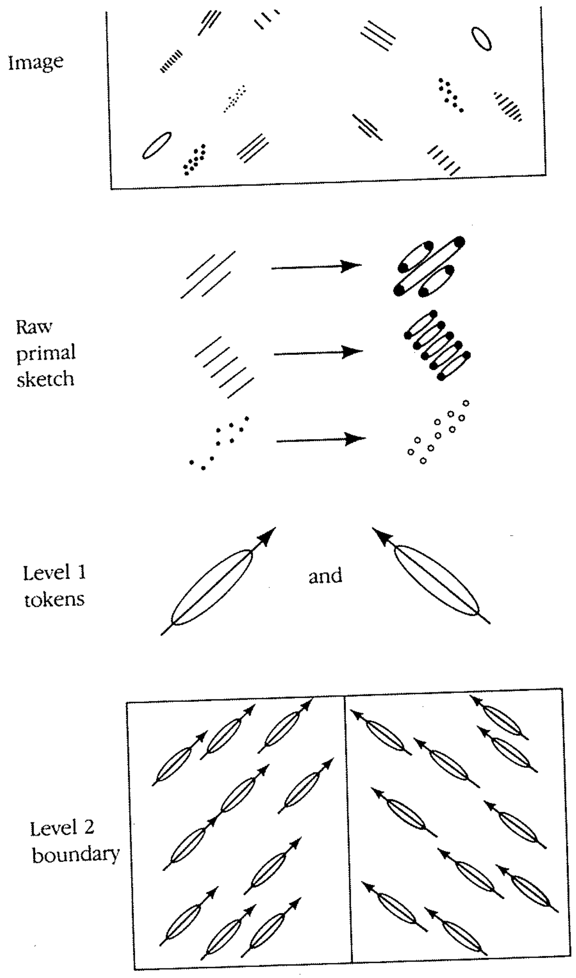


Figure 2–7. A diagrammatic representation of the descriptions of an image at different scales which together constitute the primal sketch. At the lowest level, the raw primal sketch faithfully follows the intensity changes and also represents terminations, denoted here by filled circles. At the next 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

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

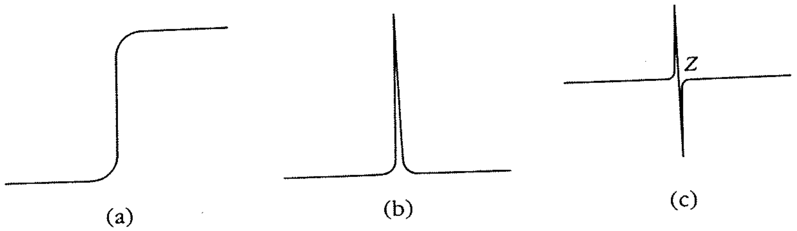


Figure 2–8. The notion of a zero-crossing. The intensity change (a) gives rise to a peak (b) in its first derivative and to a (steep) zero-crossing Z (c) in its second derivative.

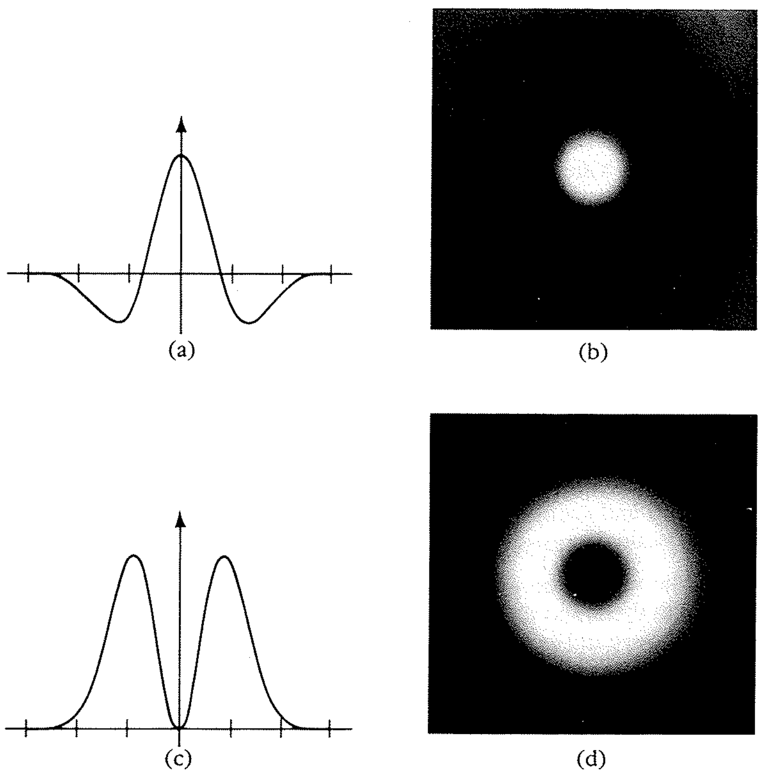
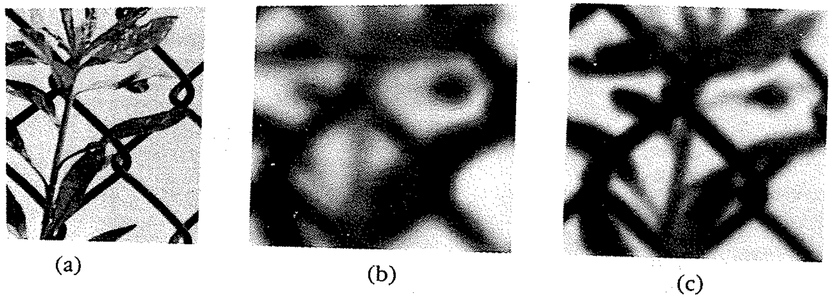


