Structure and Function in the early processing of visual information

Shimon Ullman

The Artificical Intelligence Laboratory
Massachusetts Institute of Technology

1. Introduction

A central notion in contemporary cognitive science is that mental processes involve computations defined over internal representations. This general view suggests a distinction between the study of the representation and computations performed by our cognitive systems on the one hand, and the physical brain mechanisms supporting these computations on the other. The two studies proceed along different paths, and neither is completely reducible to the other. It is the hope of cognitive science, however, that the studies of function and mechanism can complement each other, and that theories can be developed for various cognitive subsystems that will describe and explain their computational aspects, their underlying mechanisms, and the interactions between the two.

In this paper I shall describe some attempts to combine the study of brain mechanisms with computational considerations in the first stages of visual information processing. This work combines the contributions of many individuals, most notably the late David Marr, and a group of people who were fortunate to work with him, primarily at M.I.T's Artificial Intelligence Laboratory and Psychology Department.

2. Representing intensity changes in images

The first computational problem that arises in the early processing of visual information is the initial organization and representation of the input registered by the eyes. At the photoreceptors level, the input to the visual system consists of over 250 million light intensity measurements (registered by over 120 million cones and rods in each eye). This is an unwieldy large and unstructured set of measurements. We can therefore expect the visual system to construct a more economical representation of the input, that will make explicit the relevant information for later processing stages.

A reasonable candidate for the task is a representation that can be roughly described as an edge representation of the image. The idea is to make explicit the locations in the image where light intensity changes sharply from one level to another. The motivations for this type of a representation are (i) it will achieve a more concise description of the image than the original array of intensity values , and (ii) sharp changes in light intensity values usually have a physical significance. They are often associated, for example, with object boundaries, markings on objects' surfaces, and so forth An edge representation is therefore useful in making the transition from the domain of light intensities in the image to analyzing the physical structure of the visible environment. One general observation often raised in support of the edge representation approach is that many objects are recognizable from a sketch of their edges and contours alone, although in terms of the underlying light intensity distributions, the sketch and the original image are markedly different.

The representation of localized intensity changes is not the only approach that has been proposed for the first stages of analyzing visual information. One popular alternative is the Fourier analysis approach that received wide attention in the psychophysical literature following Campbell and Robson's [1968] discovery of spatial frequency tuned channels in the visual system. The approach presented here is in a sense a combination of the frequency channels and the edge detection approaches, but it is concerned primarily with the detection of intensity changes.

A large variety of techniques have been proposed in the past (primarily within the engineering field of image processing) for the detection of intensity changes in images. A major problem that has been discovered in the course of developing these techniques, is that significant intensity changes in an image can occur at a variety of scales. Some changes are gradual and smooth: they can also be described in frequency domain terminology as low frequency changes. Others are high frequency and sharply localized changes. To capture all of the significant intensity changes, it is possible to examine the image at a number of different resolutions, or scales. A low resolution "copy" will serve for capturing the gradual, gross changes, a high resolution "copy" for the fine details. Figure 1 shows an example of what it means for the same image to be examined at three different resolutions. The resolution decreases from 1a to 1c. It can be seen that in the lower resolution copies fine details are progressively blurred. The low resolution copy can be obtained by a process called gaussian filtering (and this filtering is in a sense optimal, see Marr & Hildreth 1980). This simply means that at every point a local average is taken of the intensity values, using a gaussian weighting function. The resolution of the resulting copy is controlled by the size of the gaussian. A larger guassian averages the intensity values over a wider neighborhood, and hence is less sensitive to fine details. The gaussian smoothing is also called in mathematical terms the convolution of the image with a gaussian filter, denoted by $G*I$ (where I is the image, G is the gaussian smoothing function).

As a result of the first operation we have a number of "copies" of the original image, at a number of different resolutions, as determined by the sizes of the gaussian filters (figure 2). The next step is to isolate the sharp intensity changes in each copy. We shall consider this problem first in the context of one-dimensional signals. In this case, the image I is a function of a single variable, denoted by x. A sharp change in the signal $I(x)$ can be defined as a peak in its first derivative, since the derivative, by definition, measures the signal's slope. From elementary calculus, peaks in the first derivative can also be located by zero-crossings of the second derivative (i.e. places where the second derivative changes sign). Mathematically the two criteria are equivalent, but the second characterization has certain advantages when two-dimensional signals are concerned [Marr & Hildreth 1980].

In summary, the localization of sharp changes is obtained by performing:

$$\frac{d^2}{dx^2}(G*I) \qquad (1)$$

The zero crossings in the output will indicate the locations of sharp intensity changes in the image at the scale determined by the gaussian.

This means that the image I is first passed through a gaussian smoothing function G, and then a second derivative of the result is taken. The two operations of scaling and differentiation be combined in a convenient manner. The combination is based on a mathematical identity that states that the order of differentiation and convolution can be changed without affecting the result. In mathematical notation:

$$\frac{d^2}{dx^2}(G*I) = (\frac{d^2}{dx^2}G)*I \qquad (2)$$

The implication is that the two operations can be collapsed into a single one: simply filter the image not through a gaussian function, but through $\frac{d^2}{dx^2}G$ (the second derivative of a gaussian). This function is shown in figure 3a (3b shows its fourier transform). The analogue in two dimensions would be a similar but circularly symmetric function which has the appearance of a "mexican hat". Mathematically, in the two-dimensional case the filter is $\nabla^2 G$, where ∇^2 is the laplacian and G a two-dimensional gaussian function.

The scheme is now straightforward: the representation of intensity changes is obtained from the zero-crossing in the result of passing the image through filters that have the shape of $\nabla^2 G$.

