A COMPUTER IMPLEMENTATION OF A THEORY OF HUMAN STEREO VISION

By W. E. L. GRIMSON

M.I.T. Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, Massachusetts 02139, U.S.A.

(*Communicated by S. Brenner, F.R.S. – Received 22 July 1980*)

CONTENTS

Recently, Marr & Poggio (1979) presented a theory of human stereo vision. An implementation of that theory is presented, and consists of five steps. (i) The left and right images are each filtered with masks of four sizes that increase with eccentricity; the shape of these masks is given by $\nabla^2 G$, the Laplacian of a Gaussian function. (ii) Zero crossings in the filtered images are found along horizontal scan lines. (iii) For each mask size, matching takes place between zero crossings of the same sign and roughly the same orientation in the two images, for a range of disparities up to about the width of the mask's central region. Within this disparity range, it can be shown that false targets pose only a simple problem. (iv) The output of the wide masks can control vergence movements, thus causing small masks to come into correspondence. In this way, the matching process gradually moves from dealing with large disparities at a low resolution to dealing with small disparities at a high resolution. (v) When a correspondence is achieved, it is stored in a dynamic buffer, called the $2\frac{1}{2}$-dimensional sketch.

To support the adequacy of the Marr–Poggio model of human stereo vision, the implementation was tested on a wide range of stereograms from the human stereopsis literature. The performance of the implementation is illustrated and compared

with human perception. Also statistical assumptions made by Marr & Poggio are supported by comparison with statistics found in practice. Finally, the process of implementing the theory has led to the clarification and refinement of a number of details within the theory; these are discussed in detail.

1. Introduction

If two objects are separated in depth from a viewer, then the relative positions of their images will differ in the two eyes. The process of stereo vision, in essence, measures this difference in relative positions, called the *disparity*, and uses it to compute depth information for surfaces in the scene.

The steps involved in measuring disparity are (Marr & Poggio 1979): (i) a particular location on a surface in the scene is selected from one image; (ii) that same location is identified in the other image; and (iii) the disparity between the two corresponding image points is measured. The difficulty of the problem lies in steps (i) and (ii), that is, in matching the images of the same location, the so-called correspondence problem. For the human stereo system, it can be shown that this matching takes place very early in the analysis of an image, before any recognition of what is being viewed, by means of primitive descriptors of the scene. This is illustrated by the example of random dot patterns. Julesz (1960) demonstrated that two images, consisting of random dots when viewed monocularly, may be fused to form patterns separated in depth when viewed stereoscopically. Random dot stereograms are particularly interesting because when one tries to set up a correspondence between two arrays of dots, false targets occur in profusion. A *false target* refers to a possible but incorrect match between elements of the two views. In spite of such false targets, and in the absence of any monocular or high level cues, we are able to determine the correct correspondence. Thus, the computational problem of human stereopsis reduces to that of obtaining primitive descriptions of locations to be matched from the images, and of solving the correspondence problem for these descriptions.

A computational theory of the stereo process for the human visual system was recently proposed by Marr & Poggio (1979). According to this theory, the human visual processor solves the stereoscopic matching problem by means of an algorithm that consists of five main steps. (i) The left and right images are each filtered at different orientations with bar masks of four sizes that increase with eccentricity; these masks have a cross section that is approximately the difference of two Gaussian functions, with space constants in the ratio 1:1.75. Such masks essentially perform the operation of a second directional derivative after low pass filtering or smoothing, and can be used to detect changes in intensity at different scales. (ii) Zero crossings in the filtered images are found by scanning them along lines lying perpendicular to the orientation of the mask. Since convolving the image with the masks corresponds to performing a second directional derivative, the zero crossings of the convolutions correspond to extrema in the first directional derivative of the image and thus to sharp changes in the original intensity function. (iii) For each mask size, matching takes place between zero-crossing segments of the same sign and roughly the same orientation in the two images, for a range of disparities up to about the width of the mask's central region. Within this disparity range, Marr & Poggio showed that false targets pose only a simple problem, because of the roughly bandpass nature of the filters. (iv) The output of the wide masks can control vergence movements, thus causing smaller masks to come into correspondence. In this way,

the matching process gradually moves from dealing with large disparities at low resolution to dealing with small disparities at high resolution. (v) When a correspondence is achieved, it is stored in a dynamic buffer, called the $2\frac{1}{2}$-dimensional sketch (Marr & Nishihara 1978).

An important aspect in the development of any computational theory is the design and implementation of an explicit algorithm for that theory. There are several benefits from such an implementation. One concerns the act of implementation itself, which forces one to make all details of the theory explicit. This often uncovers previously overlooked difficulties, thereby guiding further refinement of the theory.

A second benefit concerns the performance of the implementation. Any proposed model of a system must be testable. In this case, by testing on pairs of stereo images, one can examine the performance of the implementation, and hence of the theory itself, provided that the implementation is an accurate representation of that theory. In this manner, the performance of the implementation can be compared with human performance. If the algorithm differs strongly from known human performance, its suitability as a biological model is quickly brought into question (cf. the analysis of the cooperative algorithm of Marr & Poggio (1976) in Marr & Poggio (1979)).

This article describes an implementation of the Marr–Poggio stereo theory, written with particular emphasis on the matching process (Grimson & Marr 1979). For details of the derivation and justification of the theory, see Marr & Poggio (1979).

The first part of this paper describes the overall design of the implementation. Several examples of the implementation's performance on different images are then discussed, including random dot stereograms from the human stereopsis literature, such as with one image defocused, noise introduced into part of the images' spectra, and so forth. It is shown that the implementation behaves in a manner similar to humans on these special cases. The results of running the program on some natural images are also shown. Then, the theory makes some statistical assumptions; these are compared with the actual statistics found in practice. Finally, some points about the theory that were clarified as a result of writing the program are discussed.

2. Design of the program

The implementation is divided into five modules, roughly corresponding to the five steps in the summary above. These modules, and the flow of information between them, are illustrated in figure 1. Each of the components is described in turn.

2.1. *Input*

There are two aspects of the human stereo system, embedded in the Marr–Poggio theory, that must be made explicit in the input to the algorithm. The first is the position of the eyes with respect to the scene, as eye movements will be critical for obtaining fine disparity information. The second is the change in resolution of analysis of the image with increasing eccentricity.

To account for these effects, the algorithm maintains as its initial input a stereo pair of arrays, representing the entire scene visible to the viewer. This pair of arrays corresponds to the environment around the visual system, rather than some integral part of the system itself. To create this representation of the scene, photographs of natural images were digitized on an Optronix Photoscan System P1000. The sizes of these images are indicated in the legends.

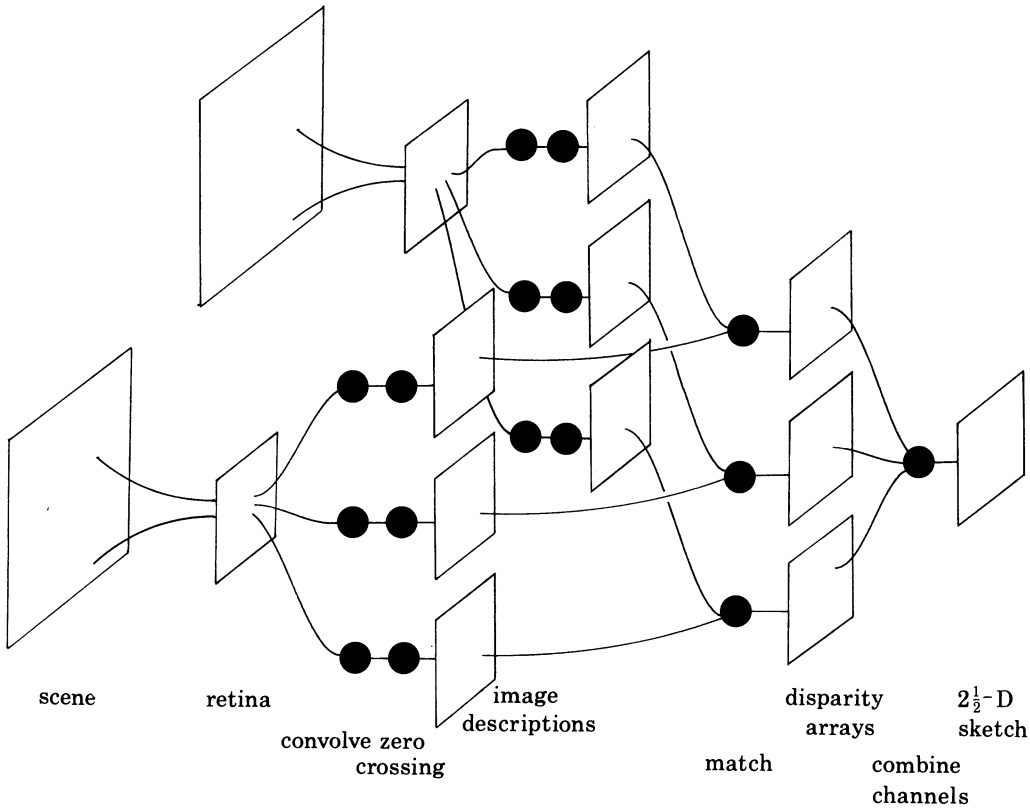


FIGURE 1. Diagram of the implementation. From the initial arrays of the scene, a retinal pair of images are extracted. These subimages reflect the current eye positions relative to the scene, and may be considered as the representation of the light intensities striking the retina. Each retinal image is convolved with a set of different-sized filters of the form $\nabla^2 G$, and the zero crossings of the output are located. For each filter size, the zero crossings descriptions are matched, based on the sign and orientation of each zero crossing point. The disparity arrays created for each filter are combined into a single disparity description, and information from the larger filters can be used to verge the eyes, thereby bringing the smaller filters into a range of operation.

Grey-level resolution is eight bits, providing 256 intensity levels. For the random dot patterns illustrated in this article, the images were constructed by computer, rather than digitized from photographs.

For a given position of the eyes relative to the scene, a representation of the images on the two retinas is extracted. The program creates this retinal representation by obtaining a second, smaller pair of images from the arrays representing the whole scene. The mapping from the scene arrays into the retinal images accounts for one of the factors inherent in the way that the human visual system is constructed. Different sections of the scenes will be mapped to the centre (fovea) of the retinal images as the positions of the eyes are varied. In this way, different sections of the scenes can separately be matched to a very fine level of disparity, by allowing the smallest channels to come into range of correspondence. Since the matching process will take place on the arrays representing the retinal images, it is important that the coordinate systems of those arrays coincide with the current positions of the eyes. Note that the portion of the scene image that is mapped into the retinal image may differ for the two eyes, depending on the relative positions of the two optical axes. In particular, there may be differences in vertical alignment as well as in horizontal alignment. There is a second factor that should

also be taken into account. Wilson & Bergen (1979) and Wilson & Giese (1977), state that the resolution of the earlier stages of the algorithm, the convolution and zero crossings, scales linearly with eccentricity. However, this aspect has not been implemented and in our situation is not critical, since the images analysed correspond to small visual angles, of the order of 4° on a side.

After the completion of this stage, the program has created a representation of the images that has accounted for eye position and if appropriate also for retinal scaling with eccentricity. For each pass of the algorithm, the matching will take place on the representation of the retinal images, thereby implicitly assuming some particular eye positions. Once the matching has been completed, the disparity values obtained may be used to change the positions of the two optic axes, thus causing a new pair of retinal images to be extracted from the representations of the scene, and the matching process may proceed again.

2.2. *Convolution*

Given the retinal representations of the images, it is necessary to transform them into a representation that the matcher may operate. The evidence of random dot patterns (Julesz 1960, 1970) suggests that the extraction of the descriptions to be matched must be, at least in part, an early visual process. Further, Marr & Poggio (1979) argue that only those points in an image that are in one-to-one correspondence with well defined locations on a physical surface can be matched by the stereo process. Thus, it is necessary to detect these points by an early visual process. A theory of this process has been developed by Marr & Hildreth (1980) and the relevant points are outlined below.

The intensity changes that correspond to well defined physical locations take place over a wide range of scales (Marr 1976). As a consequence, it is not possible to find a single filter that will be simultaneously optimal at all scales. The findings of Campbell & Robson (1968), concerning the existence of separate spatial-frequency channels in the human visual system, suggest that one should seek a method of dealing separately with the changes occurring at different scales. This is in agreement with the findings of Julesz & Miller (1975) and Mayhew & Frisby (1976), who showed that spatial-frequency-tuned channels are used in stereopsis and are independent. Marr & Hildreth argue that one first takes some local average of intensity at several resolutions, and then detects the changes in intensity that occur at each resolution. They show that the optimal smoothing filter is the Gaussian distribution. If an intensity change occurs along a particular orientation in the image, there will be a peak in the first directional derivative of intensity, and a zero crossing in the second directional derivative. Thus, the intensity changes in the image can be located by finding zero crossings in the output of a second directional derivative operator.

In the original theory (Marr & Poggio 1979), the proposed masks were oriented bar masks whose cross section was a difference of two Gaussians (the form was given by the data of Wilson & Bergen (1979)). However, a number of practical considerations have led Marr & Hildreth (1979) to suggest that the initial operators are not directional operators. The only non-directional linear second derivative operator is the Laplacian. Marr & Hildreth have shown that, provided that two simple conditions on the intensity function in the neighbourhood of an edge are satisfied, the zero crossings of the second directional derivative taken perpendicular to an edge will coincide with the zero crossings of the Laplacian along that edge.

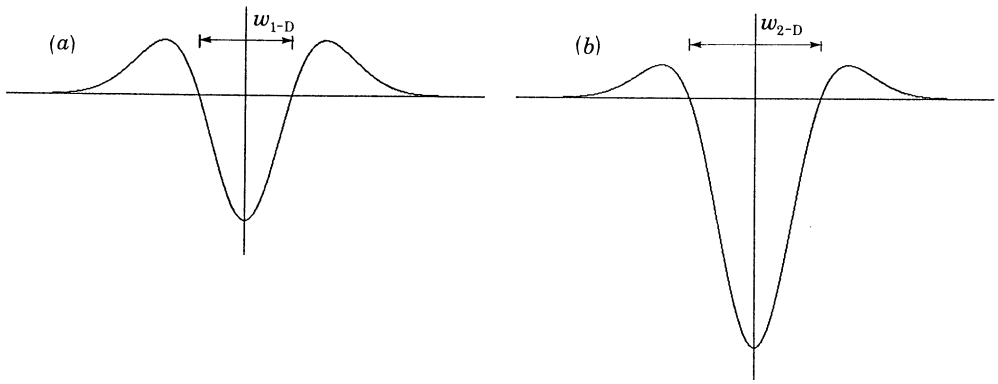


FIGURE 2. The primal sketch operator. (*a*) The one-dimensional operator. This operator applies for images whose intensity values are constant along each vertical slice of the image. (*b*) A cross section of the equivalent, rotationally symmetric, two-dimensional operator. The sizes of the operators are determined by the values of $w_{1\text{-}D}$ and $w_{2\text{-}D}$ respectively.

Therefore, theoretically, we can detect intensity changes occurring at all orientations by means of the single non-oriented Laplacian operator. Thus, Marr & Hildreth propose that intensity changes occurring at a particular scale may be detected by locating the zero crossings in the output of $\nabla^2 G$, the Laplacian of a Gaussian distribution.

It is interesting to note that, although the use of these operators was motivated by computational and practical arguments, the use of non-oriented filters was arrived at independently on psychophysical grounds by Mayhew & Frisby (1978).

The precise form of the operator is given by

$$\nabla^2 G(r, \theta) = \left(\frac{r^2 - 2\sigma^2}{\sigma^4}\right) e^{-r^2/2\sigma^2}.$$

This is a rotationally symmetric function, and its cross section is shown in figure 2. Note that its central panel width, denoted by w and defined as the width of the central negative region, is given by

$$w_{2\text{-}D} = 2\sqrt{2}\sigma.$$

If the visual input to the operator is a one-dimensional grating, then the response of the operator is equivalent to that obtained by applying the equivalent one-dimensional operator to the one-dimensional input. This equivalent one-dimensional operator is obtained by projecting $\nabla^2 G$ onto a line, and is given by

$$D_{xx}G = (2\pi)^{\frac{1}{2}} \left(\frac{x^2 - \sigma^2}{\sigma^3}\right) e^{-x^2/2\sigma^2}.$$

The central panel width of this operator is

$$w_{1\text{-}D} = 2\sigma.$$

The operator is illustrated in figure 2.

Given the form of the operators, only the size of these filters is left to be determined. At this point, it is interesting to note the evidence of Wilson & Giese (1977) and Wilson & Bergen (1979), concerning the existence and characteristics of such operators in the human visual system. They have found strong evidence that at each point in the visual field there exist at least four and possibly five independent channels in the human early visual system. The form

of the channels, as analysed by Wilson and collaborators, very closely fits the shape of a difference of two Gaussians. Further, a difference of Gaussians is a close approximation to a Laplacian applied to a Gaussian (appendix B, of Marr & Hildreth 1980). The data of Wilson & Bergen indicated filters whose sizes, specified by the width w of the filter's central region, range from 3.1′ to 21′ of visual arc. The variable w is related to the constant σ of $\nabla^2 G$ by the relation:

$$\sigma = w/(2\sqrt{2}).$$

Wilson & Bergen obtained their values by using oriented line stimuli. To obtain the diameter of the corresponding circularly symmetric centre–surround receptive field, the values of w must be multiplied by $\sqrt{2}$. Finally, we want the resolution of the initial images to roughly represent the resolution of processing by the cones, and the size of the filters to represent the size of the retinal operators. In the most densely packed region of the human fovea, the centre-to-centre spacing of the cones is 2.0 to 2.3 μm, corresponding to an angular spacing of 25″ to 29″ (O'Brien 1951). Accounting for the conversion of the data of Wilson & Bergen, and using the figure of 27″ for the separation of cones in the fovea, one arrives at values of w in the range nine to 63 image elements, and hence, values of σ in the range three to 23 image elements.

Although Wilson and collaborators have found definite evidence only for four different-sized channels, it has recently been proposed (Marr *et al.* 1980) that a further, smaller channel may be present. In the human system, this channel would consist of a single retinal receptor, although, because of the diffraction in the eye, the actual size of the equivalent operator would be larger. In the implementation, since diffraction is not a factor, this channel would have a central width of $w = 1.5'$, roughly corresponding to four image elements.

The present implementation uses four filters, each of which is a Laplacian of a Gaussian, with w values of four, nine, 17 and 35 image elements. The coefficients of the filters were represented to a precision of one part in 2048. Coefficients of less than $\frac{1}{2048}$th of the maximum value of the mask were set to zero. Thus, the truncation radius of the mask (the point at which all further mask values were treated as zero) was approximately $1.8w$, or, equivalently, 5.08σ.

The actual convolutions were performed on a LISP machine constructed at the M.I.T. Artificial Intelligence Laboratory, with use of additional hardware specially designed for the purpose (Knight *et al.* 1979). Figures 3 and 4 illustrate some images and their convolutions with various sized masks.

After the completion of this stage of the algorithm, one has four filtered copies of each of the images, each copy having been convolved with a different-sized mask.