Figure 2–9. $\nabla^2 G$ is shown as a one-dimensional function (a) and in two-dimensions (b) using intensity to indicate the value of the function at each point. (c) and (d) show the Fourier transforms for the one- and two-dimensional cases respectively. (Reprinted by permission from D. Marr and E. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond. B 207,* pp. 187-217.)

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.



−10. Blurring images is the first step in detecting intensity changes in them. (a) In the image, intensity changes can take place over a wide range of scales, and no single operator ery efficient at detecting all of them. The problem is much simplified in an image that has rred with a Gaussian filter, because there is, in effect, an upper limit to the rate at which can take place. The first part of the edge detection process can be thought of as decom-ıe original image into a set of copies, each filtered with a different-sized Gaussian, and ecting the intensity changes separately in each. (b) The image filtered with a Gaussian = 8 pixels; in (c), $\sigma = 4$. The image is 320 by 320 elements. (Reprinted by permission larr and E. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond. B 207,* pp. 187-217.)

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

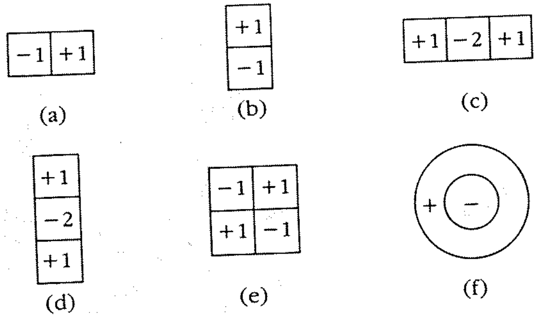
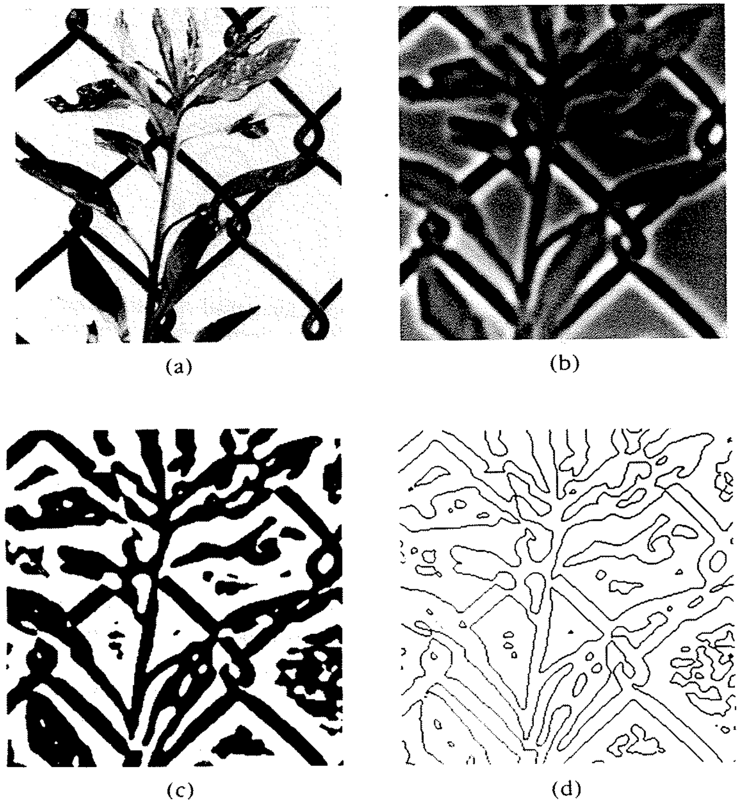