Those who have some familiarity with the physiology of the visual system would readily recognize the shape of these filters as corresponding to the shape of retinal ganglion receptive fields. In other words, the retinal structure can be viewed as approximating the convolution of the image with the $\nabla^2 G$ filters. (For more detail see Marr & Hildreth 1980, Marr & Ullman 1981).

Figure 4 shows examples of images following this retinal operation, and the resulting zero-crossings representations (generated by Ellen Hildreth). The first row shows two images prior to the filtering stage. The second row shows the images filtered through the retinal operation. It gives some idea of the form of the image as it travels up the optic nerve from the eye, via an intermediate station called the LGN, to the visual cortex in area 17 of the brain. The third row illustrates the resulting zero-crossing representations. Figure 5 shows an image (of a sculpture by Henry Moore) and its zero-crossing representation at three different resolutions.

Before turning to the physiological aspects of the zero-crossing representations, it will be of interest to note that zero-crossings in bandpass filters are known to be, in a sense, "rich in information". B. Logan of the Bell Laboratories has shown that a one-dimensional signal with a bandwidth of less than one octave can be completely reconstructed (up to an overall multiplicative constant) from its zero-crossings alone, provided that some simple conditions are met [Logan 1977]. It is not clear, however, whether the theorem can be extended to two dimensions, and under what conditions the one-octave restriction can be relaxed (this problem arises since the filters in the human visual system are probably more than an octave wide). If appropriate extensions along these lines can be made, it would imply that the zero-crossings provide not only a convenient representation that captures the significant aspects of the image, but also a complete one. That is, no essential information is lost by discarding the image and analyzing the zero-crossing representation alone. (See Marr, Poggio & Ullman 1979, for further discussion of this issue.)

3. The biological detection of zero-crossings

The analysis so far leads to the general suggestion that following the retinal operation the next step is to locate and represent a map of the zero-crossings in the output. If this suggestion is correct, then a main function of the primary visual cortex should be the construction of the zero-crossings representation. I shall next turn to consider briefly how zero-crossings may be detected by the mechanisms of the visual cortex.

The fibers of the optic nerve coming from the eye to the brain carry the image filtered through the $\nabla^2 G$ receptive fields (this is, of course, a computational idealization). This neural image is in fact carried by units of two complementary types, called on-center and off-center units. The off-center units are simply "inverted mexican hats" with negative center and positive surround. Let us now consider the retinal output in the vicinity of an edge. Figure 6a depicts a step edge, and 6b is the result of passing 6a through retinal-like receptive fields. This output contains both negative and positive values. In contrast, the optic nerve carries no negative values; the positive part of the signal is carried by the on-center units, and the negative part by the off-center ones. This means that within the system the zero-crossing itself is always flanked by two peaks of activity: of on-center cells on one side, and off-center cells on the other. The detection of a zero-crossing can easily be accomplished, therefore, by a simple combination of the on- and off-center units. When two adjacent units, one off-center, the other on-center, are active simultaneously, they indicate the existence of a zero-crossing running midway between them. Note that a point of zero value is detected in this scheme by detecting peaks of activity rather than zero activity.

The basic zero-crossing detector is shown in figure 7a. It is composed of the two sub-units (on- and off-center) combined with an "and" operation. This means that the two units are required to be active simultaneously to produce a response. The unit can be made oriented by combining a number of such detectors lying in a row (figure 7b). Such an oriented unit will exhibit many of the properties of cortical simple cells ("edge detectors") originally discovered by Hubel & Wiesel in the visual cortex of the cat [1962] and monkey [1968]. It will still lack, however, one fundamental property: cells in the visual cortex are also often selective for direction of motion. They respond well when their preferred stimulus moves in one direction, but little or not at all when it moves in the opposite direction.

4. Adding directional selectivity

With the addition of one subunit it is possible to make the basic zero-crossing detector directionally selective, and use it for the measure-

ment of visual motion. To see how, consider again the zero-crossing associated with an intensity edge (figure 6b). At the zero-crossing itself the current value is, of course, zero. It can be readily seen from the figure that if the profile now moves to the right, the value at this point will be increasing. If it move to the left, the value will be decreasing. By simply inspecting the sign of the temporal change it therefore becomes possible to determine the direction of motion. It is not difficult to establish that it is further possible to measure the speed of motion in the direction of the unit by comparing the slope of the zero-crossing and the rate of temporal change. The extra sub-unit should respond therefore to temporal changes. Ideally, it should behave like the time derivative of the signal, i.e. $\frac{d}{dt}(\nabla^2 G)$.

As it turns out, the population of retinal cells contain a natural candidate for this task. These are the so-called Y-type cells, originally discovered by Enroth-Cugell & Robson [1966]. This is a relatively small sub-population of cells that are known to be "transient". That is, they respond to a steady stimulus by a short and brisk response when the stimulus is turned on or off. The other major population of retinal cells are the sustained, X-type cells. Such a cell responds to a stationary stimulus with a sustained response that usually continues as long as the stimulus is present within its receptive field.

Our schematic model of the simplest directionally selective units is therefore constructed from three types of sub-units. As before, it has a row of on-center cells, and a row of off-center cells, both of the sustained type. In addition, it has an input from at least one transient Y-type unit (figure 7c). A more detailed discussion of this general scheme can be found in [Marr & Ullman 1981].