2.3. *Detection and description of zero crossings*

The elements that are matched between images are (i) zero crossings whose orientations are not horizontal and (ii) terminations. The exact definition and hence the detection of terminations is at present uncertain. Moreover, terminations are much rarer than zero crossings. As a consequence, only zero crossings are used as input to the matcher.

Since, for the purpose of obtaining disparity information, horizontally oriented segments may be ignored, the detection of zero crossings can be accomplished by scanning the convolved image horizontally for adjacent elements of opposite sign, or for three horizontally adjacent elements, the middle one of which is zero, the other two containing convolution values of

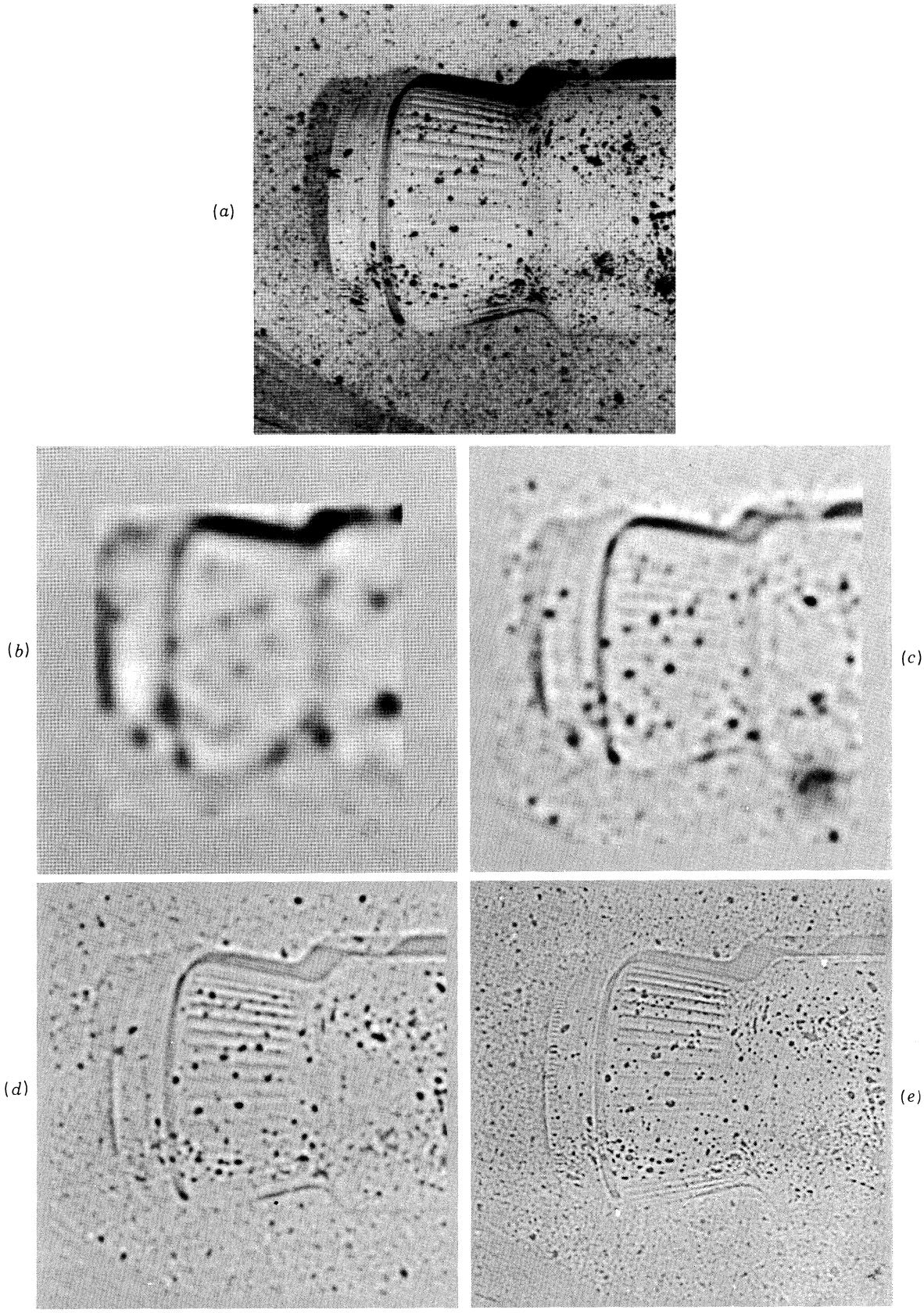


FIGURE 3. Examples of convolutions with $\nabla^2 G$. A natural image is indicated in (*a*). Below are examples of the convolved image, after application of different sized $\nabla^2 G$ operators, with central panel widths of (*b*) 36, (*c*) 18, (*d*) nine and (*e*) four picture elements. The original image was 480 picture elements on a side.

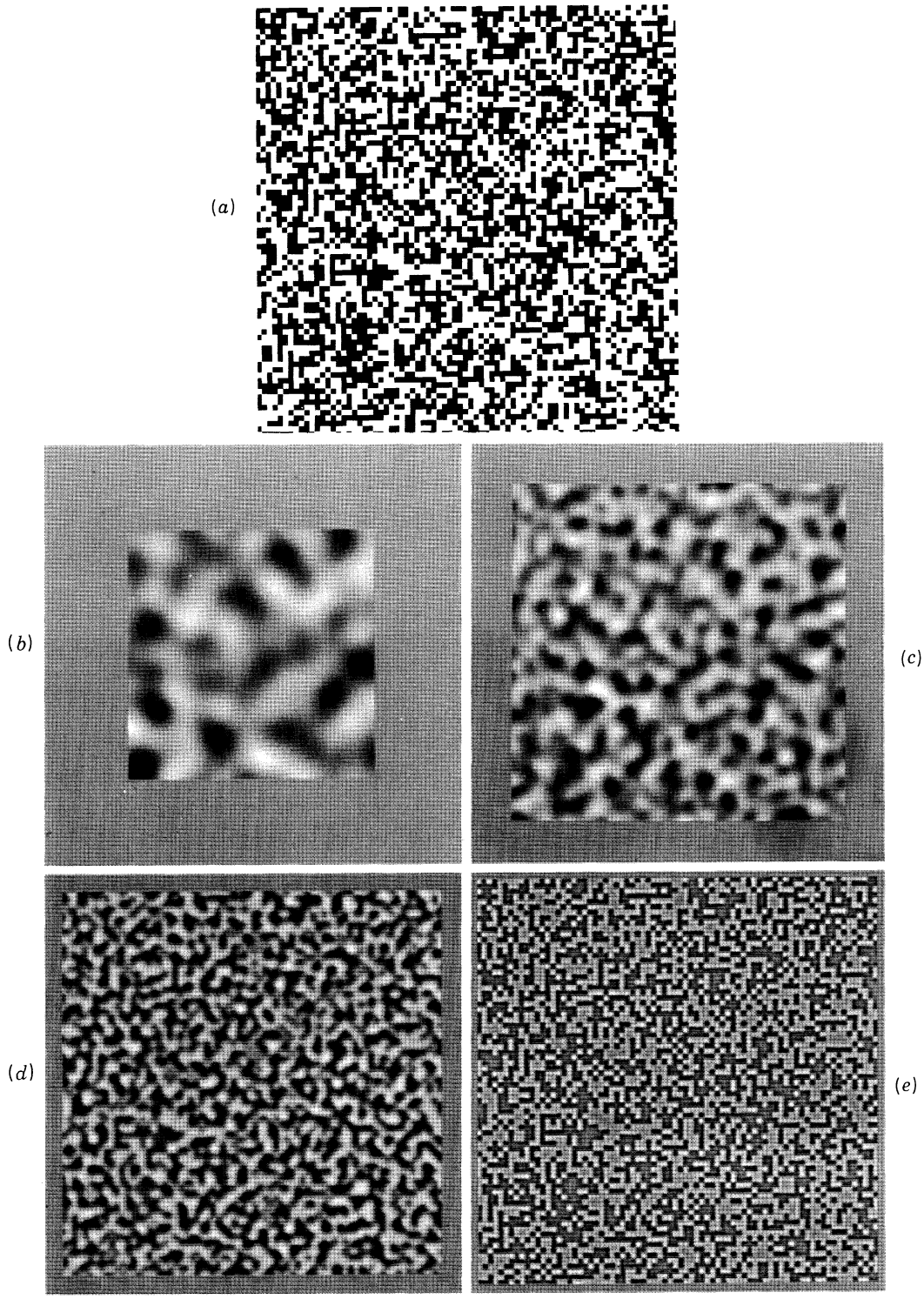


FIGURE 4. Examples of convolutions with $\nabla^2 G$. A random dot pattern is indicated in (*a*). Below are examples of the convolved image, after application of different sized $\nabla^2 G$ operators, with central panel widths of (*b*) 36, (*c*) 18, (*d*) nine and (*e*) four picture elements. The original image was 320 picture elements on a side.

opposite sign. This gives the position of zero crossings to within an image element. Note that there is no theoretical limit on the accuracy with which the zero crossings may be localized For the purposes of matching, however, a resolution of one pixel suffices.

In addition to their location, the sign of the zero crossings (whether convolution values change from positive to negative or negative to positive as scanned from left to right) and a rough estimate of the local, two-dimensional orientation of pieces of the zero-crossing contour are recorded. In the present implementation, the orientation at a point on a zero-crossing segment is computed as the direction of the gradient of the convolution values across that segment, and recorded in increments of 30°. Figures 5 and 6 illustrate zero crossings obtained in this way from the convolutions of figures 3 and 4. Positive zero crossings are shown white, and negative crossings in black. This zero-crossing description is computed for each image and for each size of mask.

2.4. *Matching*

The matcher implements the second of the matching algorithms described by Marr & Poggio (1979, p. 315). For each size of filter, matching consists of six steps:

(i) fix the eye positions;

(ii) locate a zero crossing in one image;

(iii) divide the region about the corresponding point in the second image into three pools;

(iv) assign a match to the zero crossing based on the potential matches within the pools;

(v) disambiguate any ambiguous matches;

(vi) assign the disparity values to a buffer.

These steps may be repeated several times during the fusion of an image. Given positions for the eyes, these matching steps are performed, with the results stored in a buffer. These results may be used to refine the eye positions, causing a new set of retinal images to be extracted from the scene, and the matching steps are performed again.

The first step consists of fixing the two eye positions. The alignment between the two zero-crossing descriptions, corresponding to the positions of the eyes, is determined in two ways. The initial offsets of the descriptions are arbitrarily set to zero. Thereafter, the offsets of the two optical axes are determined by accessing the current disparity values for a region and using these values to adjust the vergence of the eyes. In my implementation, this is done by modifying the extraction of the retinal images from the images of the entire scene, accounting for the positions of the eyes.

Once the eye positions have been fixed, and the retinal images extracted, the zero-crossing descriptions are obtained as in figures 5 and 6. For a zero-crossing description obtained from a particular filter size, the matching is performed by locating a zero crossing and performing the following operation. Given the location of a zero crossing in one image, a region about the same location in the other image is partitioned into three pools. These pools form the region to be searched for a possible matching zero crossing and consist of two larger convergent and divergent regions, and a smaller one lying centrally between them. Together these pools span a disparity range equal to $2w_{1\text{-}D}$, where $w_{1\text{-}D}$ is the width of the central excitatory region of the corresponding one-dimensional convolution filter.

The following criteria are used for matching zero crossings in the left and right filtered images, for each pool.

(i) The zero crossings must come from convolutions with the same size filter.

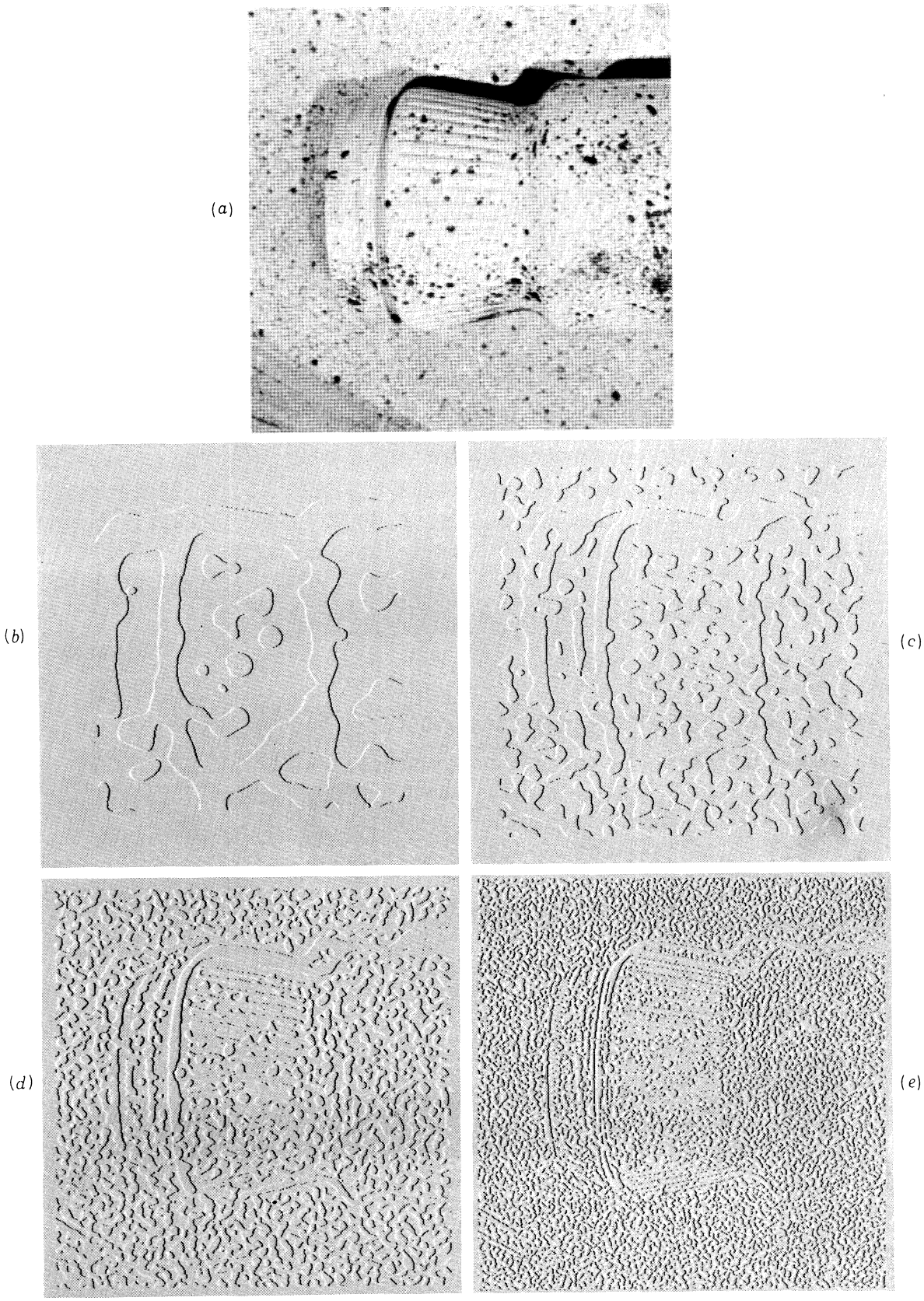


FIGURE 5. Examples of zero crossings. A natural image is indicated in (*a*). Below are examples of the zero crossings, obtained from different-sized $\nabla^2 G$ operators, with central panel widths of (*a*) 36, (*b*) 18, (*c*) nine and (*d*) four picture elements. The positive zero crossings are shown as white, the negative ones as black.

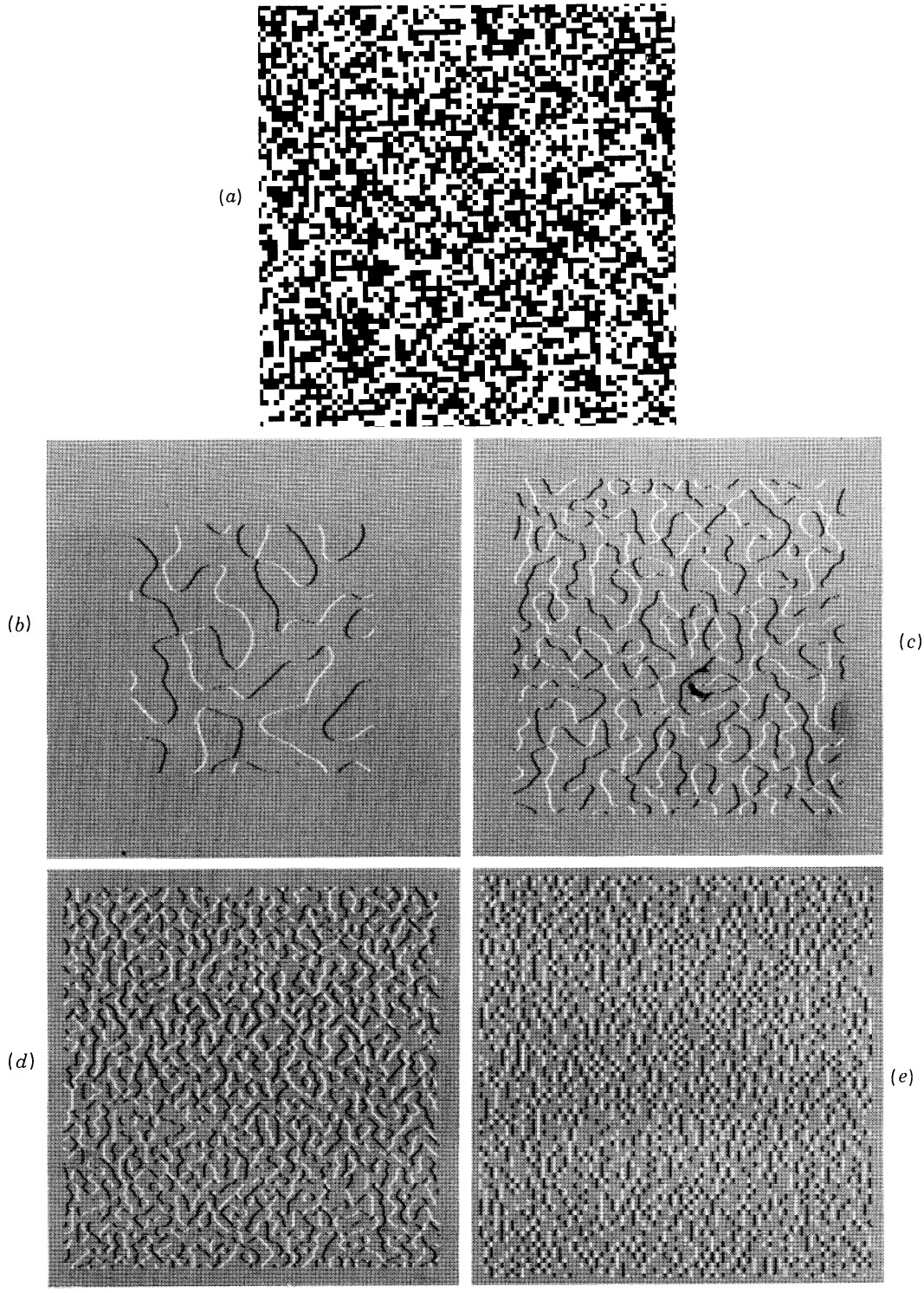


FIGURE 6. Examples of zero crossings. A random dot pattern is indicated in (*a*). Below are examples of the zero crossings, obtained from different-sized $\nabla^2 G$ operators, with central panel widths of (*b*) 36, (*c*) 18, (*d*) nine and (*e*) four picture elements. The positive zero crossings are shown as white, the negative ones as black.