Figure 2–11. The spatial configuration of low-order differential operators. Operators like $\partial/\partial x$ can be roughly realized by filters with the receptive fields illustrated in the figure. (a) $\partial/\partial x$ can be thought of as measuring the difference between the values at two neighboring 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 $[(\partial I/\partial x)^2 + (\partial I/\partial y)^2]^{-1/2}$ 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 $w_{2-\mathrm{D}} = 8$ (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 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.

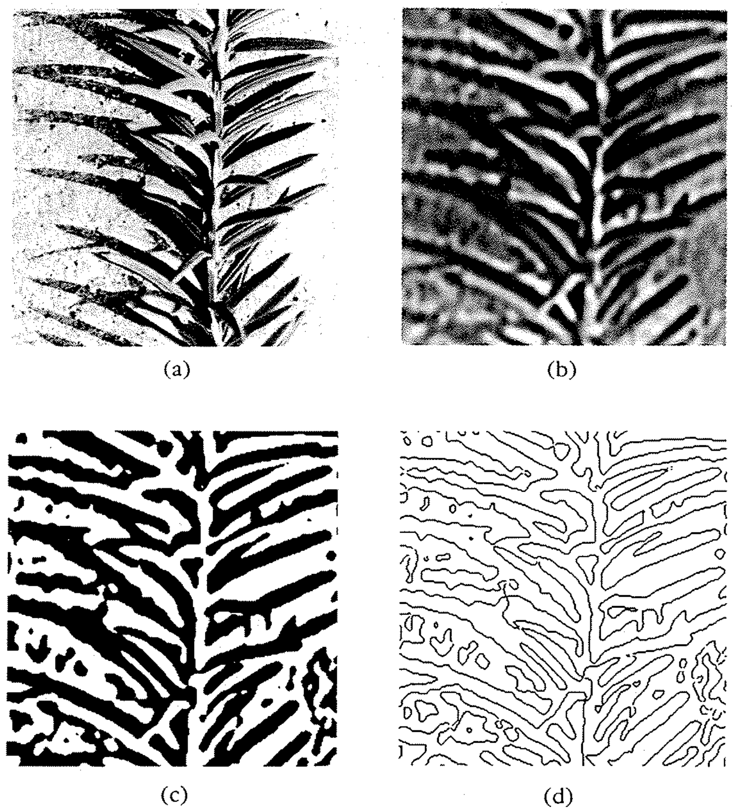


Figure 2–13.