This general scheme for zero-crossing and motion detection was driven primarily by computational considerations. Physiologically, although Y-type units were often described as transient, it was not clear whether they can also be described as at least approximating the required time derivative operation. We therefore compared the response required by the computational scheme with physiological response (taken from Rodieck & Stone, 1965, Dreher & Sanderson 1973; see Marr & Ullman 1981 for details). Some comparisons are shown in figure 8 (for X cells) and 9 (for Y cells). The top row in figure 8 is the convolution of various profiles (edge, thin bar, wide bar) with $\nabla^2 G$. On-center cells are expected to carry the positive part of these profiles, and off-center the negative part. In the next two rows the positive part of the signal is compared with recordings from on-center cells, and in the last rows the negative part is compared with recordings from off-center cells. Similar comparisons are shown in figure 9 between the computational model, based on $\frac{d}{dt}(G*I)$, and physiological recordings. It can be seen that even in the cases where the profiles are rather complicated, the general agreement is good.

Finally in this section, figure 10 shows an example of applying the motion detection scheme described above to a moving random texture. Figures 10a and b show a pair of random dot patterns. A central square in 10a is shifted in 10b slightly to the right, while the backgrounds of the two figures are uncorrelated. When these figures are presented to human observers in a rapid alternation, the central square is immediately perceived to move back and forth against a background of uncorrelated motion. Figure 10c shows the zero-crossings representation of 10a. Figure 10d is the result of the motion analysis of the zero-crossing (the light dots indicate the direction of motion of the zero-crossings). In figure 10e the light dots where removed from the area where coherent motion (to the right) was found. The motion assignment was correct, with the exception of a few isolated points, and as a result the moving square was detected.

I have sketched above some aspects of an evolving theory of early visual information processing. The main goal has been not to present a comprehensive review of the theory, but to illustrate an attempt aimed at combining the study of structure and function in the early stages of visual perception. Major parts of the theory were consequently left out of the discussion, most notably, the use of the early representations in stereo vision [Marr & Poggio 1979, Grimson 1981].

Finally, I would like to end with two brief cautionary notes. The first has to do with the specific problem of analyzing image contours. Even if the zero-crossing analysis is along the right track, it provides only the first stages in the analysis of edges and image contours. Figure 11 illustrates examples of contours that are easily perceived but cannot be captured by any simple intensity-based analysis of the image. In figure 11a all the lines lie along the 45 deg. diagonals. The horizontal and vertical boundaries which are apparent in the image are produced not by abrupt intensity changes, but by certain grouping processes. Figure 11b is an example of so called "cognitive contours". They do not exist in the image, and cannot be detected by simple intensity-based operations. These example serves to illustrate that even a seemingly simple and elementary task such as the detection of image contours, requires in fact complex processing that is still far from being completely understood.

The second and more general comment has to do with the integration of theories of function and structure in more complex systems. The examples I have outlined come from a system that is relatively simple and easy to explore. Its anatomical structure is orderly, the input to the system is relatively easy to control and manipulate experimentally, and much is known about its physiology. Even under these favorable conditions, the integration of structure and function proves to be exceedingly difficult. What is the hope, then, for achieving comprehensive theories of structure and function for higher, more complicated, cognitive systems?

The task is certainly formidable, but it is probably worthy of exploration at least in certain instances, since it appears unlikely that the structure of complex systems can be understood without some guidelines supplied by computational theories. It has to be admitted, however, that given the difficulties of the task it is unclear whether coherent and detailed theories combining structure and function can be achieved at present beyond the simplest cognitive systems.

Acknowledgment: I wish to thank E. Hildreth and K. Stevens for their invaluable help.

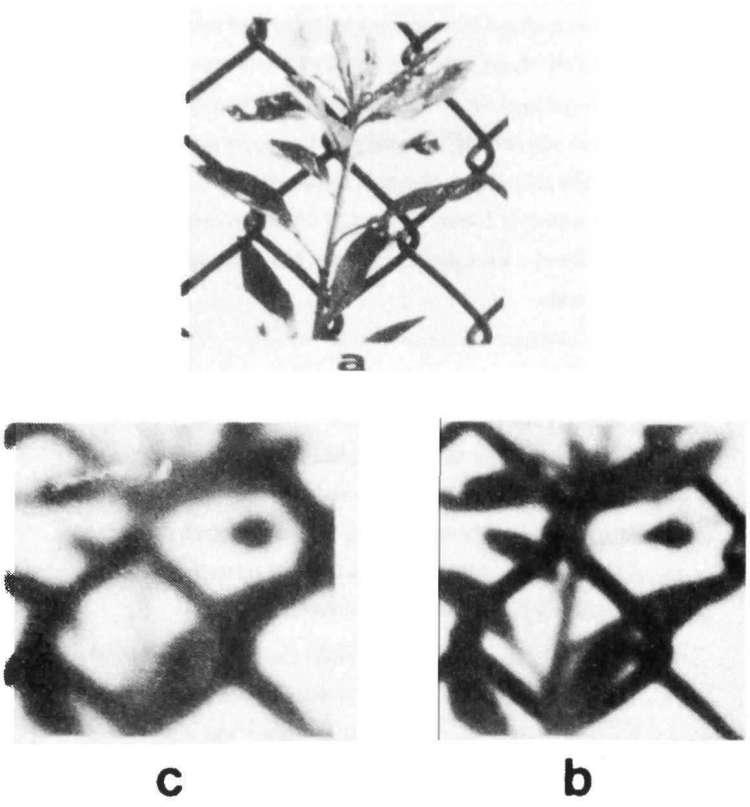


Fig. 1

The same image at three different resolutions.