(ii) The zero crossings must have the same sign.

(iii) The zero crossing segments must have roughly the same orientation.

A match is assigned on the basis of the responses of the pools. If exactly one zero crossing of the appropriate sign and orientation (within 30°) is found within a pool, the location of that crossing is transmitted to the matcher. If two candidate zero crossings are found within one pool (an unlikely event), the matcher is notified and no attempt is made to assign a match for the point in question. If the matcher finds a single crossing in only one of the three pools, that match is accepted, and the disparity associated with the match is recorded in a buffer. If two or three of the pools contain a candidate match, the algorithm records that information for future disambiguation.

Once all possible unambiguous matches have been identified, an attempt is made to disambiguate double or triple matches. This is done by scanning a neighbourhood about the point in question, and recording the sign of the disparity of the unambiguous matches within that neighbourhood. (The sign of the disparity refers to the sign of the pool from which the match comes: divergent, convergent or zero.) If the ambiguous point has a potential match of the same sign as the dominant type within the neighbourhood, then that is chosen as the match (this is the 'pulling' effect). Otherwise, the match at that point is left ambiguous.

There is the possibility that the region under consideration does not lie within the $\pm w$ disparity range handled by the matcher. This situation is detected and handled by the following operation. Consider the case in which the region does lie within the disparity range $\pm w$. Excluding the case of occluded points, every zero crossing in the region will have at least one candidate match (the correct one) in the other filtered image. On the other hand, if the region lies beyond the disparity range $\pm w$, then the probability of a given zero crossing having at least one candidate match will be less than one. In fact, Marr & Poggio show that the probability of a zero crossing having at least one candidate match in this case is roughly 0.7. Hence, the following operation can be performed. For a given eye position, the matching algorithm is run for all the zero crossings. Any crossing for which there is no match is marked as such. If the percentage of matched points in any region is less than a threshold of 0.7 then the region is declared to be out of range, and no disparity values are accepted for that region.

The size of the regions used for checking the statistics of matching zero crossings should be proportional to the density of the zero crossings, to ensure a fixed confidence level. Typically, the regions were roughly 25 picture elements on a side, for a channel with filter size $w = 9$.

The overall effect of the matching process, as driven from the left image, is to assign disparity values to most of the zero crossings obtained from the left image. An example of the output appears in figure 7. In this array, a zero crossing at position (x,y) with associated disparity d has been placed in a three-dimensional array with coordinate (x,y,d). For display purposes, the array is shown in the figures as viewed from a point some distance away. The heights in the figure correspond to the assigned disparities. (For graphical convenience, the disparities have been inverted.)

After completion of this stage of the implementation, a disparity array for each filter size has been obtained. The disparity values are located only along the zero-crossing contours obtained from that filter.

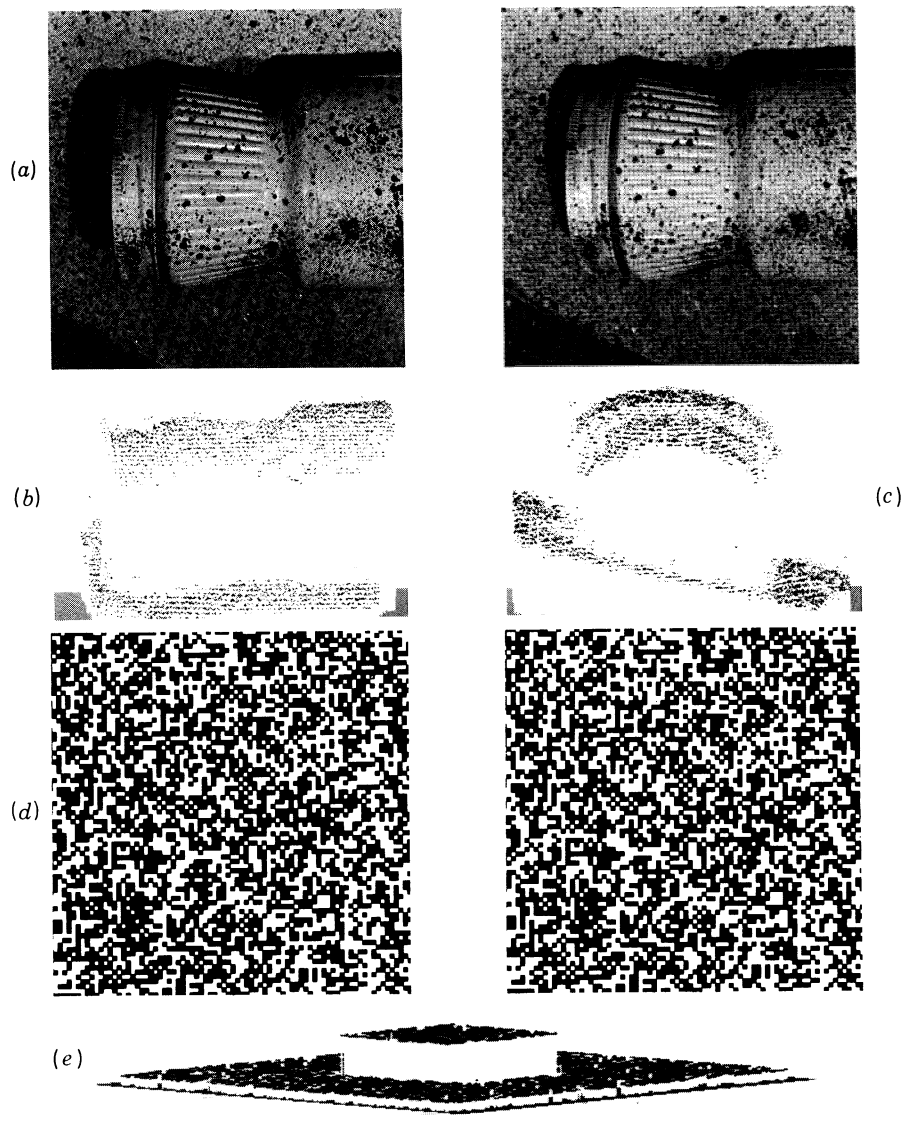


FIGURE 7. Results of the algorithm. The top stereo pair (*a*) is an image of a painted coffee jar. (*b*), (*c*). Two orthographic views of the disparity map. The disparities are displayed as $[x, y, c-ad(x,y)]$, where c is a constant and $d(x,y)$ is the difference in the location of a zero crossing in the right and left images. For purposes of illustration, a has been adjusted to enhance the features of the disparity map. The jar is viewed in (*b*) from the lower edge of the image and in (*c*) from the left edge of the image. Note that the background plane appears tilted in the disparity map. This agrees with the fused perception. The second stereo pair (*d*) is a 50% density random dot pattern. (*e*) The disparity map as viewed orthographically from some distance away. All disparity maps are those obtained from the $w = 4$ channel.

2.5 *Vergence control*

The Marr–Poggio theory states that to obtain fine resolution disparity information, it is necessary that the smallest channels obtain a matching. Since the range of disparity over which a channel can obtain a match is directly proportional to the size of the channel, this means that the eyes must move to ensure that the corresponding zero-crossing descriptions from the two images are within a matchable range. The disparity information required to bring the smallest channels into their matchable range is provided by the larger channels. That

is, if a region of the image is declared to be out of range of fusion by the smaller channels, one can frequently obtain a rough disparity value for that region from the larger channels, and use this to verge the eyes. In this way, the smaller channels can be brought into a range of correspondence.

Thus, after the disparities from the different channels have been combined, there is a mechanism for controlling vergence movements of the eyes. This operates by searching for regions of the image that have disparity values from the larger channels, but do not have disparity values for the smallest channel. The values from the large channel are used to provide a refinement to the current eye positions, thereby bringing the smaller channels into range of correspondence. Possible mechanisms for extracting the disparity value from a region of the image include using the peak value of a histogram of the disparities in that neighbourhood, using the average of the local disparity values, or using the median of the local disparity values. In the current implementation, the search for such a region proceeds outwards from the fovea.

It should be noted here that, although Marr & Poggio state that disparity information from coarser channels can drive eye movements, they do not rule out that other information can also do this. There may be other modules of the visual system that can initiate eye movements. Kidd *et al.* (1979) for example, found that certain types of texture boundaries can initiate eye movements. However, such effects are somewhat orthogonal to the question of the adequacy of the matching component of the Marr–Poggio theory, since they affect the input to the matcher but not the actual performance of the matching algorithm itself.

2.6. *The $2\frac{1}{2}$-dimensional sketch*

Once the separate channels have performed their matching, the results are combined and stored in a buffer, called the $2\frac{1}{2}$-D sketch. There are several possible methods for accomplishing this. As far as the Marr–Poggio theory is concerned, the important point is that some type of storage of disparity information occurs. Perhaps the strongest argument for this is the fact that up to 2° of disparity can be held fused in the fovea, although the matching range for a single fixation of the eyes is only 30′.

I have considered two different possibilities for the way in which information from the different channels is combined. The method used in the current implementation will be described below. A more biologically feasible method will be outlined in the discussion.

One of the critical questions concerning the form of the $2\frac{1}{2}$-D sketch is whether its coordinates are consistent with those of the scene or the retinal images. For all the cases illustrated in this article, the sketch was constructed by directly relating the coordinates of the sketch to the coordinates of the scene arrays. That is, as disparity information was obtained, it was stored in a buffer at the position corresponding to the position in the original scene from which the underlying zero crossing came. Since disparity information about the scene is extracted from several eye positions, explicit information about the positions of the eyes is required in order to store this information into a buffer. This is probably inappropriate as a model of the human system, but it suffices for demonstrating the effectiveness of the matching module.

The actual mechanism for storing the disparity values requires some combination of the disparity maps obtained for each of the channels. Currently, the sketch is updated, for each region of the image, by writing in the disparity values from the smallest channel that is within range of fusion. Vergence movements are possible to bring smaller channels into a range of

matching for some region. Further, for those regions of the image for which none of the channels can find matches, modification of the eye positions over a scale larger than that of the vergence movements is possible. By this method, one can attempt to bring those regions of the image into a range of fusion. There are several possibilities for the actual method of driving the vergence movements. Two of these were outlined in the previous section.

The final output of the algorithm consists of a representation of disparity values in the image, specified along zero-crossing segments from the smallest channel that was used to analyse that part of the scene.

2.7. *Summary of the process*

The complete algorithm, as currently implemented, uses four filter sizes. Initially, the two views of the scene are mapped into a pair of working arrays. These arrays are convolved with each filter. The zero crossings and their orientation are computed for each channel. The initial alignments of the eyes determine the initial registration of the images. The matching of the descriptions from each channel is performed for this alignment. Any points with either ambiguous matchings or no match are marked as such.

Next, the percentage of unmatched points is checked, for all square neighbourhoods of a particular size. This size is chosen so as to ensure that the measurement of the statistics of matching within that neighbourhood is statistically sound. Only the disparity points of those regions whose percentage of unmatched points is below a certain threshold are allowed to remain. All other points are removed. These values are stored in a buffer. At this stage, vergence movements may take place, with use of information from the larger channels to bring the smaller channels into a range where matching is possible. Further, if there are regions of the image that do not have disparity values for any channel, an eye movement may take place in an attempt to bring those portions of the image into a range where at least the largest filter can perform its matching.

Note that the matching process takes place independently for each of the four channels. Once the matching of each channel is complete, the results are combined into a single representation of the disparities.

The final output is thus a disparity map, with disparities assigned along most portions of the zero-crossing contours obtained from the smallest filters used. The accuracy of the disparities thus obtained depends on how accurately the zero crossings have been localized, which may, of course, be to a resolution much finer than the initial array of intensity values that constitutes the image.

3. Examples and assessment of performance

Since random dot stereograms (Julesz 1960, 1971) contain no visual cues other than the stereoscopic ones, they are a useful tool for studying the stereo components of the human visual system in isolation. One test of the adequacy of the algorithm as representative of human stereo vision is to compare human perception and the performance of the algorithm on such patterns. Since random dot stereograms have known disparity values, these patterns can also be used to assess the correctness of the performance of the algorithm.

Table 1 lists some of the matching statistics for various random dot patterns. These are illustrated in figures 8–13 and discussed below.

TABLE 1. MATCHES

pattern	density (%)	total	exact	pixel	wrong	percentage wrong
square	50	11847	11830	14	3	0.03
square	25	9661	9632	22	7	0.07
square	10	5286	5264	20	2	0.04
square /	5	3500	3498	0	2	0.06
wedding	50	11162	11095	61	6	0.06
noise-*w*4	50	2270	1909	346	15	0.7
noise-*w*9	50	8683	6621	1868	194	2
noise-*w*4-1	50	63	28	24	11	17
noise-*w*9-1	50	8543	5194	2864	485	6
90% correct	50	9545	9091	263	191	2
80% correct	50	4343	4120	143	80	2
70% correct	50	134	127	2	5	4
diagonal correct	50	6573	6325	271	157	2

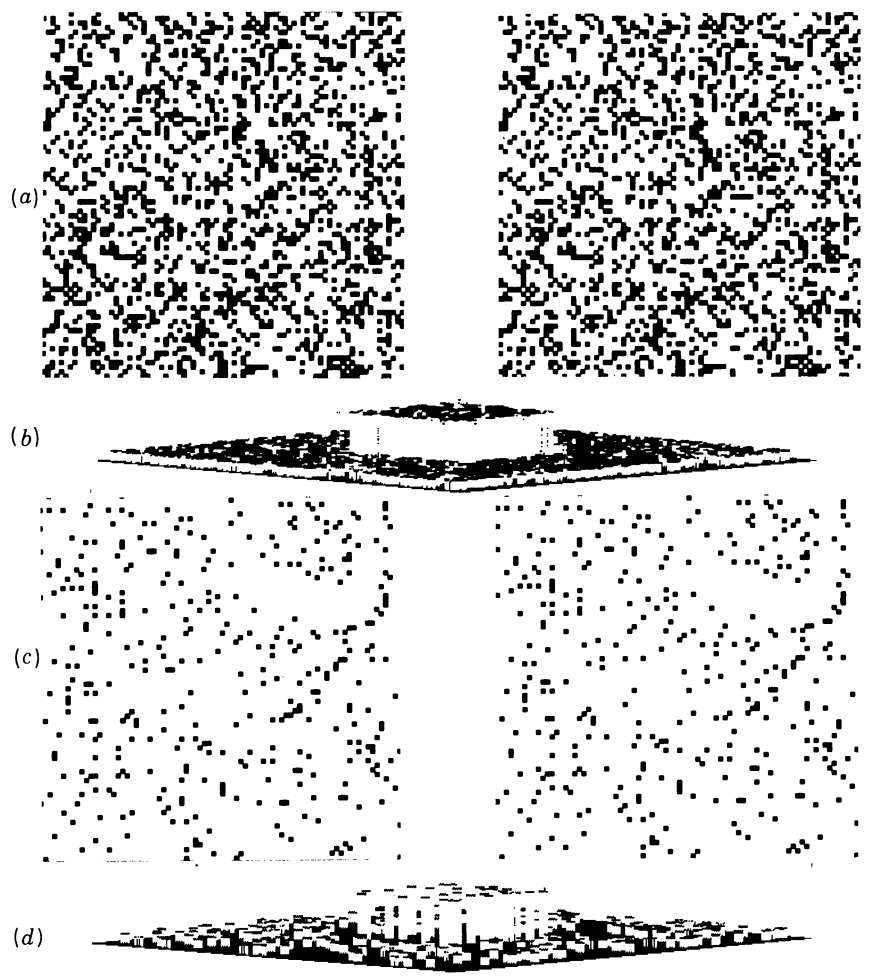


FIGURE 8. The top stereo pair (*a*) is a 25% density random dot pattern. The disparity map (*b*) below it is displayed as in figure 7. The bottom stereo pair (*c*) is a 5% density random dot pattern. Its disparity map (*d*) is shown below it. Both disparity maps are obtained from the $w = 4$ channel.

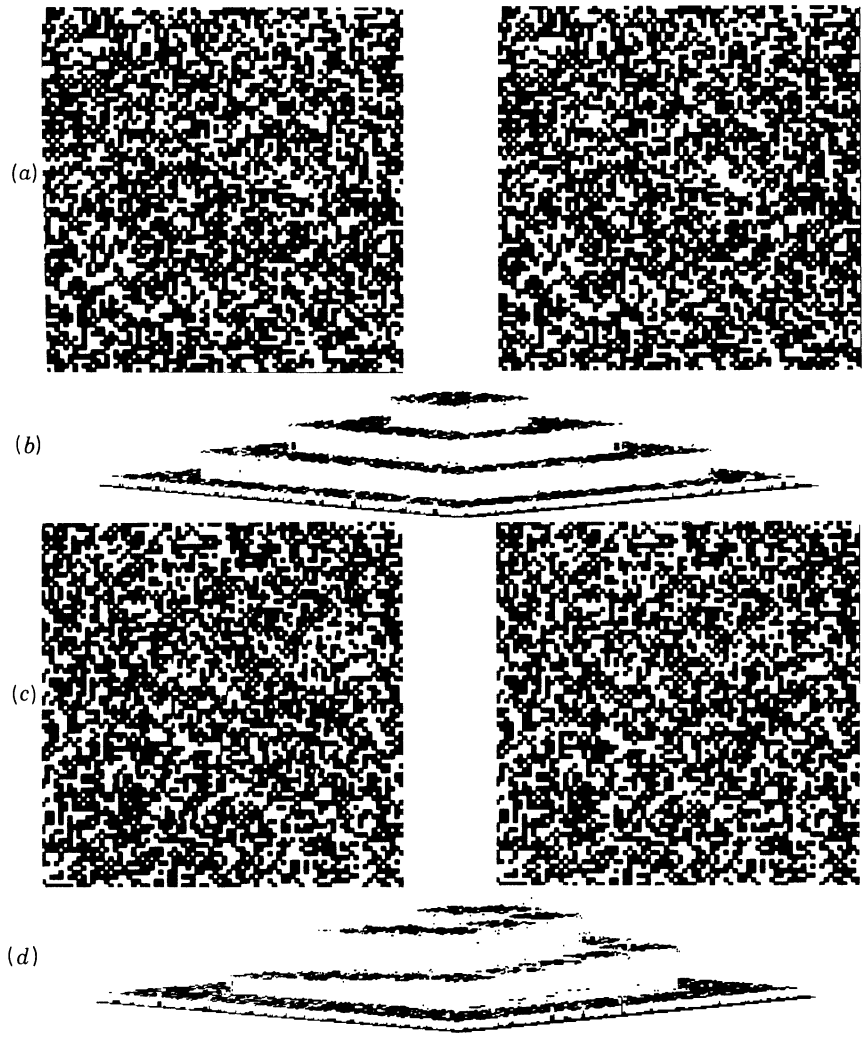


FIGURE 9. The top stereo pair (*a*) is a 50 % density wedding cake, composed of four planar levels. The disparity map (*b*) is shown below it. The bottom stereo pair (*c*) is a 50 % density spiral staircase. The disparity map (*d*) is shown below it, in a manner similar to figure 7. Both disparity maps are obtained from the $w = 4$ channel.