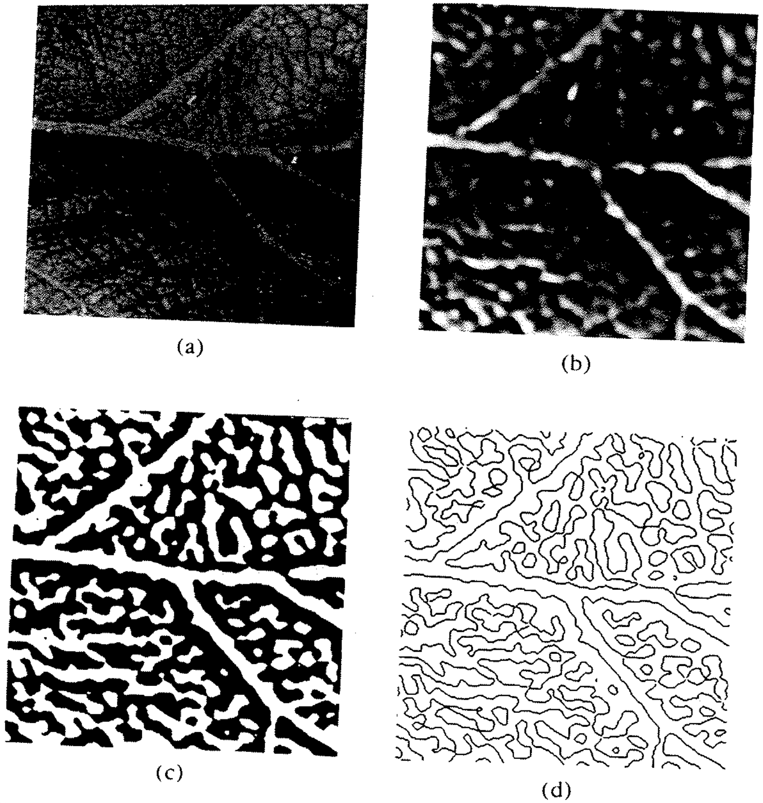


Figure 2–14.

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 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.

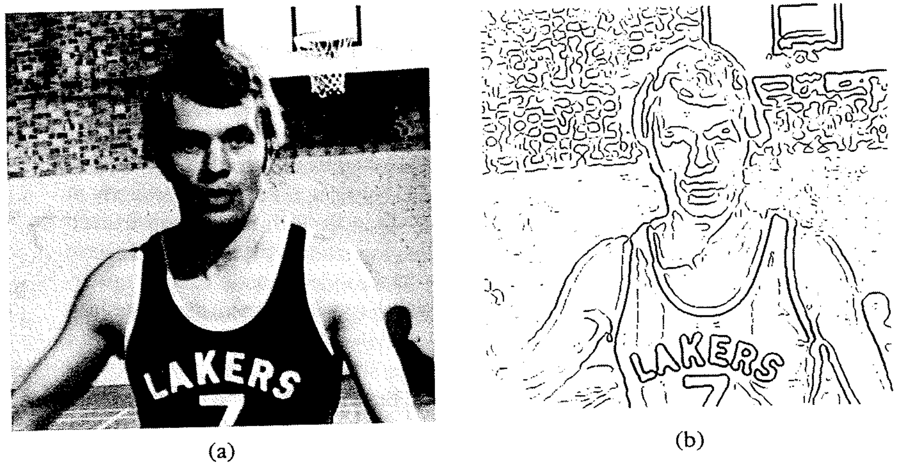


Figure 2–15. Another example of zero-crossings; here, the intensity of the lines has been made to vary with the slope of the zero-crossing, so that it is easier to see which lines correspond to the greater contrast. (Courtesy BBC Horizon.)

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 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).

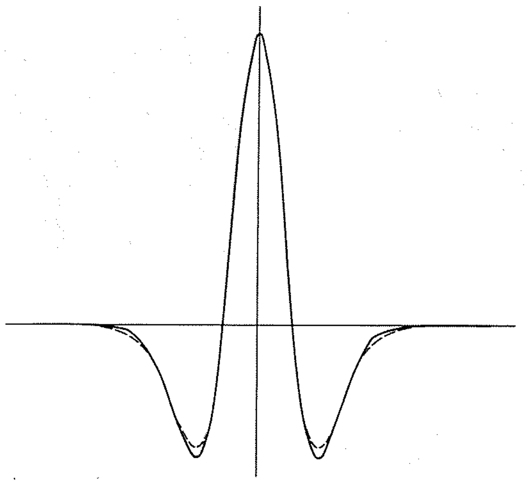


Figure 2–16. The best engineering approximation to $\nabla^2 G$ (shown by the continuous line), obtained by using the difference of two Gaussians (DOG), occurs when the ratio of the inhibitory to excitatory space constraints is about 1:1.6. The DOG is shown here dotted. The two profiles are very similar. (Reprinted by permission from D. Marr and E. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond. B 207*, pp. 187-217.)

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.

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 $w_{2-D}/\sqrt{2}$ apart, where w_{2-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