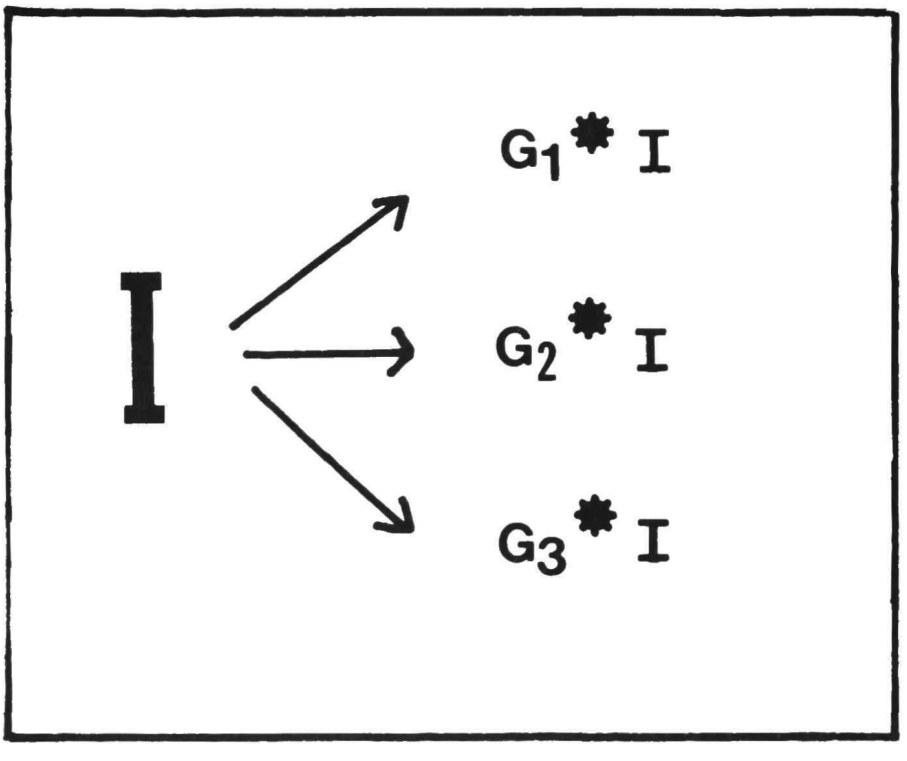


Figure 2. Different resolution copies of the original image are obtained by convolving the image with gaussian filters of different sizes.

Campbell, F.W. & Robson, J.G. Application of Fourier analysis to the visibility of gratings. *J. Physiol. (London) 197,* 551–556.

Dreher, B. & Sanderson, K.J. 1973 Receptive field analysis: responses to moving visual contours by single lateral geniculate neurons in the cat. *J. Physiol., Lond. 234,* 95–118.

Enroth-Cugell, C. & Robson, J. D. 1966 The contrast sensitivity of retinal ganglion cells of the cat. *J. Physiol. (Lond.) 187,* 517-522.

Grimson, W.E.L. 1981 A computer implementation of a theory of human stereo vision. *Phil. Trans. Roy. Soc., B. 292 (1058),* 217–253.

Hubel, D.H. & Wiesel, T.N. 1962 Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *J. Physiol. London, 160,* 106–154.

Hubel, D.H. & Wiesel, T.N. 1968 Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. London, 195,* 215–243.

Logan, B.F. 1977 Information in the zero–crossings of bandpass signals. *Bell Sys. Tech. J., 56,* 487–510.

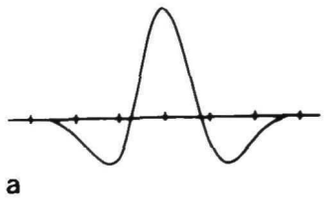
Marr, D. & Poggio, T. 1979 A computational theory of human stereo vision. *Proc. Roy. Soc. Lond. B 204,* 301–328.

Marr, D. Poggio, T. & Ullman, S. 1979 Bandpass channels, zero–crossings, and early visual information processing. *J. Opt. Soc. Am., 69(6),* 914–916.

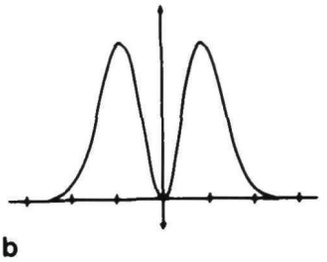
Marr, D. & Hildreth, E. 1980 Theory of edge detection. *Proc. R. Soc. Lond. B,* 187–217.

Marr, D. & Ullman, S. 1981 Directional selectivity and its use in early visual processing. *Proc. Roy. Soc. Lond. B, 211* 151–180.

Rodieck, R. W. & Stone, J. 1965 Analysis of receptive fields of cat retinal ganglion cells. *J. Neurophysiol. 28,* 833-849.



a



b

Figure 3. *a*. The shape of $\frac{d^2}{dx^2}G$. *b*. Its Fourier transform.