The first pattern consisted of a central square separated in depth from a second plane. The statistics of matching are labelled by the 50 % square row in table 1. The pattern had a dot density of 50 % and its analysis is shown in figure 7. Each dot was a square with four image elements on a side. For the algorithm, this corresponds to a dot of approximately 2′ of visual arc. The total pattern was 320 image elements on a side. The central plane of the figure was shifted 12 image elements in one image relative to the other. The final disparity map assigned after the matching of the smallest channel had the following statistics. The number of zero-crossing points in the left description that were assigned a disparity was 11847. Of these 11847, 11830 were disparity values that were exactly correct, and an additional 14 deviated by one image element from the correct value. Approximately 0.03 % of the matched points, or roughly three points in 10 000, were incorrectly matched.

A similar test was run on patterns with a dot density of 25, 10 and 5 %. These are shown in table 1 in the rows labelled 25 % square, 10 % square and 5 % square. The results are

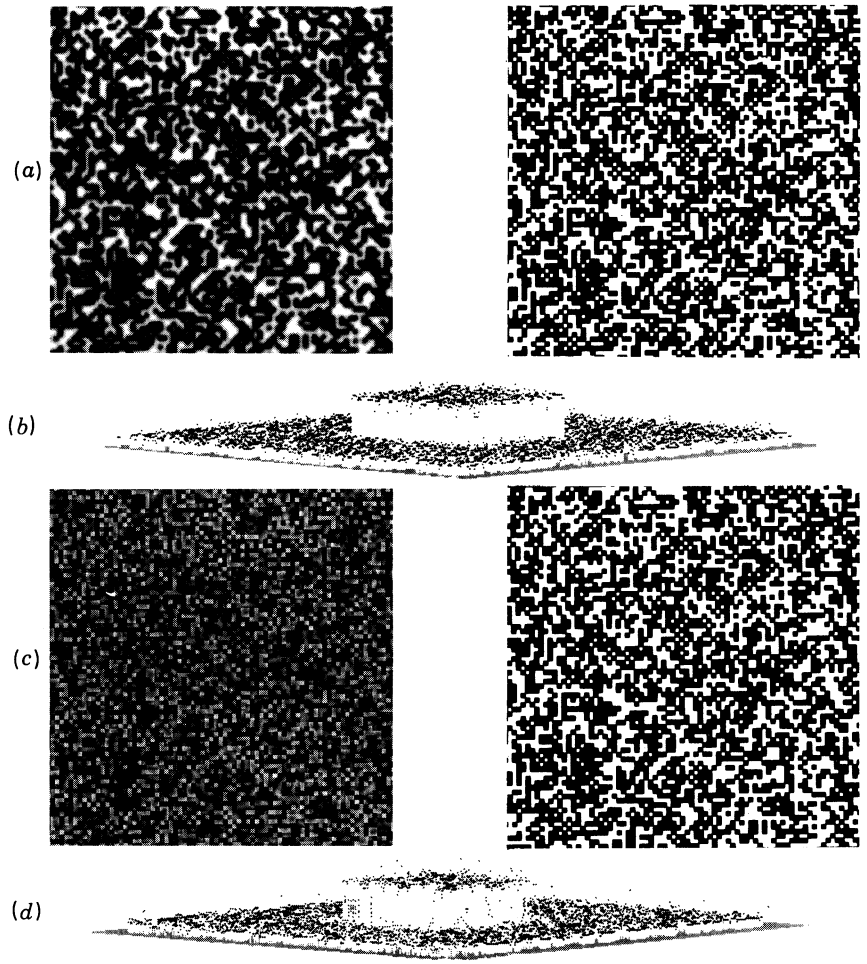


FIGURE 10. The top stereo pair (*a*) is a 50% density pattern in which the left image has been blurred. The disparity map (*b*) is shown below it. It can be seen that two planes are still evident, although they are not as sharply defined as in figure 7 or figure 8. The disparity map is that obtained from the $w = 4$ channel. The bottom stereo pair (*c*) is a 50% density pattern. The left image has had high pass filtered noise added to it so that the maximum magnitude of the noise is equal to the maximum magnitude of the image. The disparity map (*d*) shown is that obtained by the $w = 9$ channel.

illustrated in figure 8. For each of these cases, the number of incorrectly matched points was extremely low, their error rates lying around 0.05%. Those points that were assigned incorrect disparities all occurred at the border between the two planes, that is, along the discontinuity in disparity. This was also true for the 50% density case.

Figure 9 shows a more complex random dot pattern, consisting of a wedding cake built from four different planar layers, each separated by eight image elements, or two dot widths. The matching statistics are shown in table 1 in the row labelled wedding. In this case, the number of zero-crossing points assigned a disparity was 11162. Of these points, 11095 were assigned a disparity value that was exactly correct, and an additional 61 deviated from the correct value by one image element. Approximately 0.06% of the points were incorrectly matched. Again, these incorrect points all occurred at the boundaries between the planes. A second complex pattern is illustrated in figure 9. The object is a spiral staircase with a range of continuously varying disparities.

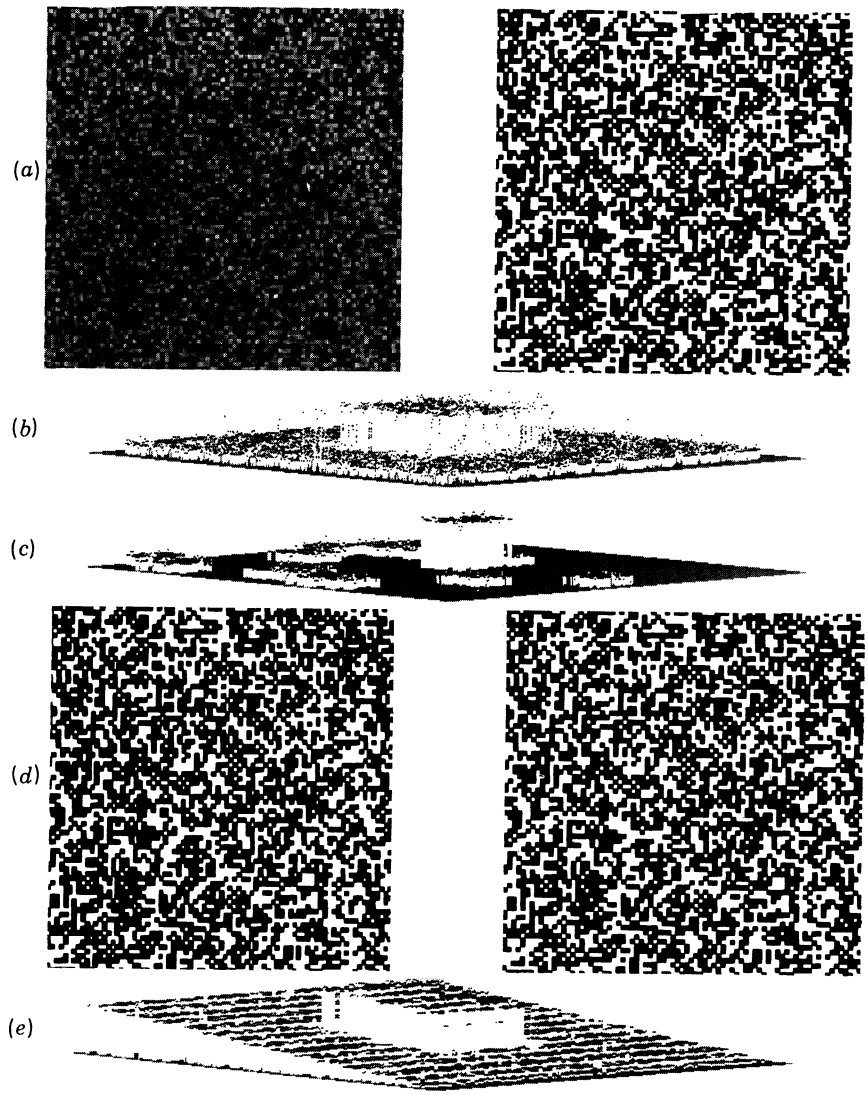


FIGURE 11. The top stereo pair (*a*) is a 50% density pattern. The left image has had high pass filtered noise added to it so that the maximum magnitude of the noise is half the maximum magnitude of the image. The top disparity map (*b*) is that obtained from the $w = 9$ channel, while the next (*c*) disparity map is that obtained from the $w = 4$ channel. It can be seen that the $w = 4$ channel obtains a matching only in a few sections of the image. The bottom stereo pair (*d*) is a 50% density pattern in which the left image has been compressed in the horizontal direction. The disparity map (*e*) from the $w = 4$ is displayed below. It can be seen that the two planes are still evident, although the entire pattern appears slanted. This is in agreement with human perception.

There are a number of special cases of random dot patterns that have been used to test various aspects of the human visual system. The algorithm was also tested on several of these stereograms. They are outlined below and a comparison is given between the performance of the algorithm and of humans with good stereo vision.

It is known that, if one or both of the images of a random dot stereogram are blurred, fusion of the stereogram is still possible (Julesz 1971, p. 96). To test the algorithm in this case, the left half of a 50% density pattern was blurred by convolution with a Gaussian filter. This is illustrated in figure 10. The disparity values obtained were not as exact as for without blurring.

Rather, there was a distribution of disparities about the known correct values. As a result, the percentage of points that might be considered incorrect (more than one image element deviation from the correct value) rose to 6 %. The qualitative performance of the algorithm is still correct, however, representing two planes separated in depth. It is interesting to note that the slight distribution of disparity values about those corresponding to the original planes is consistent with the human perception of a pair of slightly warped planes. For larger filters, there was little difference between the performance of the algorithm on this stereogram and its performance on stereograms that have not been blurred.

Julesz & Miller (1975) showed that fusion is also possible in the presence of some types of masking noise. In particular, if the spectrum of the noise is sufficiently far from the spectrum of the pattern, fusion of the pattern is still possible. Within the framework of the Marr–Poggio theory, this is equivalent to stating that, if one introduces noise of such a spectrum as to interfere with one of the stereo channels, fusion is still possible among the other channels, provided that the noise does not have a substantial spectral component overlapping other channels as well. This was tested on the algorithm by high pass filtering a second random dot pattern, to create the noise, and adding the noise to one image. In the case illustrated in figures 10 and 11, the spectrum of the noise was designed to interfere maximally with the smallest channel. For the patterns labelled in table 1 by noise–$w4$ and noise–$w9$, the noise was added such that the maximum magnitude of the noise was equal to the maximum magnitude of the original image. Noise–$w4$ illustrate the performance of the smallest channel. Noise–$w9$ illustrates the performance of the next larger channel. It can be seen that some fusion is still possible in the smallest channel, although it is patchy. The next larger channel also obtains fusion. In both the accuracy of the disparity values are less than usual. This is to be expected, since the introduction of noise tends to displace the positions of the zero crossings. For the pattern labelled by noise–$w4$–1 and noise–$w9$–1 in table 1, the noise was added such that the maximum magnitude was twice that of the maximum magnitude of the original image. Here, matching in the smallest channel is almost completely eliminated (noise–$w4$–1). Yet matching in the next larger channel is only marginally affected (noise–$w9$–1).

The implementation was also tested on addition of low pass filtered noise to a random dot pattern, with results similar to that of adding high pass filtered noise. Here, the larger channels are unable to obtain a good matching, while the smaller channels are relatively unaffected.

If one of the images of a random dot pattern is compressed in the horizontal direction, the human stereo system is still able to achieve fusion (Julesz 1971, p. 213). The algorithm was tested on this case, and the results are shown in Figure 11. It can be seen that the program still obtains a reasonably good match. The planes are now slightly slanted, which agrees with human perception.

If some of the dots of a pattern are decorrelated, it is still possible for a human observer to achieve some kind of fusion (Julesz 1971, p. 88). Two different types of decorrelation were tested. In the first type, increasing percentages of the dots in the left image were decorrelated at random. In particular, 10, 20 and 30 % were tried, and are illustrated in figure 12. For 10 %, (table entry 90 % correct) it can be seen that the algorithm was still able to obtain a good matching of the two planes, although the total number of zero crossings assigned a disparity decreased, and the percentage of incorrectly matched points increased. When the percentage of decorrelated dots was increased to 20 % (table entry 80 % correct), the number

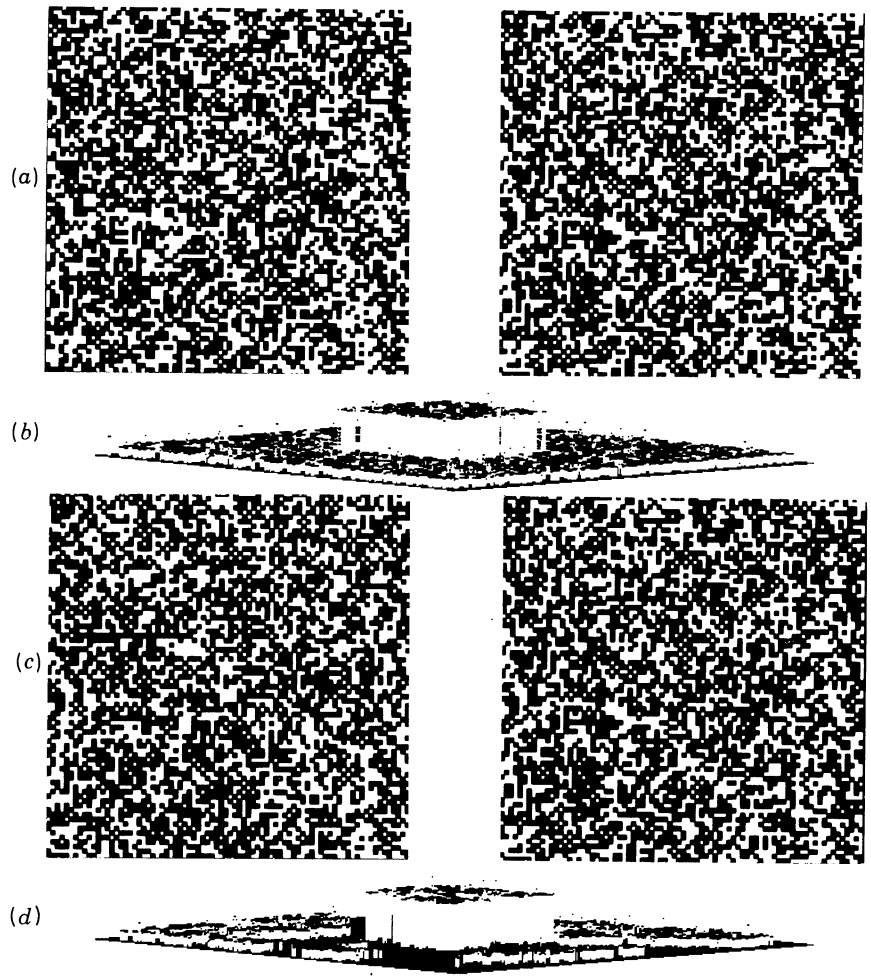


FIGURE 12. The top stereo pair (*a*) is a 50% density pattern in which the left image has had 10% of the dots decorrelated. The disparity map (*b*) is shown below. The bottom stereo pair (*c*) is a 50% density pattern in which the left image has had 20% of the dots decorrelated. The disparity map (*d*) is shown below. Note that in this case there are large regions of the image for which no match was made.

of matched points decreased again, although the percentage of those that were incorrectly matched remained about the same. Finally, when the percentage of decorrelated dots was increased to 30% (table entry 70% correct), the algorithm found virtually no section of the image that could be fused.

The failure of the algorithm to match the 30% decorrelated pattern is caused by the component of the algorithm that checks that each region of the image is within range of correspondence. Recall that to distinguish between the case of two images beyond range of fusion (for the current eye positions), which will have only randomly matching zero crossings, and the case of two images within range of fusion, the theory requires that the percentage of unmatched points is less than approximately 0.3. For the pattern with 30% decorrelation, each region of the image will, on the average, have roughly 30% of its zero crossings with no match and the algorithm will decide that the region is out of matching range. Thus, the algorithm cannot distinguish a correctly matched region of a degraded pattern from the matches that would be made between two random patterns. Hence, no disparities are accepted

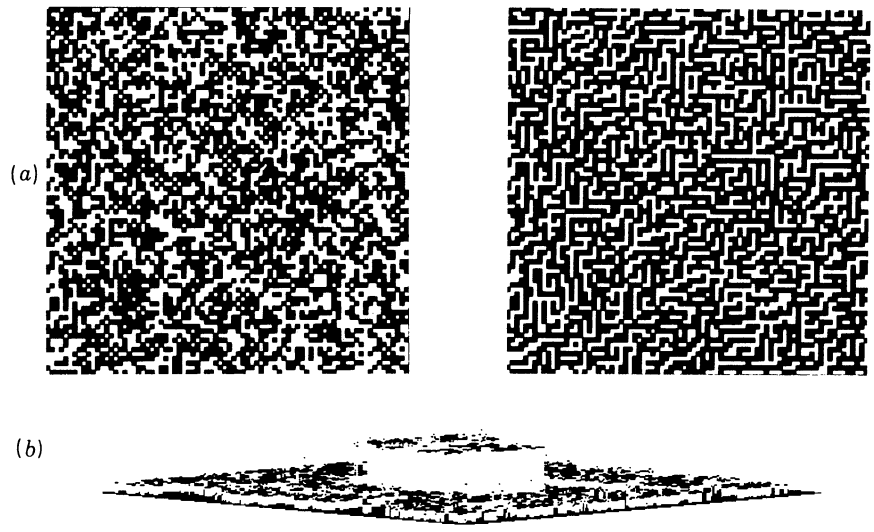


FIGURE 13. The stereo pair (*a*) is a 50 % density pattern in which the right image has been diagonally decorrelated. Along one set of diagonals, every triplet of white dots has been broken by the insertion of a black dot, and along the other set of diagonals every triplet of black dots has been broken by the insertion of a white dot. The disparity map (*b*) is shown below.

for this region. It is interesting to note that many human subjects can achieve some kind of fusion up to about 20 % decorrelation, the fusion becoming weaker as the decorrelation increases, being eliminated for patterns with 30 % decorrelation.

Fascinatingly, one can also decorrelate the pattern by breaking up all white triplets along one set of diagonals, and all black triplets along the other set of diagonals (Julesz 1971, p.87). The table entry 'diagonal correct' indicates the matching statistics for this case. Again, it can be seen that the program still obtains a good match, as do human observers. The performance of the algorithm is illustrated in figure 13. This is a particularly fascinating example, since at first glance it would appear extremely unlikely that the two patterns could be fused. Yet the program is quite consistent with human perception on this example, obtaining a good matching of the two images.