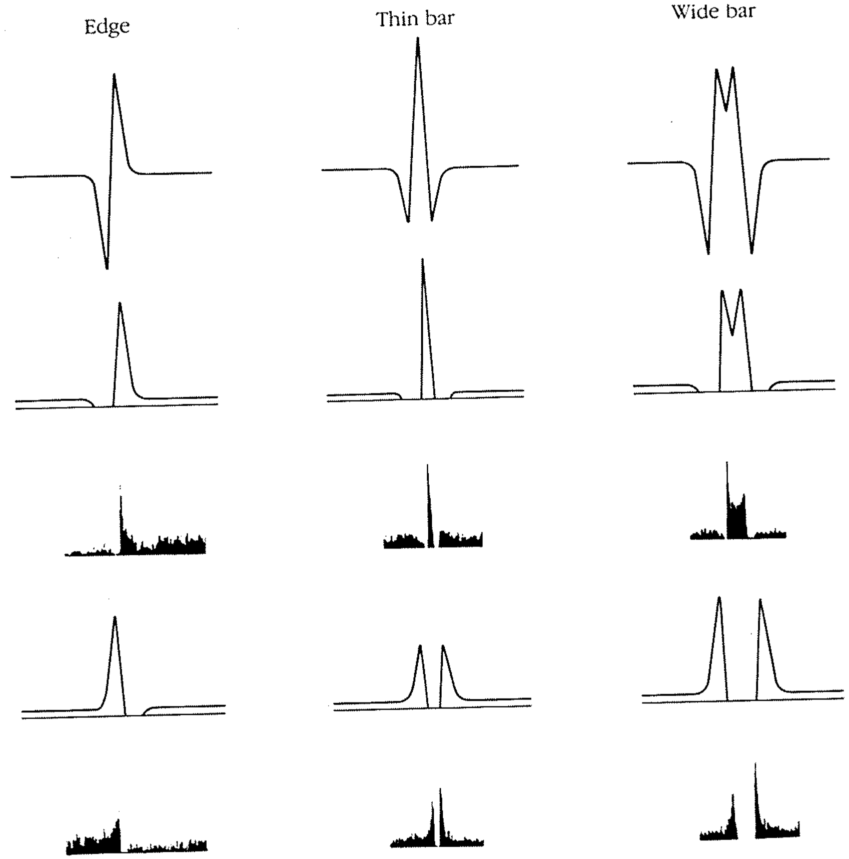


Figure 2–17. Comparison of the predicted responses of on- and off-center X cells with electrophysiological recordings. The first row shows the response to $\nabla^2 G * I$ for an isolated edge, a thin bar (bar width = $0.5w_{1-D}$, where $_{1-D}$ is the width of the central excitatory region of the receptive 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.)

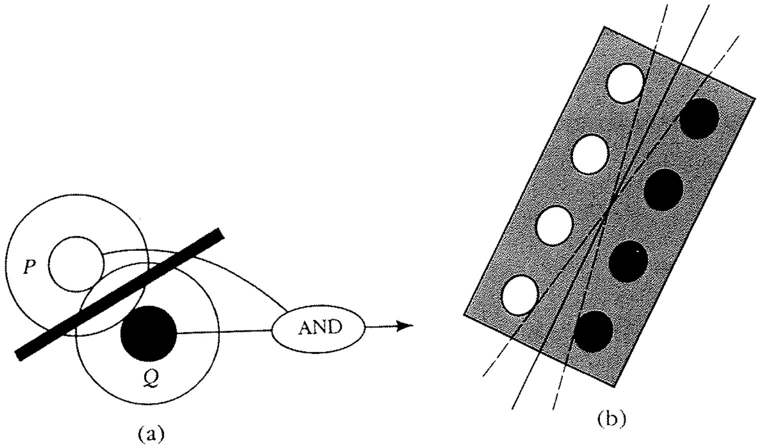


Figure 2–18. A mechanism for detecting oriented zero-crossing segments. In (a), if *P* represents an on-center geniculate X-cell receptive field, and *Q* an off-center, then a zero-crossing must pass between them if both are active. Hence, if they are connected to a logical AND gate as shown, the gate 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 207,* pp. 187-217.)

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.

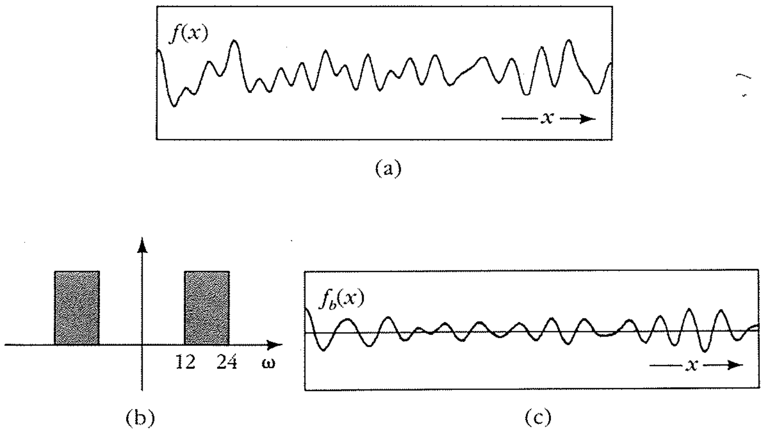


Figure 2–19. The meaning of Logan's theorem. (a) A stochastic, band-limited Gaussian signal $f(x)$. (b) The passband—in the frequency domain—of an ideal one-octave band-pass filter. (c) The result $f_b(x)$ of filtering (a) with the filter described by (b). Provided that (c) has no zeros in common with 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

output only when all of its inputs are positive.