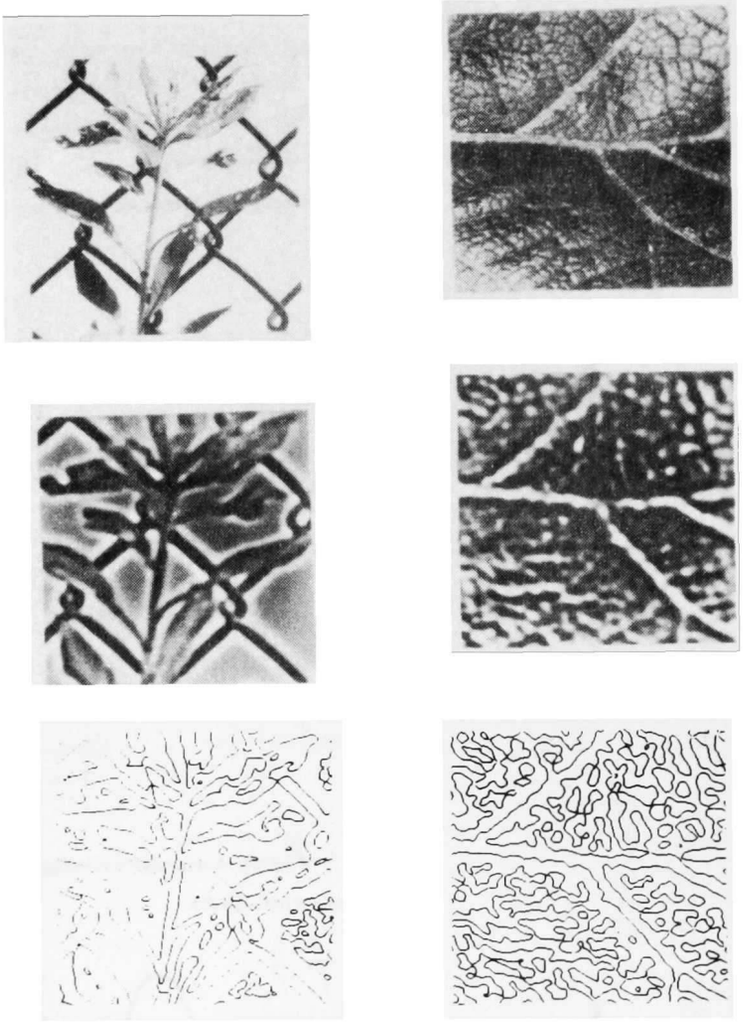


Figure 4. Examples of zero-crossing representations. First row: the original images. Second row: the images following the convolution with $\nabla^2 G$. Third rwo: the resulting zero-crossings representations.

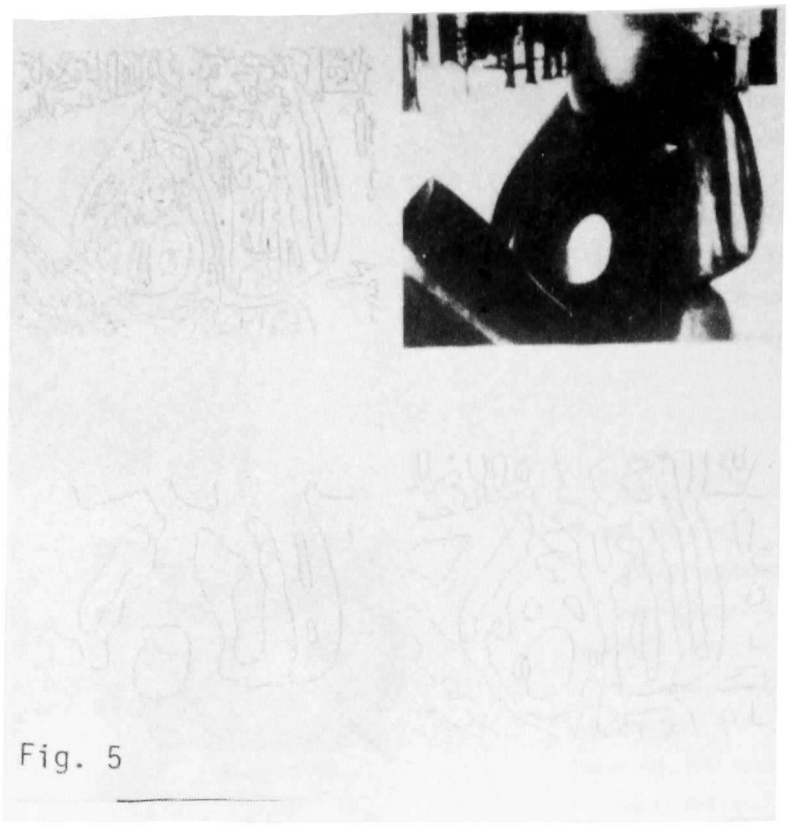


Fig. 5

Zero-crossing representations of the same image at three different resolutions.