4. Natural images

The algorithm was also tested on some natural images. In such cases, an exact evaluation of the performance of the algorithm is difficult. A qualitative comparison is, however, possible, and the results of the algorithm may be seen in the figure 14.

5. Statistics

The theory of the Marr–Poggio matcher is based on an assumption concerning the distribution of intervals between zero crossings. This leads to assumptions concerning the worst possible occurrences of false targets. The empirical occurrence of false targets has been measured in random dot patterns and the worst occurrences of false targets are indicated in table 2. The theoretical worst case bounds used by Marr & Poggio appear for comparison.

From the table, it can be seen that the assumptions of the Marr–Poggio theory are not

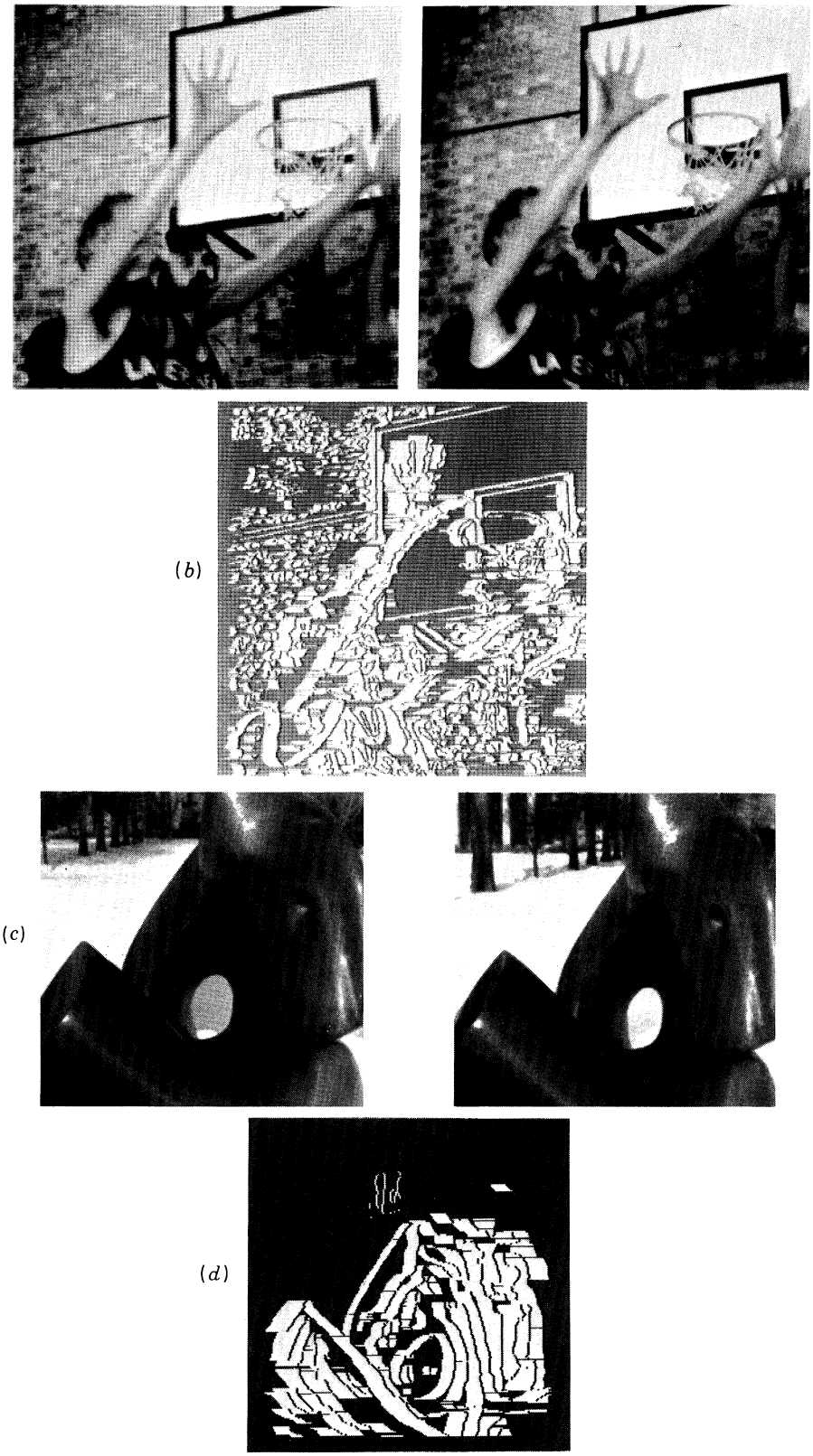


FIGURE 14. Examples of natural images. The top stereo pair (*a*) is a scene of a basketball game. The disparity map (*b*) is represented in such a manner that the width of the white bars, terminated by a black dot, corresponds to the disparity of the point. It can be seen that the disparity values are all qualitatively correct with the arm of the foremost player emerging from the background of the basket and the wall. The images were 480 pixels on a side. The bottom stereo pair (*c*) is a scene of sculpture by Henry Moore. The disparity array (*d*) is represented as in the top stereo pair. It can be seen that the disparity values obtained by the program roughly correspond to the shape of the surface. The images were 320 pixels on a side.

very reliable, compared with the empirical statistics found for random dot patterns. This is in part due to the fact that the analysis performed by Marr & Poggio was based on their assumption of oriented filters. The current implementation uses non-oriented filters to obtain the zero crossings and uses orientation as a matching criterion after the descriptions have been obtained. It then becomes of interest to check whether a proper accounting of the use of non-oriented filters will make statistical predictions more consistent with the empirically observed statistics.

TABLE 2. STATISTICS

parameter	worst case behaviour, parameter	without orientation, empirical	with orientation, empirical
average distance between zero crossings of same sign	$2w$	$1.85w$	$5.56w$
probability of candidates in at most one pool	> 0.50	0.38	0.81
probability of candidates in two pools	< 0.45	0.60	0.19
probability of candidates in all three pools	< 0.05	0.02	0.001
given a candidate near zero, probability of no other candidates	> 0.9	0.75	0.93

For non-oriented filters, the derivation is very similar to that used by Marr & Poggio (1979). Assume that $f(x, y) = \nabla^2 G * I(x, y)$ is a white Gaussian process, where $I(x, y)$ is the image intensity. The problem is to find the distribution of intervals between alternate zero crossings, taken along a horizontal slice of the image.

Assume that there is a zero crossing at the origin, and let $P_1(\tau)$, $P_2(\tau)$ be the probability densities of the distances to the first and second zero crossings. P_1 and P_2 are approximated by the following formulae (Rice 1945, §3.4; Longuet-Higgins 1962, eq. 1.2.1, 1.2.3; Leadbetter 1969).

$$P_1(\tau) = \frac{1}{2\pi}\left(\frac{\psi(0)}{-\psi''(0)}\right)^{\frac{1}{2}} \frac{M_{23}(\tau)}{H(\tau)}\left[\psi^2(0)-\psi^2(\tau)\right]\left\{1+H(\tau)\operatorname{arccot}[-H(\tau)]\right\},$$

$$P_2(\tau) = \frac{1}{2\pi}\left(\frac{1\psi(0)}{-\psi''(0)}\right)^{\frac{1}{2}} \frac{M_{23}(\tau)}{H(\tau)}\left[\psi^2(0)-\psi^2(\tau)\right]\left\{1-H(\tau)\operatorname{arccot}[H(\tau)]\right\},$$

where $\psi(\tau)$ is the autocorrelation of the filter $\nabla^2 G$, a prime denotes differentiation with respect to τ, and

$$H(\tau) = \frac{M_{23}(\tau)}{\sqrt{[M_{22}(\tau)-M_{23}(\tau)]}},$$

$$M_{22}(\tau) = -\psi''(0)\left[\psi^2(0)-\psi^2(\tau)\right]-\psi(0)\,\psi'^2(\tau),$$

$$M_{23}(\tau) = \psi''(\tau)\left[\psi^2(0)-\psi^2(\tau)\right]+\psi(\tau)\,\psi'^2(\tau).$$

It is now necessary to compute the autocorrelation $\psi(\tau)$. The filter is given by

$$\nabla^2 G(r) = \left(\frac{r^2-2\sigma^2}{\sigma^4}\right) \mathrm{e}^{-r^2/2\sigma^2}.$$

Using the two-dimensional Fourier transform and the slice projection theorem (Mersereau & Oppenheim 1974), one finds that the autocorrelation function for this filter is given by

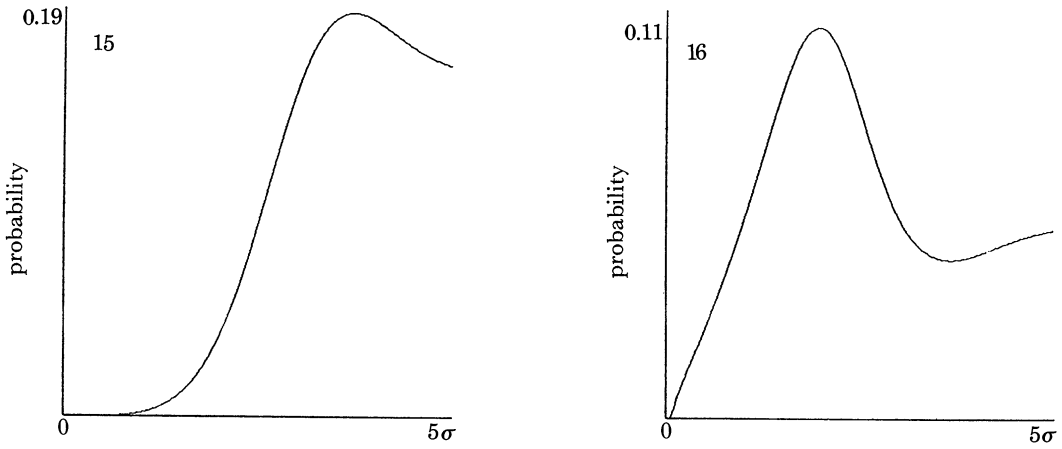


FIGURE 15. Probability distribution of first zero crossing. This is a graph of the probability of obtaining a zero crossing at a certain distance, given a zero crossing at the origin.

FIGURE 16. Probability distribution of second zero crossing. This is a graph of the probability of obtaining a second zero crossing at a certain distance, given a zero crossing at the origin.

$$\psi(\tau) = \pi^{\frac{5}{2}}\sigma\left(\frac{11}{\sigma} - 5\frac{\tau^2}{\sigma^3} + \frac{1}{4}\frac{\tau^4}{\sigma^5}\right)e^{-\tau^2/4\sigma^2}.$$

The formulae of Rice may then be applied to this autocorrelation function, and the probability distributions obtained in this way are shown in figures 15 and 16.

In this case, the expected worst case behaviour of multiple zero crossings within a particular range of matching is somewhat different. In particular, the predicted density of zero crossings is somewhat higher than in the Marr–Poggio case. At first sight, this would seem to suggest that the situation is worse than that given by the Marr & Poggio analysis. The use of orientation as a matching criterion, however, has yet to be included. To do this, some estimate of the distribution of orientation of zero crossings is needed. The matching algorithm segments the orientation distribution into blocks of 30°. The simplest estimate is given by assuming that the orientations are uniformly distributed, in which instance the probabilities given by the statistical analysis should be adjusted by a factor of $\frac{1}{6}$. However, there is no real justification for assuming that the orientations will be uniformly distributed. In fact, the distribution tends to be more strongly weighted towards vertical orientations. Hence, rather than adjusting by a factor of $\frac{1}{6}$, a more pessimistic factor of $\frac{1}{3}$ will be used.

A table comparing predicted and empirical statistics of the distribution of similar zero crossings is shown in table 3. A number of interesting comparisons can be made from this. First, consider the case in which orientation is not used as a matching criterion. The number of multiple targets predicted by the formula of Rice agrees well with the empirical statistics found in practice. For example, if the range of matching neighbourhood is taken as $\pm 2.044\sigma$, Rice's formula predicts a worst case probability of 0.36 double targets. The empirical statistics in this case are 0.33. If the range is extended to $\pm 2.36\sigma$, Rice's formula predicts a worst case probability of 0.49 and the empirical statistics are 0.44. For a range of $\pm 2.84\sigma$, the prediction probability is 0.65 and the empirical statistic is 0.62.

Secondly, the use of orientation as a matching criteria can greatly improve the problem of false targets. For example, let $w_{1\text{-}D}$ be the central panel width of the one-dimensional projection

of the $\nabla^2 G$ operator. This is related to the central panel width, $w_{2\text{-}D}$, of the two-dimensional operator by

$$w_{2\text{-}D} = \sqrt{2}\, w_{1\text{-}D}$$

and these are related to the space constant σ of the operator by

$$w_{2\text{-}D} = 2\sqrt{2}\, \sigma.$$

If the matching range has the value $\pm w_{1\text{-}D}$ then using orientation to within 30° as a matching criterion reduces the percentage of false targets to 0.091. Even for a matching range of $\pm w_{2\text{-}D}$, the percentage of false targets rises only to 0.20.

TABLE 3. STATISTICS

parameter	without orientation, expected	without orientation, empirical	with orientation, expected	with orientation, empirical
average distance between zero crossings of same sign	5.29σ	5.24σ	15.87σ	18.56σ
probability of candidates in, at most, one pool	> 0.35	0.38	> 0.78	0.81
probability of candidates in two pools	< 0.64	0.60	< 0.21	0.19
probability of candidates in all three pools	< 0.01	0.02	< 0.01	0.001
given a candidate near zero, probability of no other candidates	> 0.81	0.75	> 0.94	0.93

This raises an interesting possibility concerning the range of the matching neighbourhood. In the original theory, Marr & Poggio (1979) considered the possibility of avoiding the false targets problem almost entirely by reducing the probability of its occurrence, while maintaining a range of matching consistent with estimates of the size of Panum's area. According to their analysis, however, if the matching range is so restricted as to reduce the probability of false targets to less than 0.05, the range is too small by a factor of 2. If the range of matching is adjusted to account for the size of Panum's area, then the probability of false targets rises to 0.50 and it is necessary to introduce a disambiguation mechanism to resolve the false targets problem.

Given the statistical analysis derived above, one can propose a matching mechanism in which the zero crossings are obtained from non-oriented filters and the orientation of the zero crossing is used as a criterion for matching. In this case, matching over a range consistent with Panum's area will result in very few false targets (on the order of 0.10), and there is no need to introduce a disambiguation mechanism.

6. DISCUSSION

Implementing a computational theory offers us the opportunity of testing its adequacy. In this I have found that the performance of the implementation coincides well with that of human subjects over a broad range of random dot test cases obtained from the literature, including defocusing of, compression of, and the introduction of various kinds of masking noise to one image of a random dot stereo pair.

In running the program, a number of interesting points concerning the form of the algorithm have arisen. These are discussed below.

(i) The neighbourhood over which a search for a matching zero crossing is conducted is

broken into three pools, corresponding to convergent, divergent and zero disparity. In the present implementation, the pools are used to deal with the ambiguous case of two matching zero crossings, while the disparity values associated with a match are represented to within an image element. A second possibility is to use the pools not only to disambiguate multiple matches, but also to assign a disparity to a match. Thus, a single disparity value, equal to the disparity value of the midpoint of the pool, would be assigned for a matching zero crossing lying anywhere within the pool. In this scheme, only three possible disparities could be assigned to a zero crossing; zero, corresponding to the middle pool, or $\pm w/2$, corresponding to the divergent or convergent pools.

Interestingly, computer experiments show that either scheme will work. For a single disparity value for each pool, the disparities assigned by the smallest channel are within an image element of those obtained by means of exact disparities for each match. This modification was tried on both natural images and random dot patterns, and suggests that the accuracy with which the pools represent the match is not a critical factor.

(ii) The points that were incorrectly matched in the test cases all lay along depth discontinuities. The major reason for this is connected with occlusion of regions. Note that at any depth discontinuity there will be an occluded region that is present in one image, but not in the other. Any zero crossings within that region cannot, of course, have a matching zero crossing in the other image. However, there is a certain probability of such a zero crossing being matched incorrectly to a random zero crossing in the other image. In principle, the algorithm detects regions that are occluded, by checking the statistics of the number of unmatched zero crossings and using such results to mark all zero crossing matches in the region as unknown. However, for a region that contains a depth discontinuity, only part of the region will have the above characteristics. Zero crossings in the rest of the region will have a unique match. Thus, when the statistical check on the number of unmatched points is performed, it is possible for the entire region to be considered in range, and thus all matches, including the incorrect ones of the occluded region, will be accepted.

(iii) It is interesting to comment on the effect of depth discontinuities for the different sizes of mask. For random dot patterns, the zero crossings obtained from the larger masks tend to outline blobs or clusters of dots. Thus in general, the positions of the zero crossings do not correspond to single elements of the underlying image. Suppose that the dot pattern consists of one plane separated in depth from a second plane. In such a case, one might well find a zero crossing that belongs at one end to dots on the first plane, and at the other end to dots belonging to the second plane. Such zero crossings will be assigned disparities that reflect, to within the resolution of the channel, the structure of the image. The zero crossings lying between the two ends will, however, receive disparities that smoothly vary from one extreme to the other. The largest channel would thus not see a plane separated in depth from a second plane, but rather a smooth hump.

For the smaller mask this does not occur, as the zero crossing contours tend to outline individual dots or connected groups of dots. Thus the disparities assigned are such that the dots belong to one plane or the other and the final disparity map is one of two separated planes.

To achieve perfect results from stereo, it is probably necessary to include in the $2\frac{1}{2}$-dimensional sketch a way of dealing competently with discontinuities. Some initial work has already been done in this direction (Grimson 1980). Interestingly, when one looks at a 5% random

dot stereogram portraying a square in front of its background, one sees vivid subjective contours at its boundary, although the output of the matcher does not account for this.

(iv) An integral part of most computational theories, proposed as models of aspects of the human visual system, is the use of computational constraints based on assumptions about the physical world (Marr & Poggio 1979; Marr & Hildreth 1980; Ullman 1979). The constraints so derived are critical in the formation of the computational theory and in the design of an algorithm for solving the problem. An interesting question to raise is whether the algorithm explicitly checks that the constraints imposed by the theory are satisfied. For example, Ullman's rigidity constraint in the analysis of structure from motion is explicitly checked by his algorithm. For the Marr–Poggio stereo theory, two constraints were outlined, uniqueness and continuity of disparity values. It is curious that, in the algorithm used to solve the stereo problem, the continuity constraint is explicitly checked while the uniqueness constraint is not. Uniqueness of disparity is required in one direction of matching, since only those zero crossing segments of one image that have exactly one match in the second image are accepted. However, it may be the case that more than one element of the right image could be matched to an element of the left image. When matching from the right image to the left, the same is true. Note that one could easily alter the algorithm to include the checking of uniqueness, thereby retaining only those disparity values corresponding to zero crossing segments with a unique disparity value when matched from both images. However, the evidence of Braddick, discussed in the next section, would indicate that this is not so. Hence, in the Marr–Poggio stereo theory, although both the requirement of uniqueness and continuity are subsumed, only one of these two constraints is explicitly checked by the algorithm. The reason the other constraint is not checked is probably because it is physically very unlikely to be violated.