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 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.

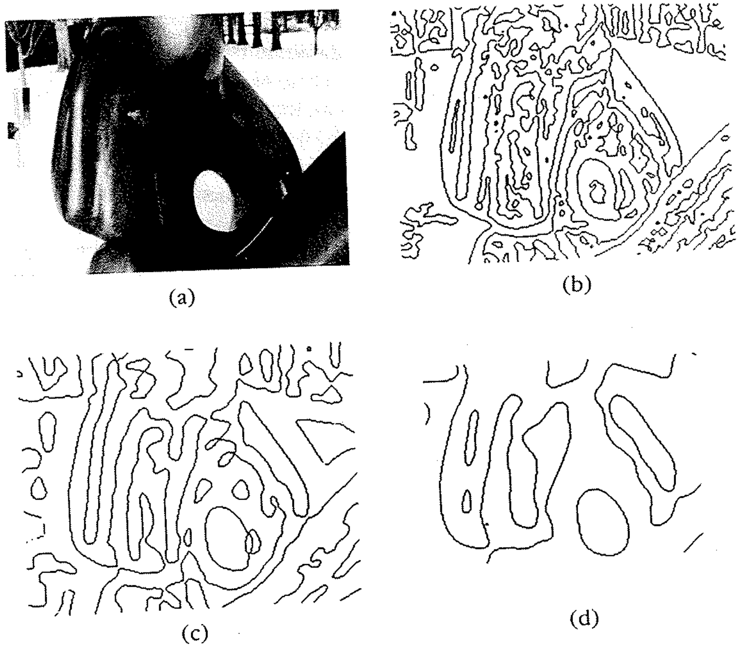


Figure 2–20. The image (a) has been convolved with $\nabla^2 G$ having $w_{2\text{-D}} = 2\sqrt{2}\,\sigma = 6$, 12, and 24 pixels. These filters span approximately the range of filters that operate in the human fovea. (b), (c), and (d) show the zero-crossings thus obtained. Notice the fine detail picked 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.)

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.