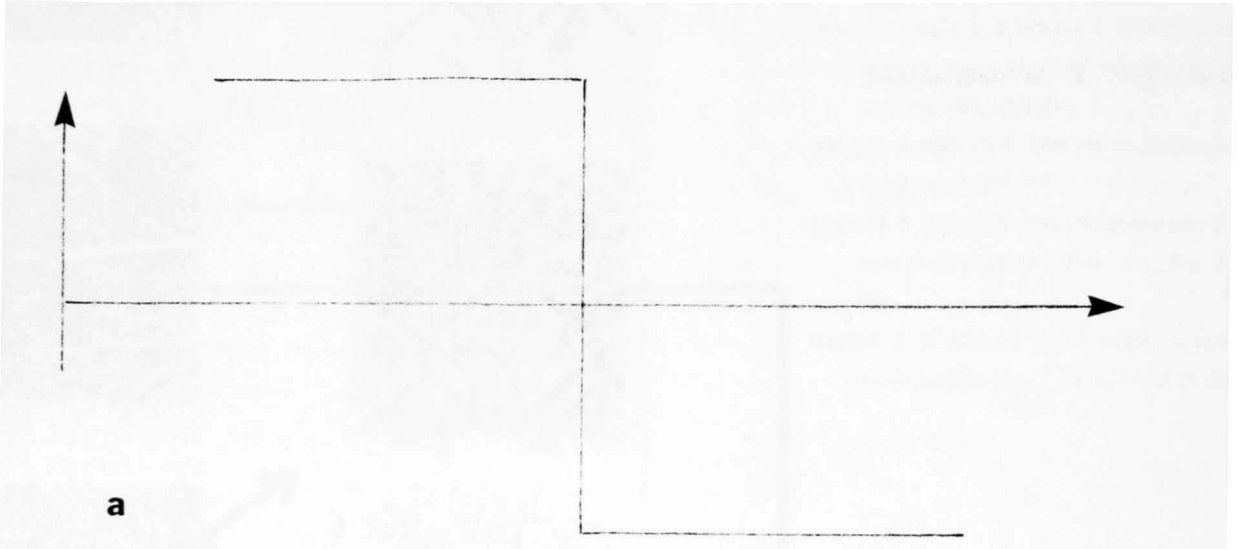
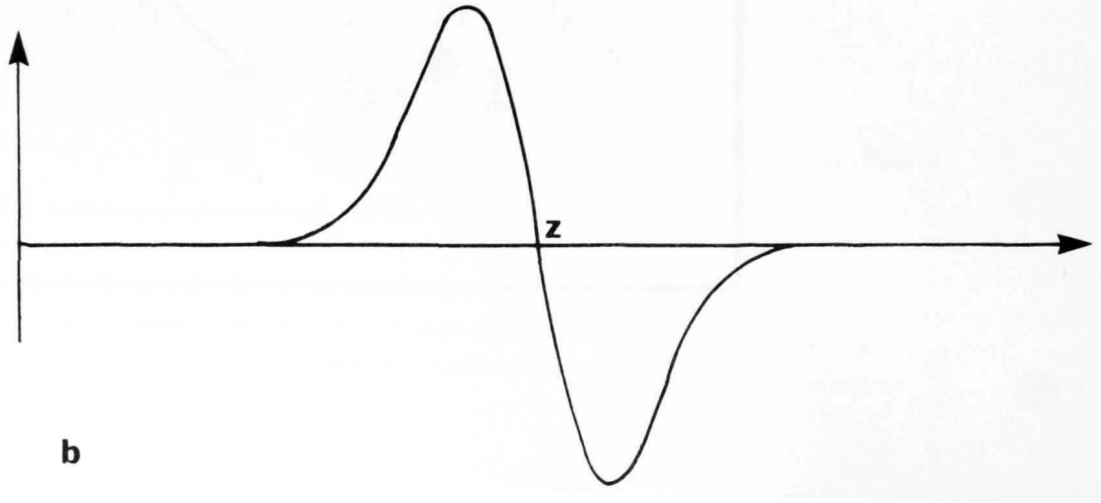


Figure 6. A step edge before (a) and after the convolutiion with the retinal operator (b).



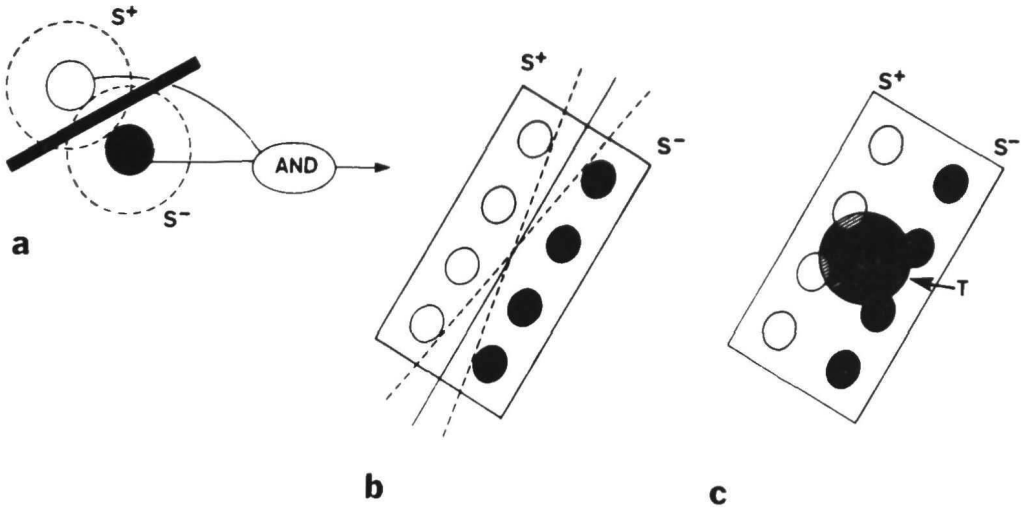


Figure 7 A schematic diagram of the basic zero-crossing detector. *a*. On-center and off-center units are *AND*ed together. *b*. A row of such subunits makes the detector orientation-specific. *c*. With the addition of a time derivative subunit the detector becomes directionally selective.