(v) There are a number of questions concerning the form of the $2\frac{1}{2}$-D sketch, which have yet to be firmly answered. Some of these problems, and the results of experimentation with the implementation as it relates to them, are indicated below.

The first critical question concerns whether the sketch uses the coordinates of the scene or of the working arrays. In the first case, the coordinates of the sketch would be directly related to the coordinates of the arrays of the entire scene. The advantage of this is that, since disparity information about the scene is extracted from several eye positions, the representation of the disparities over the entire scene can readily be updated. However, this advantage also raises a difficulty. To store this information into a buffer with coordinate system connected to the image of the scene, explicit information about the positions of the eyes is required. This is fine computationally, but, for a model of the human visual system, it may be that such information is not available to the stereo process.

In the second case, no such problem arises. Here, the coordinates of the sketch are directly related to the coordinates of the retinal images. Such a system is called retinocentric, since it reflects the current positions of the eyes. As such, it does not require explicit knowledge of the eye positions relative to some fixed coordinate system within the scene, and thus it seems to be the most natural representation. This then raises the question of how information about disparities in the scene are maintained across eye movements.

The second question concerns the use of a fovea. Different sections of the images are analysed at different resolutions, for a given position of the optical axes. An important consequence of this is that the amount of buffer space required to store the disparity will vary widely in the visual field, being much greater for the fovea than for the periphery. This also suggests

the use of a retinocentric representation, because, if one used a frame that had already allowed for eye-movements, it would have to have foveal resolution everywhere. Not only does such a buffer waste space, but it does not agree with our own experience as perceivers. If such a buffer were used, we should be able to build up a perceptual impression of the world that was everywhere as detailed as it is at the centre of the gaze, and this is clearly not the case.

The final point about the $2\frac{1}{2}$-D sketch is that it is intended as an intermediate representation of the current scene. It is important for such a representation to pass on its information to higher level processes as quickly as possible. Thus, it probably cannot wait for a representation to be built up over several positions of the eyes. Rather, it must be refreshed for each eye position.

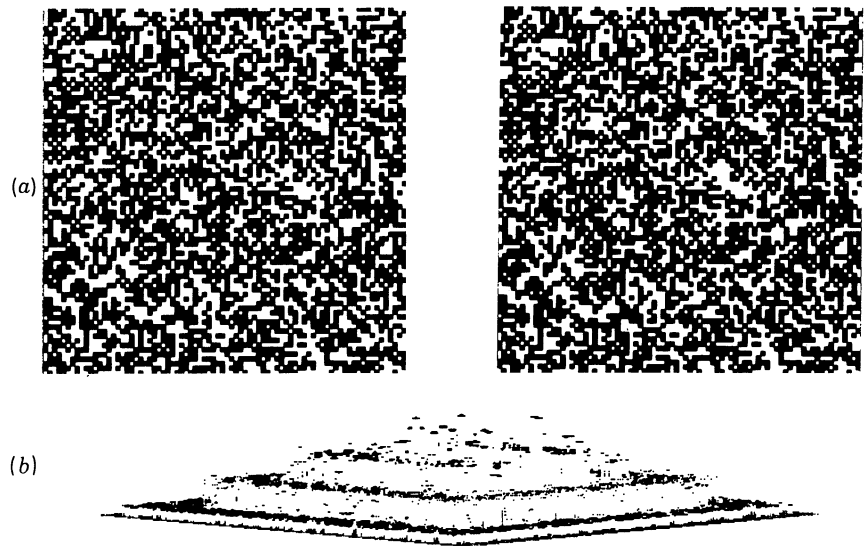


FIGURE 17. Single fixation example. The top images (*a*) are a random dot 'wedding cake'. The disparity map (*b*) is that obtained by combining the different sized channels for a single fixation of the eyes. In this case, the eyes were fixated at the level of the bottom, outermost plane. It can be seen that the disparity values for the bottom plane are very sharp, since at this level the smallest channels are able to make the correspondences. For the region in the centre of the image, the disparity values are much sparser and less accurate. This is since at this level, only the larger channels are within range of correspondence. The reader may check this perception by viewing the stereogram, while fixating only on the bottom plane.

All of these factors combine to suggest a refinement to the implementation, as outlined above. In particular, a retinocentric representation, which represents disparities with decreasing resolution as eccentricity increases, should be used.

For the cases illustrated in this article, the $2\frac{1}{2}$-D sketch was created by storing fine resolution disparity values into a representation with a coordinate system identical to that of the scene. As we have argued above, a second alternative is to store values from all channels into a retinocentric representation, using disparity values from the smaller channels where available, and the coarser disparities from the larger channels elsewhere. In this way, a disparity representation for a single fixation of the eyes may be constructed, with disparity resolution varying across the retina. Such a method of creating the $2\frac{1}{2}$-D sketch has been tested on the implementation, and is indicated in figure 17.

(vi) In the first part of this paper, we have seen that the algorithm performs well on a wide range of random dot patterns, and in fact is consistent with human perception on that

set of patterns. We have also seen that the algorithm can perform well on some natural images. However, there are some differences between using random dot patterns as input and using natural images as input, which can result in a difference in performance on natural images.

First, random dot patterns consist of synthesized intensity values, while the intensity values of natural images are subject to many factors of the imaging process. This can serve to add 'noise' to the intensity values, and this in turn can affect the positions and orientations of the zero crossings. Secondly, the vertical alignment of random dot patterns is a trivial matter, while for natural images the vertical alignment is important. When random dot stereograms are synthesized, it is simple to ensure that there is exact vertical alignment between the elements. However, in natural images, this is not so. Even if the two images of a natural scene are aligned vertically with respect to some object in the scene, the process of projection may cause other regions of the scene to be slightly misaligned vertically, unless the optical axes intersect.

The algorithm can be modified to account for this vertical deviation by allowing both horizontal and vertical alignment of the images to be controlled. That is, when the positions of the eyes are specified to the algorithm, they include a vertical as well as a horizontal displacement relative to one another. For example, suppose that the disparity values from one of the larger channels specify a particular horizontal alignment of the eyes. The algorithm will make this adjustment and match the zero-crossing descriptions of the smaller channels accordingly. If the smaller channels do not obtain a match, it may be because of a slight vertical misalignment, and the matching is repeated for this horizontal adjustment, with a slight vertical alignment of the two images also taking place. In all the cases tested on the implementation, the total range of vertical deviation across the image was small, of the order of two or three image elements.

(vii) In the first part of this paper, I investigated the performance of the Marr–Poggio algorithm on a wide range of random dot patterns, and indicated that its performance was consistent with that of human perception. When turning to natural images, we have seen that the algorithm also seems to perform well. However, there are situations in which the algorithm can return disparity values inconsistent with other information in the image. The question is whether this reflects a basic error in the theory or its implementation, or whether there are other aspects of the visual process interacting with stereo that have not been accounted for in this implementation.

The results of testing the implementation on the broad range of images demonstrate that the matching module is acceptable as an independent module. In particular, the agreement between the performance of the algorithm and that of human observers on the many random dot patterns demonstrates that the matching module is acceptable, since in these cases all other visual cues have been isolated from the matcher.

When turning to natural images, it is reasonable to expect that other visual modules may affect the input to the matcher and that they may alter the output of the matcher. For example, the evidence of Kidd *et al.* (1979) concerning the ability of texture boundaries to drive vergence eye movements indicates that other visual information besides disparity may alter the position of the eyes and thus the input to the matcher. However, it does not necessarily imply that the theory of the matcher itself is incorrect.

Interestingly, the performance of the implementation supports this point. The implementation, which is considered a distinct module, also performs very well on random dot patterns, where

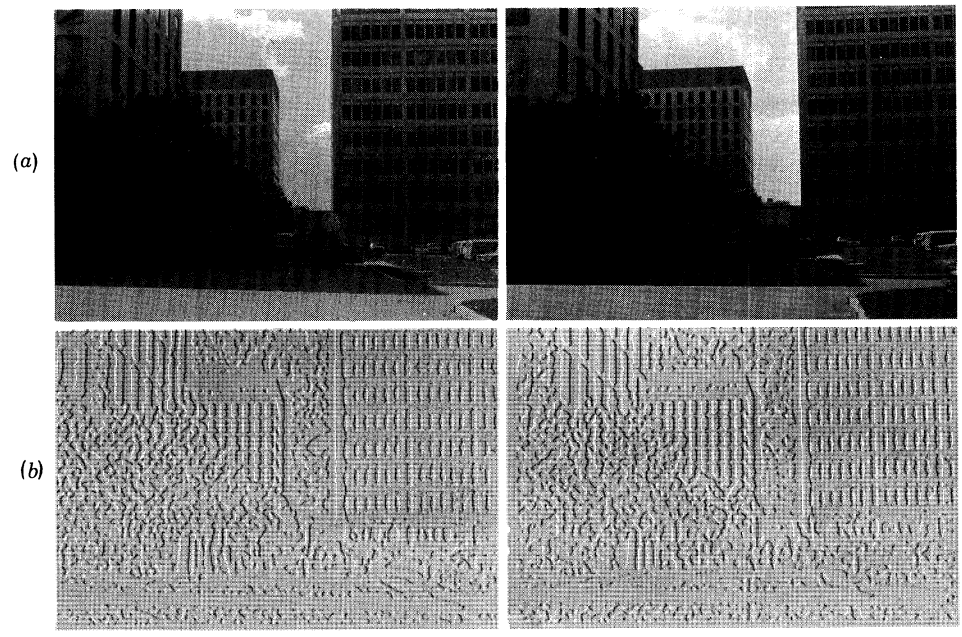


FIGURE 18. The false targets problem. (*a*) A stereo pair of a group of buildings. (*b*) The zero-crossing descriptions of these images. The regular pattern of the windows of the rear building causes difficulties for the matcher. If the alignment of the eyes corresponds to fixating at the level of the building, the algorithm matches the zero crossings corresponding to the windows correctly. If the alignment of the eyes corresponds to fixating at the level of the trees in front of the building, the algorithm matches the zero crossings corresponding to the windows incorrectly. Experiments indicate that under similar conditions humans have a similar perception.

there is no possibility of interaction with other visual processes. For many natural images, this is still true. However, occasionally, it is the case that a natural image provides some difficulty for the implementation. A particular example of this occurs in the image of figure 18. Here, the regular pattern of the windows provides a strong false targets problem. In running the implementation, the following behaviour was observed. If the initial vergence position was at the depth of the building, the zero crossings corresponding to the windows were all assigned a correct disparity. If, however, the initial vergence position was at the depth of the trees in front of the building, the windows were assigned an incorrect disparity, due to the regular pattern of zero crossings associated with them. Clearly, this seems wrong. Yet the question to ask is whether the implementation is wrong. Curiously, if one fuses the zero-crossing descriptions without eye movements, human observers have the same problem: if the eyes are fixated at the level of the building, all is well; if the eyes are fixated at the level of the trees, the windows are incorrectly matched. I would argue that this implies that the implementation, and hence the theory of the matching process is in fact correct. Given a particular set of zero crossings, the module finds any acceptable match and writes it into a buffer. When the output of this buffer is sent to the $2\frac{1}{2}$-D sketch, it must be made consistent with other sources of information feeding the $2\frac{1}{2}$-D sketch. In this case, it is possible that some later processing module is capable of altering the disparity values, based on other information unavailable to the stereo process, and the correct depth is written into the $2\frac{1}{2}$-D sketch.

Thus, I would suggest that future refinements to the Marr–Poggio theory must account for the interactions of other aspects of visual information processing with the input and output of the matching module.

7. Development of the implementation

We have indicated earlier that one reason for implementing a computational theory is that it offers us the opportunity to test the theory's adequacy.

A second reason for implementing a computational theory is that the implementation serves as a useful feedback device for the theory, indicating errors or omissions in the theory, as well as indicating areas whose difficulty had not been previously appreciated. Throughout the course of the development of the stereo implementation, a number of interesting observations were made. Some of these indicated equivalent methods of implementation that had interesting properties with regard to alternative theories of the process. Others served to correct assumptions made by the theory. Still other observations arose at surprising places, places where no difficulty was expected in the implementation process. In many of these cases, in finding a way around the problem, decisions of wide ranging effect were made. This is particularly true of the question of zero crossings and the question of non-oriented filters. Thus, we shall see an example of a problem of implementation causing major changes in the theory of early visual processing. Without the act of implementation, such effects might not have been found. These observations are discussed in the following section.

(i) Although the Marr–Poggio matcher is designed to match from one image into the other, there is no inherent reason why the matching process cannot be driven from both eyes independently. In fact, there may be some evidence that this is so, as is shown by the following experiment of Braddick (1978) on an extension to Panum's limiting case. First, a sparse random dot pattern was constructed. From this pattern, a partner was created by displacing the entire pattern by slight amounts to both the left and the right. Thus, for each dot in the right image, there corresponded two dots in the left image, one with a small displacement to the left and one with a small displacement to the right. The perception obtained by viewing such a random dot stereogram is one of two superimposed planes.

Suppose that the matching process were driven from only one image, for example, from the right image to the left. In this case, the implementation would not be able to account for Braddick's results, since all the zero crossings would have two possible candidates. However, suppose that the matching process were driven independently from both the right and left images, an unambiguous match from either side being accepted. In this case, although every zero crossing in the right image would have an ambiguous match, the program would obtain a unique match for each zero crossing in the left image.

Braddick's case has been tested on the program, and the results are shown in figure 19. It can be seen that the results of the implementation are that of two transparent planes, as in human perception.

An interesting idea is that there may be half stereo blind people who can see two planes when the images are presented such that the double image is in one eye, and who see none or one fuzzy plane when the double image is in the other eye.

(ii) Although this point has been extensively treated elsewhere (Marr & Hildreth 1979; Hildreth 1980), it is interesting to recount the historical development of the use of non-oriented filters.

In the original implementation, oriented filters were used rather than non-oriented ones, in part because the original Marr–Poggio theory was based on them. This was motivated by physiological considerations, but actual practice showed that there were severe difficulties with using such filters. Two effects were particularly noticeable. The first is that such bar-

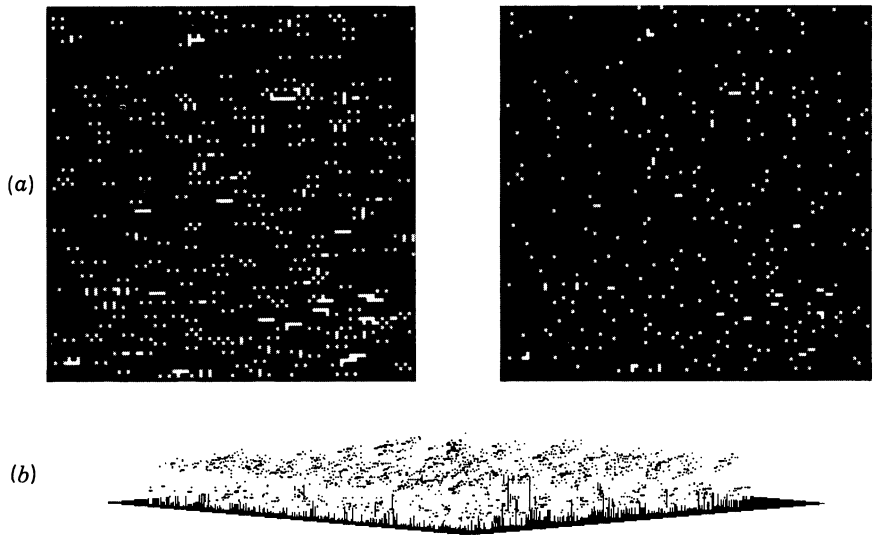


FIGURE 19. Panum's special case. The pair (*a*) is a special case of Panum's limit. The left image is formed by superimposing two slightly displaced copies of the right image. The disparity map (*b*) is shown below, and consists of two superimposed planes.

shaped filters tend to smear or stretch zero crossings in the direction of the orientation of the filter relative to the actual edges in the image. For example, a bar filter that is oriented vertically tends to convert a circular object in a scene into an oval zero-crossing contour. This is not desirable, since any matching performed on such zero crossings will result in disparity values being assigned to locations in the image for which there is no evidence that such assignments are valid.

Some experiments with the aspect ratio of such bar-shaped filters indicated, as might be expected, that the lower the aspect ratio the smaller the amount of stretching of the zero-crossing contours.

A second effect concerned an edge with an orientation different from that of the filter. For this, the resulting zero crossings suffered from the following problem. Rather than giving a straight zero-crossing contour along the orientation of the edge, the filter produced a zero-crossing contour whose overall orientation was that of the edge, but that curved about the edge in a snake-like fashion. Thus, the resulting zero-crossing contour had a significant number of segments consisting of components in the direction of the filter orientation, rather than in the direction of the edge. Again, any matching process that assigns disparities based on such descriptors will be assigning disparity values that do not accurately reflect the structure of the underlying surface.

Both of these effects led to a questioning of the necessity for oriented filters, and, in fact, such filters were replaced by circularly symmetric ones. A comprehensive analysis of non-oriented filters was developed by Marr & Hildreth (1980).

What is of interest here is the fact that attempts at implementing the early version of the stereo theory led to practical difficulties. In overcoming these problems, a major modification of the theory took place.

(iii) We have already seen in §4.3 that the statistical analysis performed by Marr & Poggio

is not consistent with the observed statistics of zero crossings. This is due to the change in operators from the oriented operators used by Marr & Poggio, to the non-oriented operators used in the implementation discussed here. By redoing the statistical analysis for non-oriented filters, I have been able to propose a modification to the matching algorithm that simplifies its operation.

(iv) An earlier implementation of the theory did not use zero crossings, but rather attempted to create a symbolic description of the changes in an image by means of peaks of the convolved output. Several difficulties were encountered.

One difficulty concerned the side lobe effect. By this, I mean that even for an isolated edge the convolved values would exhibit not only a peak corresponding to the edge, but a pair of side lobe peaks of the opposite sign. This not only made the matching task much harder, but also added to the symbolic description zero crossings that did not reflect actual changes in the image intensities. Any matching based on such descriptors was therefore likely to assign disparity values that did not reflect the structure of the underlying part of the surface. Moreover, it was not possible to distinguish locally between '*real*' peak locations and '*false*' or side lobe peaks.

A second difficulty concerned the difference between a one-dimensional peak and a two-dimensional peak. Since the matching takes place along horizontal slices of the convolved image, one could define a peak as any local extremum along that slice. However, the symbolic descriptions generated in this manner will differ greatly from those generated by locating local extrema in two dimensions, where the point is required to be a local extremum along both axial directions. This is in contrast to the case of zero crossings, where the zero crossings generated by scanning along a horizontal direction are virtually identical to those generated by examining a two-dimensional neighbourhood. In fact, the zero crossings not generated by the one-dimensional scan are not relevant to the stereo matching process, since they correspond to horizontally oriented zero-crossing segments which do not have a precise disparity associated with them. Returning to the question of peaks, we see that, if only the two-dimensional peaks are matched, the density of such features is much smaller than that of zero crossings. On the other hand, matching of the one-dimensional peaks may lead to difficulties, since the locations of such peaks need not be sharply localized.