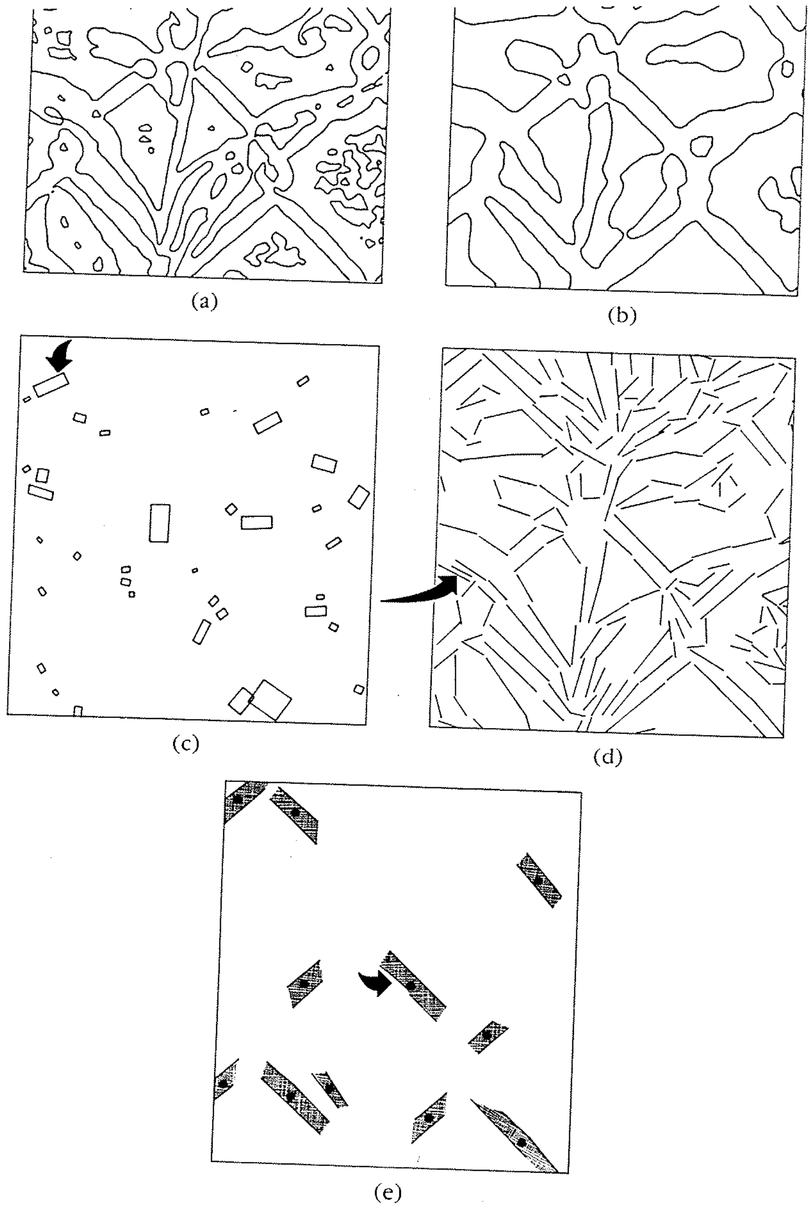


Figure 2–21. *(opposite)* The raw primal sketch as computed from two channels. (a), (b) The zero-crossings obtained from the image of Figure 2–12 by using masks with w_{2-D} = 9 and 18 pixels. Because there are no zero-crossings in the larger channel that do not correspond to zero-crossings in the smaller 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.

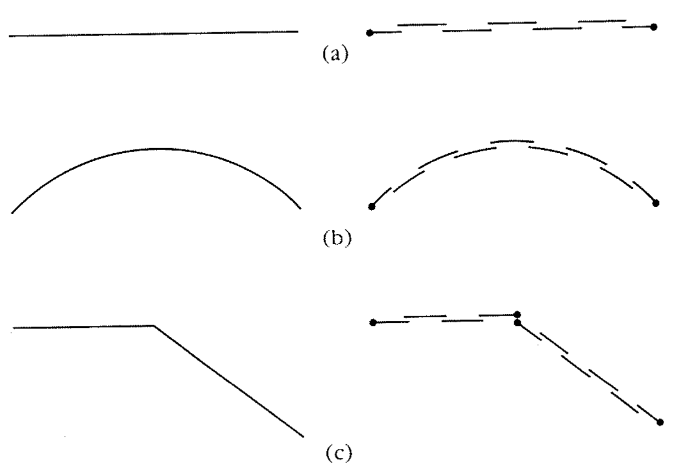


Figure 2–22. The raw primal sketch represents a straight line as a termination, several oriented segments, and a second termination (a). If the line is replaced by a smooth curve, the orientations of the inner segments will gradually change (b). If the line changes its orientation suddenly in the middle (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.

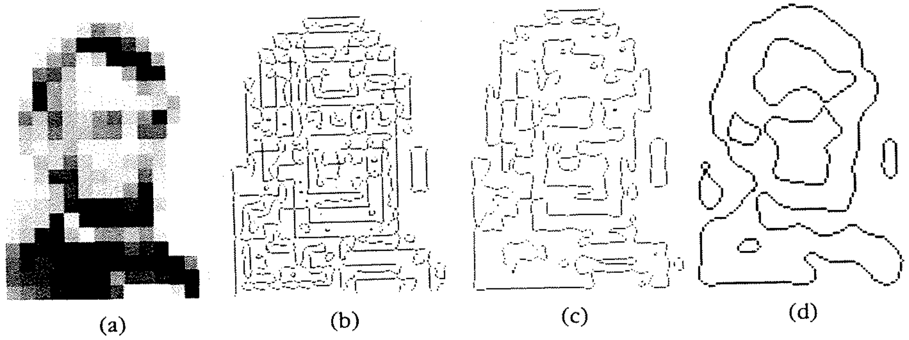


Figure 2–23. We cannot sense the primitive zero-crossings, only the description to which they give rise in the raw primal sketch. This can be seen in L. D. Harmon's discretely sampled and quantized image of Abraham Lincoln (a). No amount of voluntary effort allows us to see Lincoln without defocusing 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