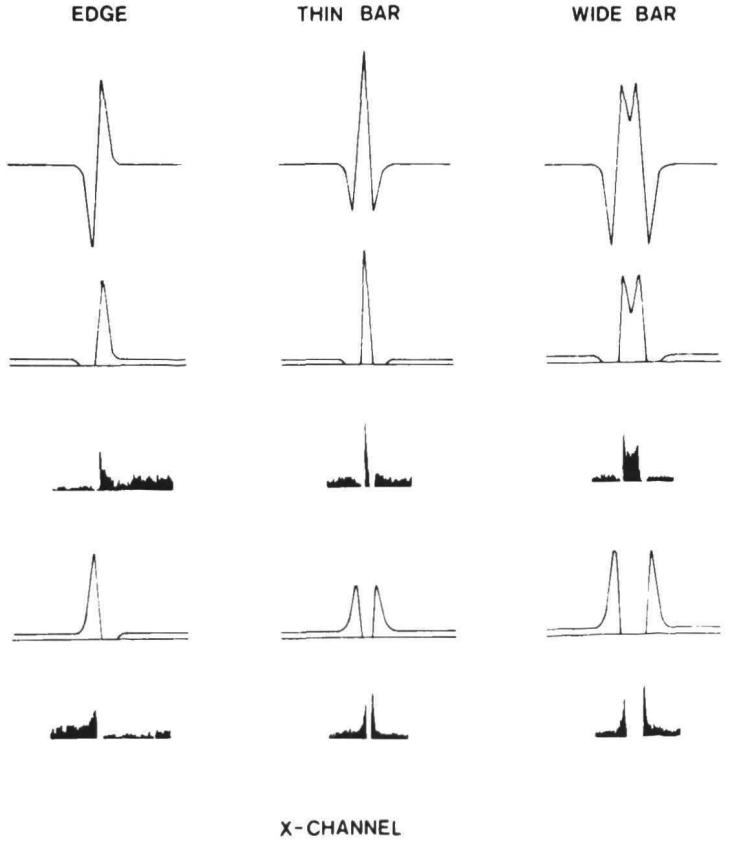


Figure 8. A comparison between the computational model and physiological recordings, for on- and off-center X-type units. For details see text

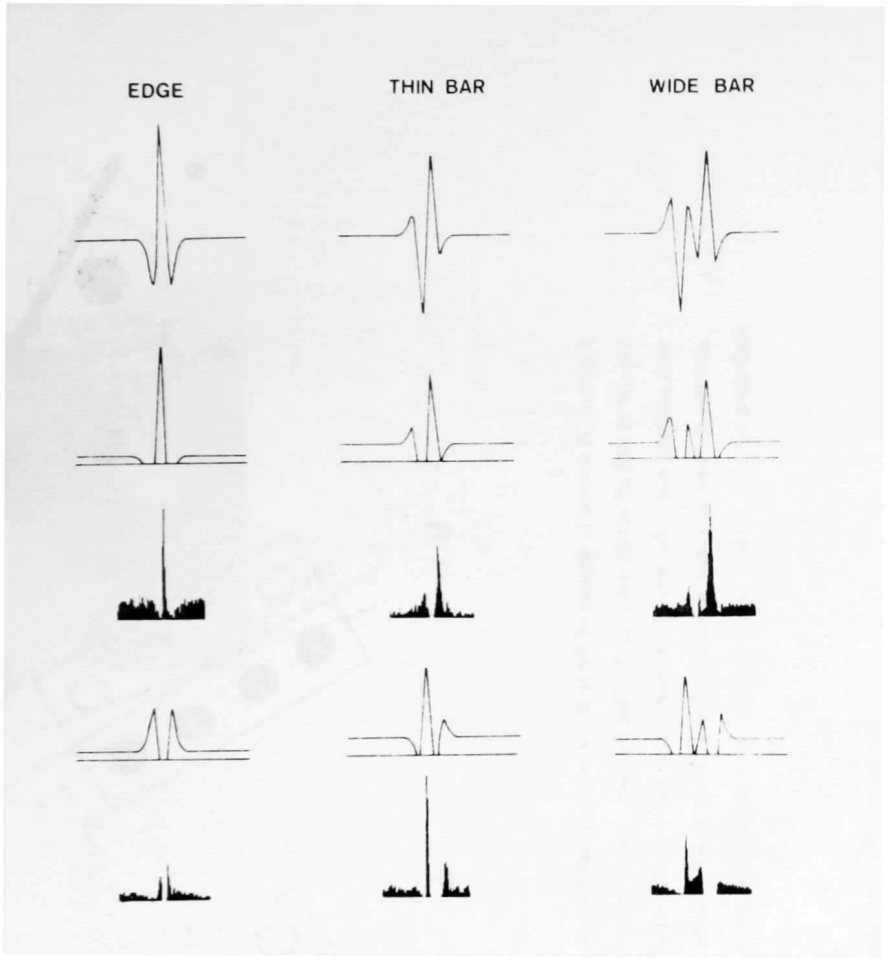


Figure 9. A comparison between the computational model and physiological recordings, for on- and off-center Y-type units. For details see text.