All of these difficulties led to the use of zero crossings as a matching primitive rather than peaks.

This is a particularly interesting illustration of the role of an implementation in developing a computational theory. In this case, the implementation led to questions about a particular aspect of the process whose resolution had wide reaching effects. Thus, the stereo implementation brought to light a problem that had not previously been considered, and the resolution of that problem has significantly altered the theory of several other processes (for example, the theory of edge detection and the primal sketch (Marr & Hildreth 1979)).

8. A FINAL EXAMPLE

As a final demonstration of the stereo implementation, figure 20 illustrates that stereopsis is not strictly a terrestrial phenomenon. The figure illustrates a stereo image of the Martian surface, together with an interpolated disparity map.

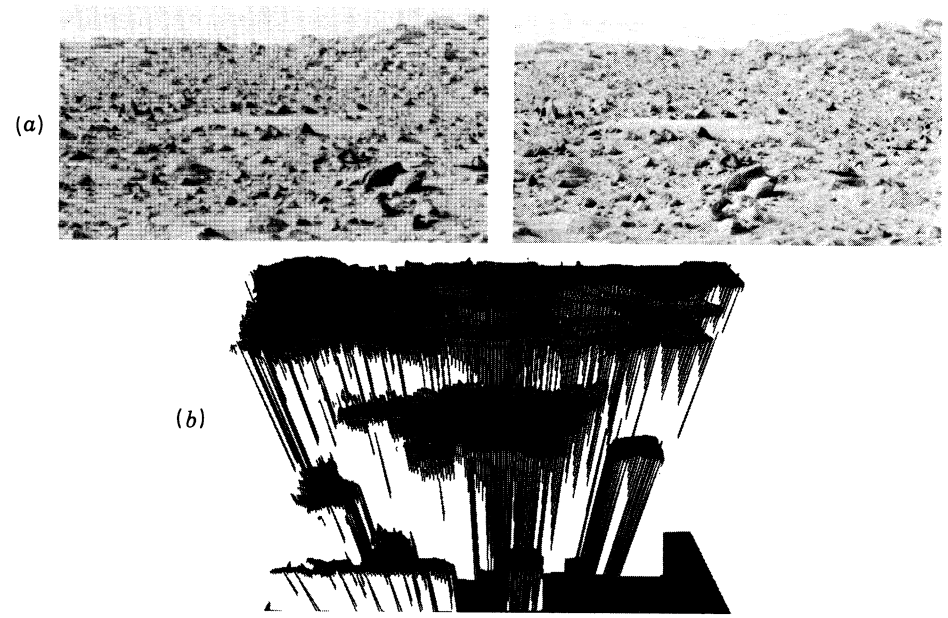


FIGURE 20. The interpolated Martian surface. The top pair of images (*a*) are a stereo pair of the Martian surface. (*b*) The disparity map, which has been interpolated between the known disparity points of the Marr–Poggio algorithm (see Grimson (1980) for details). The height of a point above the reference plane corresponds to the distance from the viewer to the point in the image. The total range of disparity in this image is roughly 200 pixels. Some sections of the foreground were not matched by the algorithm, and have not been interpolated. It is interesting to note that the disparity map contains a series of sharp breaks in disparity, corresponding to occluding hills in the image. These breaks are not evident in the monocular images, yet are clearly visible when the two images are fused.

Without David Marr and Tomaso Poggio, this work would have been impossible. Ellen Hildreth, Keith Nishihara and Shimon Ullman provided many useful comments and suggestions.

REFERENCES

Braddick, O. 1978 Multiple matching in stereopsis. (M.I.T. internal report.)

Campbell, F. W. & Robson, J. 1968 Application of Fourier analysis to the visibility of gratings. *J. Physiol., Lond.* **197**, 551–566.

Grimson, W. E. L. 1980 Computing shape using a theory of human stereo vision. Ph.D. thesis, M.I.T.

Grimson, W. E. L. & Marr, D. 1979 A computer implementation of a theory of human stereo vision. In *Proceedings: Image Understanding Workshop (Palo Alto, California)*, pp. 41–47. Arlington, Virginia: Science Applications.

Hildreth, E. C. 1980 Implementation of a theory of edge detection. *MIT Artificial Intelligence, Tech. Rep.*, no. 579.

Julesz, B. 1960 Binocular depth perception of computer-generated patterns. *Bell System Tech. J.* **39**, 1125–1162.

Julesz, B. 1971 *Foundations of cyclopean perception.* Chicago: University of Chicago Press.

Julesz, B. & Miller, J. E. 1975 Independent spatial-frequency-tuned channels in binocular fusion and rivalry. *Perception* **4**, 125–143.

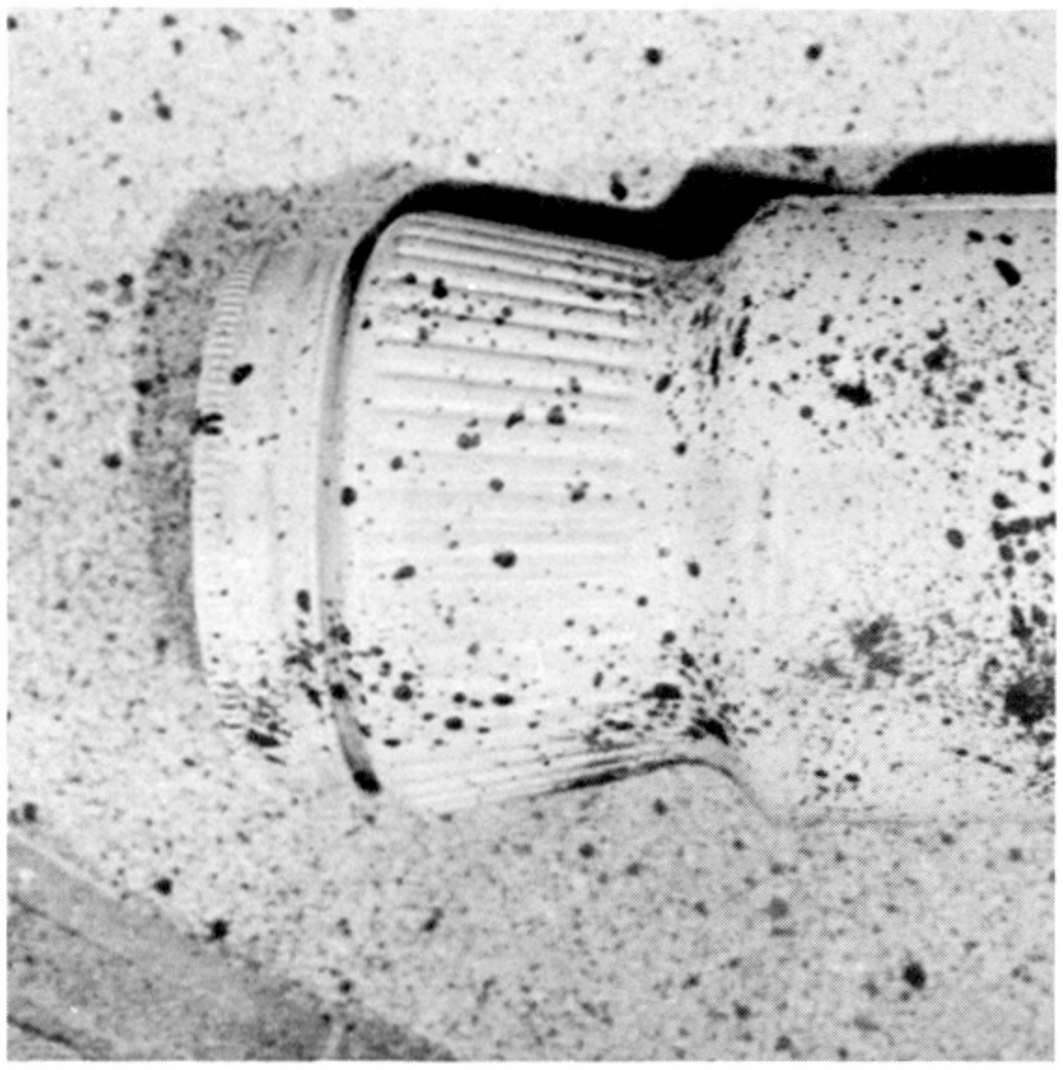
Kidd, A. L., Frisby, J. P. & Mayhew, J. E. W. 1979 Texture contours can facilitate stereopsis by initiating appropriate vergence eye movements. *Nature, Lond.* **280**, 829–832.

Knight, T. F., Moon, D. A., Holloway, J. & Steele, G. L. 1979 *CADR MIT Artificial Intelligence Lab. Memo.* no. 528.

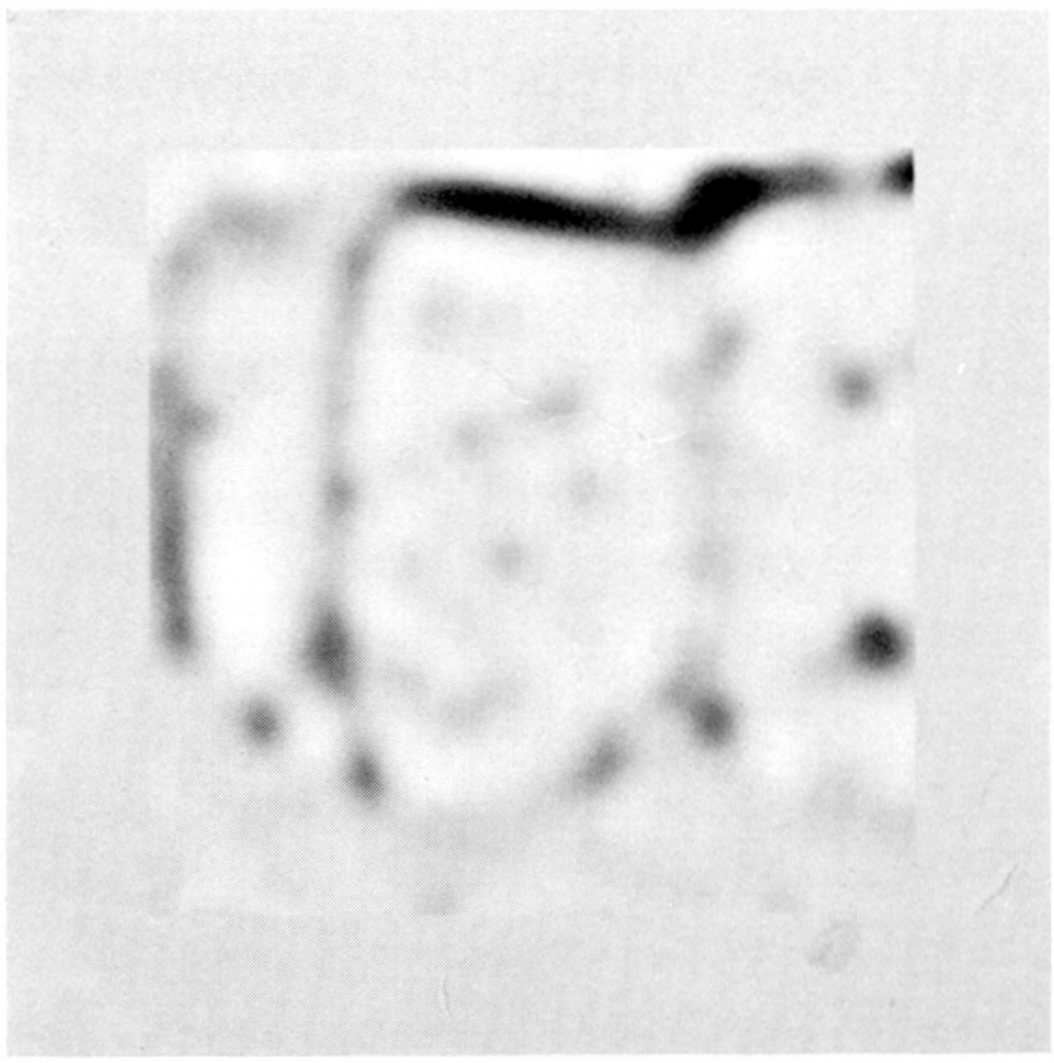
Leadbetter, M. R. 1969 On the distributions of times between events in a stationary stream of events. *Jl R. Statist. Soc.* B **31**, 295–302.

Longuet-Higgins, M. S. 1962 The distribution of intervals between zeros of a stationary random function. *Phil. Trans. R. Soc. Lond.* A **254**, 557–599.

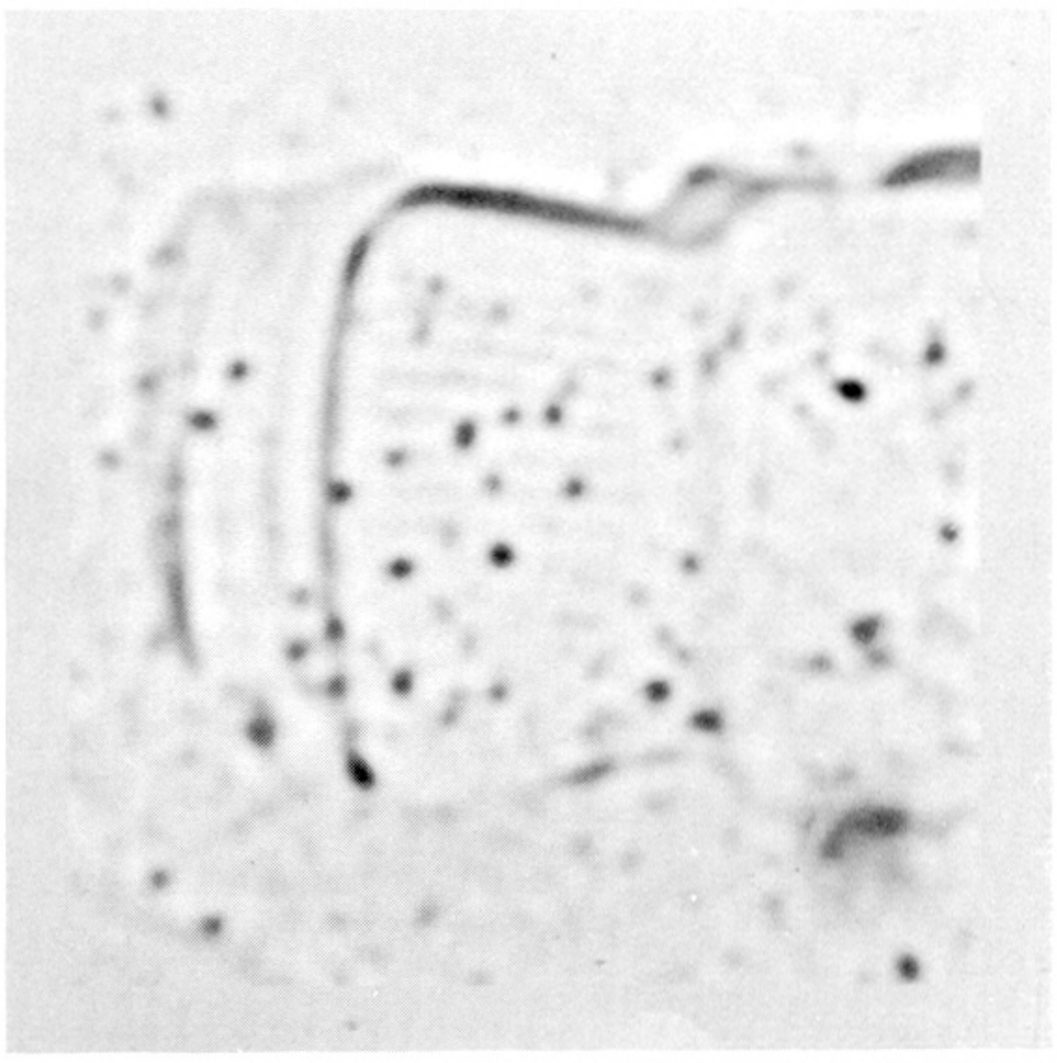
Marr, D. 1976 Early processing of visual information. *Phil. Trans. R. Soc. Lond.* B **275**, 483–534.
Marr, D. & Hildreth, E. 1980 Theory of edge detection. *Proc. R. Soc. Lond.* B **207**, 187–217.
Marr, D. & Nishihara, H. K. 1978 Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond.* B **200**, 269–294.
Marr, D. & Poggio, T. 1976 Cooperative computation of stereo disparity. *Science, N.Y.* **194**, 283–287.
Marr, D. & Poggio, T. 1979 A computational theory of human stereo vision. *Proc. R. Soc. Lond.* B **204**, 301–328.
Marr, D., Poggio, T. & Hildreth, E. 1980 The smallest channel in early human vision. *J. Opt. Soc. Am.* **70**, 868–870
Mayhew, J. E. W. & Frisby, J. P. 1976 Rivalrous texture stereograms. *Nature, Lond.* **264**, 53–56.
Mayhew, J. E. W. & Frisby, J. P. 1978 Stereopsis masking in humans is not orientationally tuned. *Perception* **7**, 431–436.
Mersereau, R. M. & Oppenheim, A. V. 1974 Digital reconstruction of multi-dimensional signals from their projections. *Proc. Inst. Elect. Electron. E* **62**, 1319–1338.
O'Brien, B. 1951 Vision and resolution in the central retina. *J. Opt. Soc. Am.* **41**, 882–894.
Rice, S. O. 1945 Mathematical analysis of random noise. *Bell Syst. Tech. J.* **24**, 46–156.
Ullman, S. 1979 *The interpretation of visual motion.* Cambridge, Massachusetts: M.I.T. Press.
Wilson, H. R. & Bergen, J. R. 1979 A four mechanism model for threshold spatial vision. *Vision Res.* **19**, 19–32.
Wilson, H. R. & Giese, S. C. 1977 Threshold visibility of frequency gradient patterns. *Vision Res.* **17**, 1177–1190.