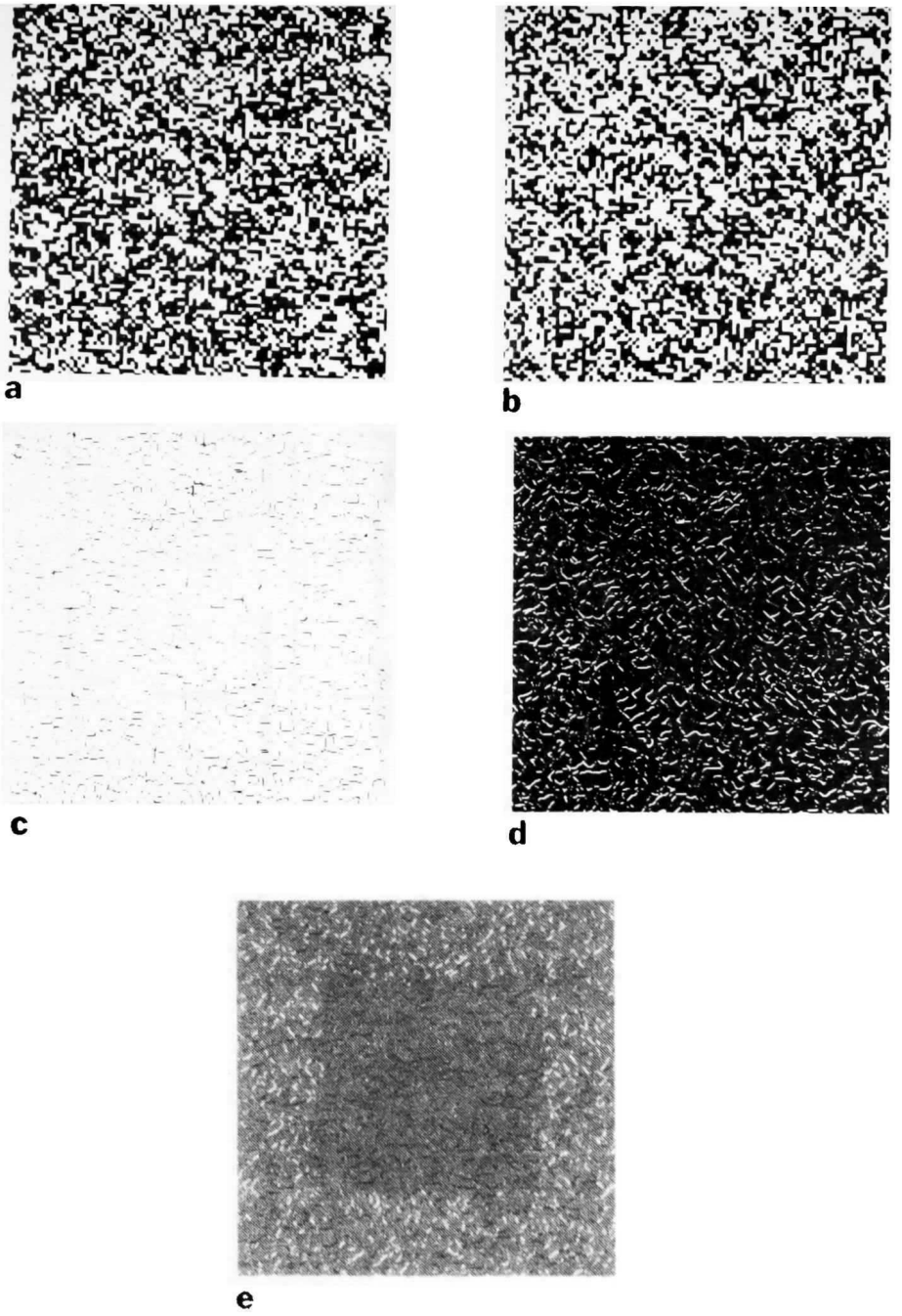
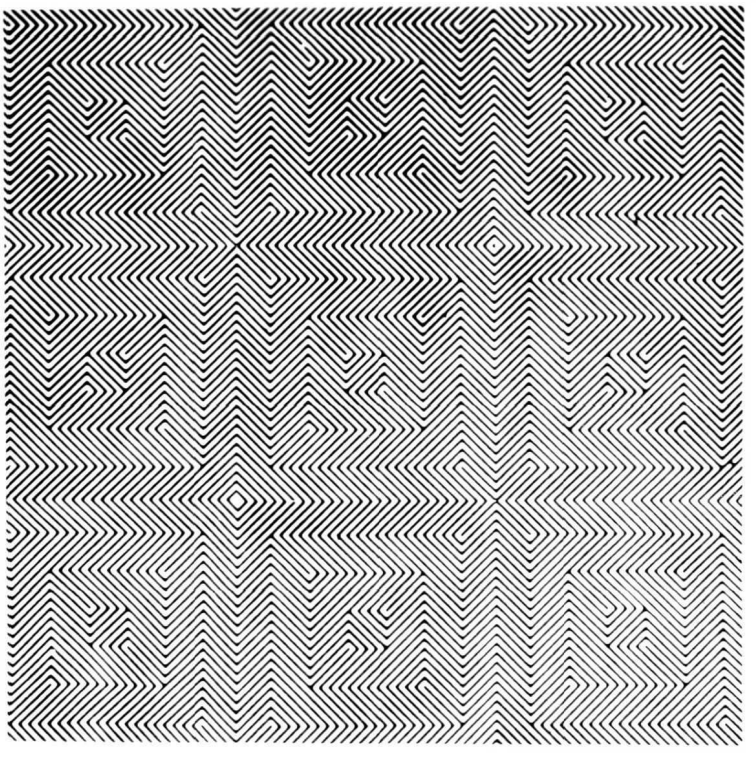


Figure 10. Detecting the motion of a moving random texture. See text for details.

(a)



(b)

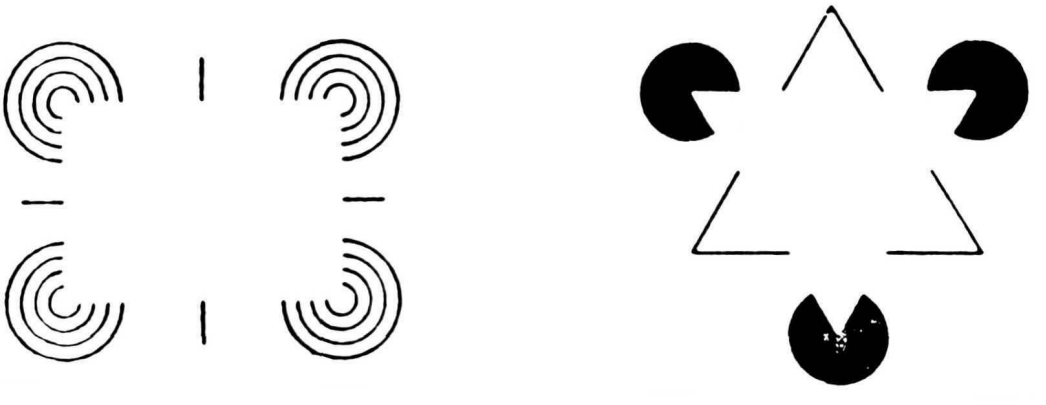


Figure 11. Contours that cannot be detected by simple intensity based operators.