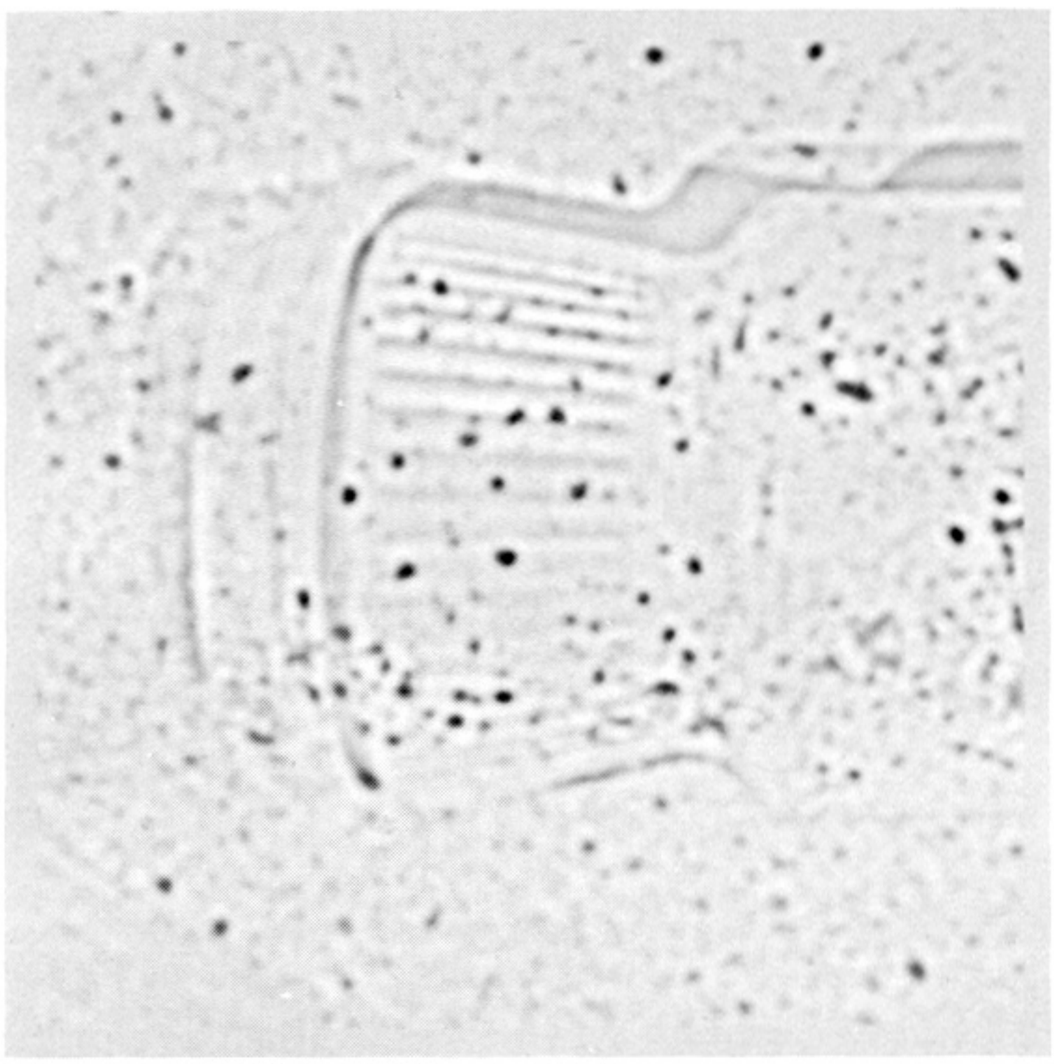
(*a*)



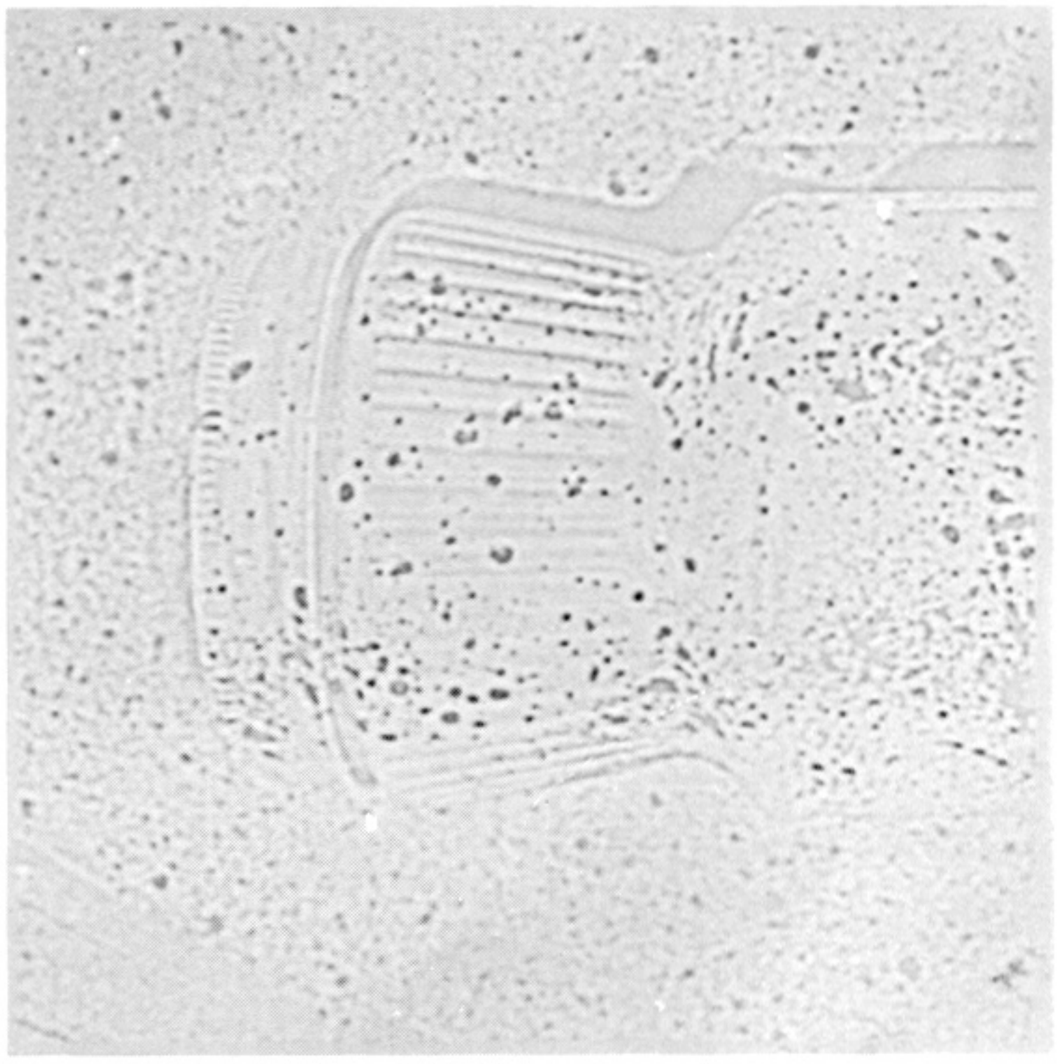
(*b*)



(*c*)



(*d*)

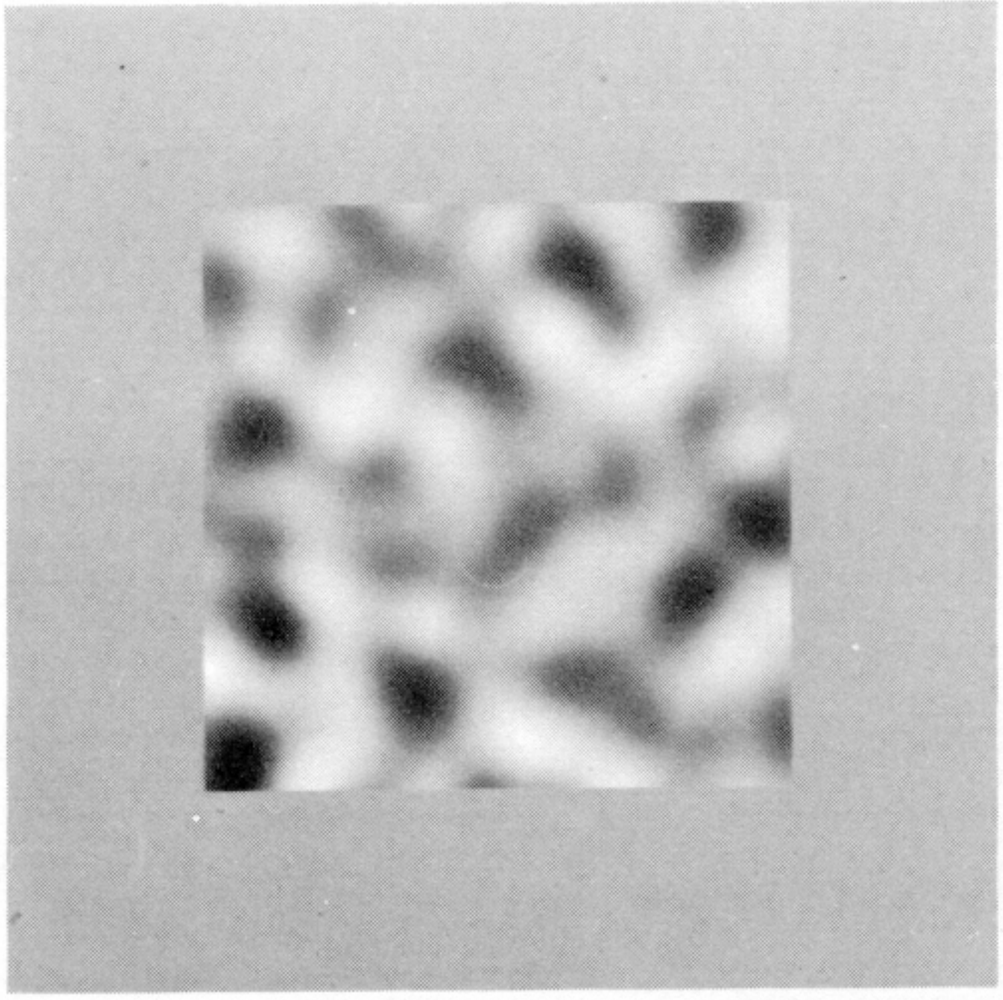


(*e*)

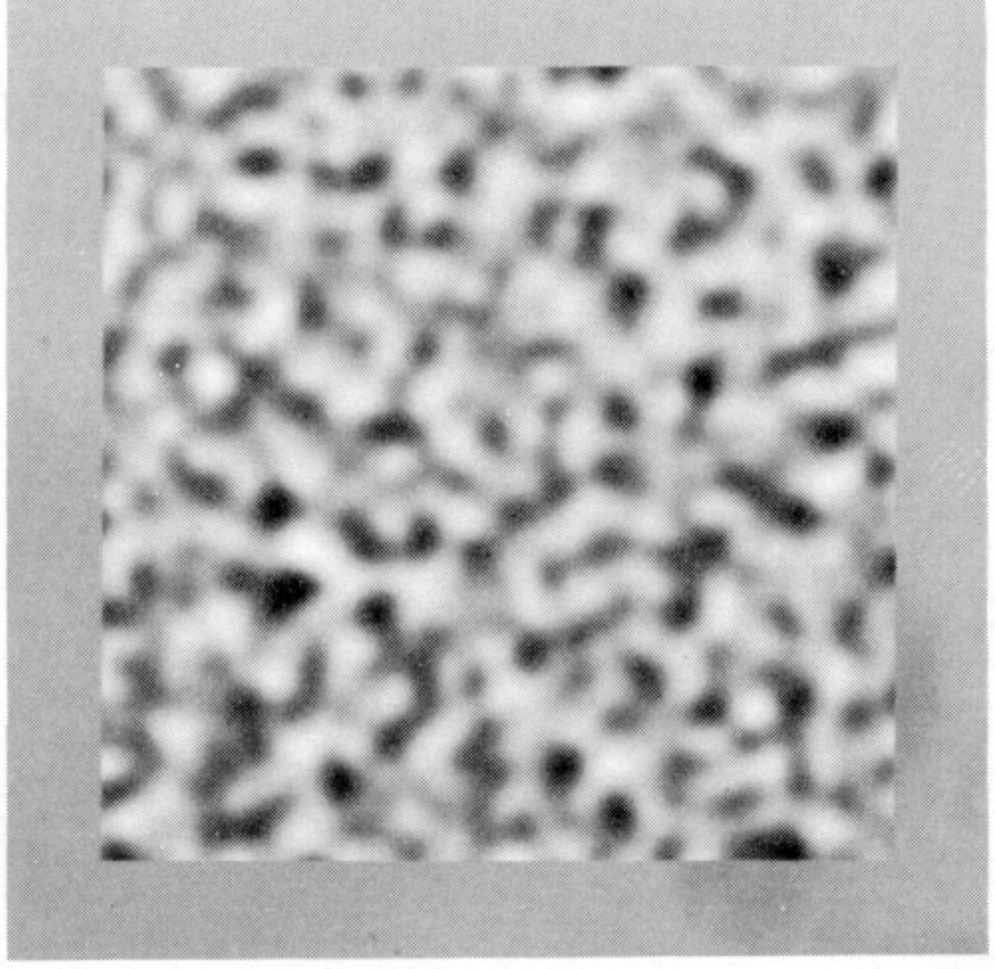
FIGURE 3. Examples of convolutions with $\nabla^2 G$. A natural image is indicated in (*a*). Below are examples of the convolved image, after application of different sized $\nabla^2 G$ operators, with central panel widths of (*b*) 36, (*c*) 18, (*d*) nine and (*e*) four picture elements. The original image was 480 picture elements on a side.

(a)

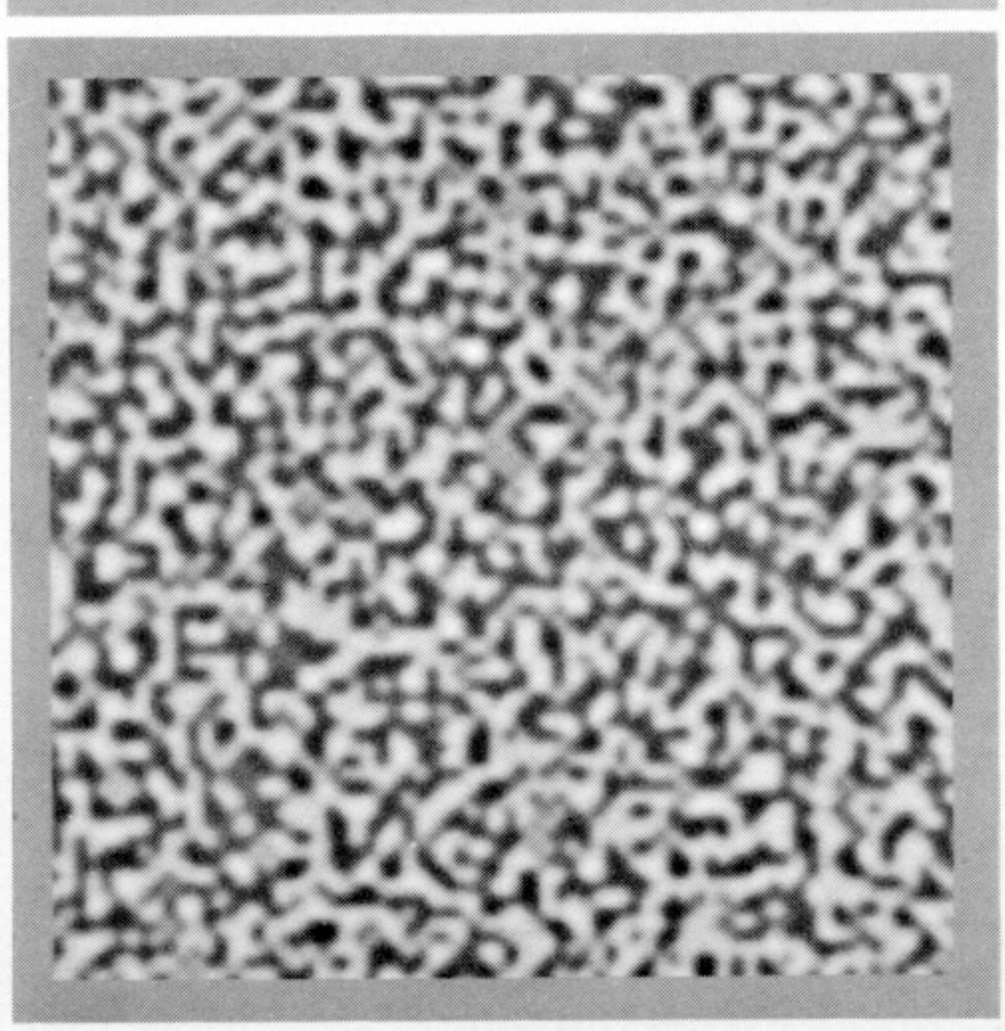
(b)



(c)



(d)



(e)

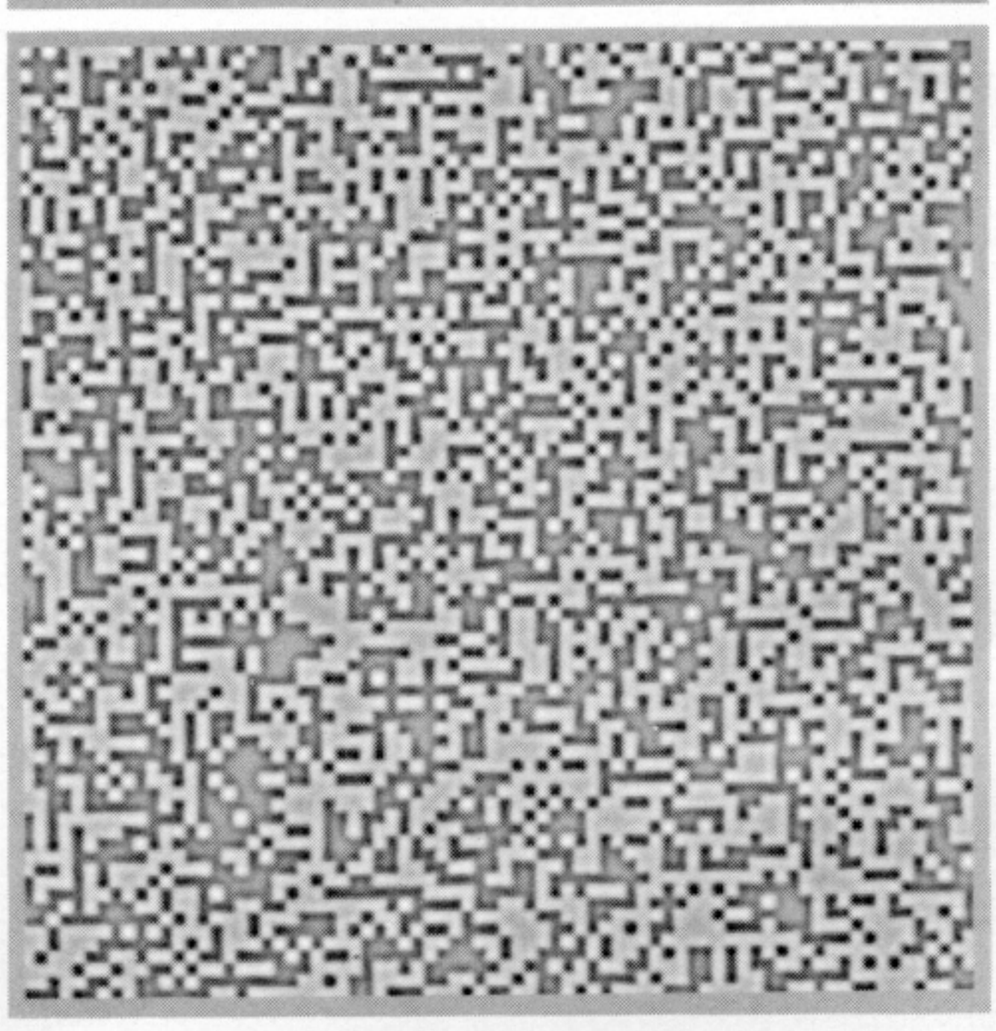
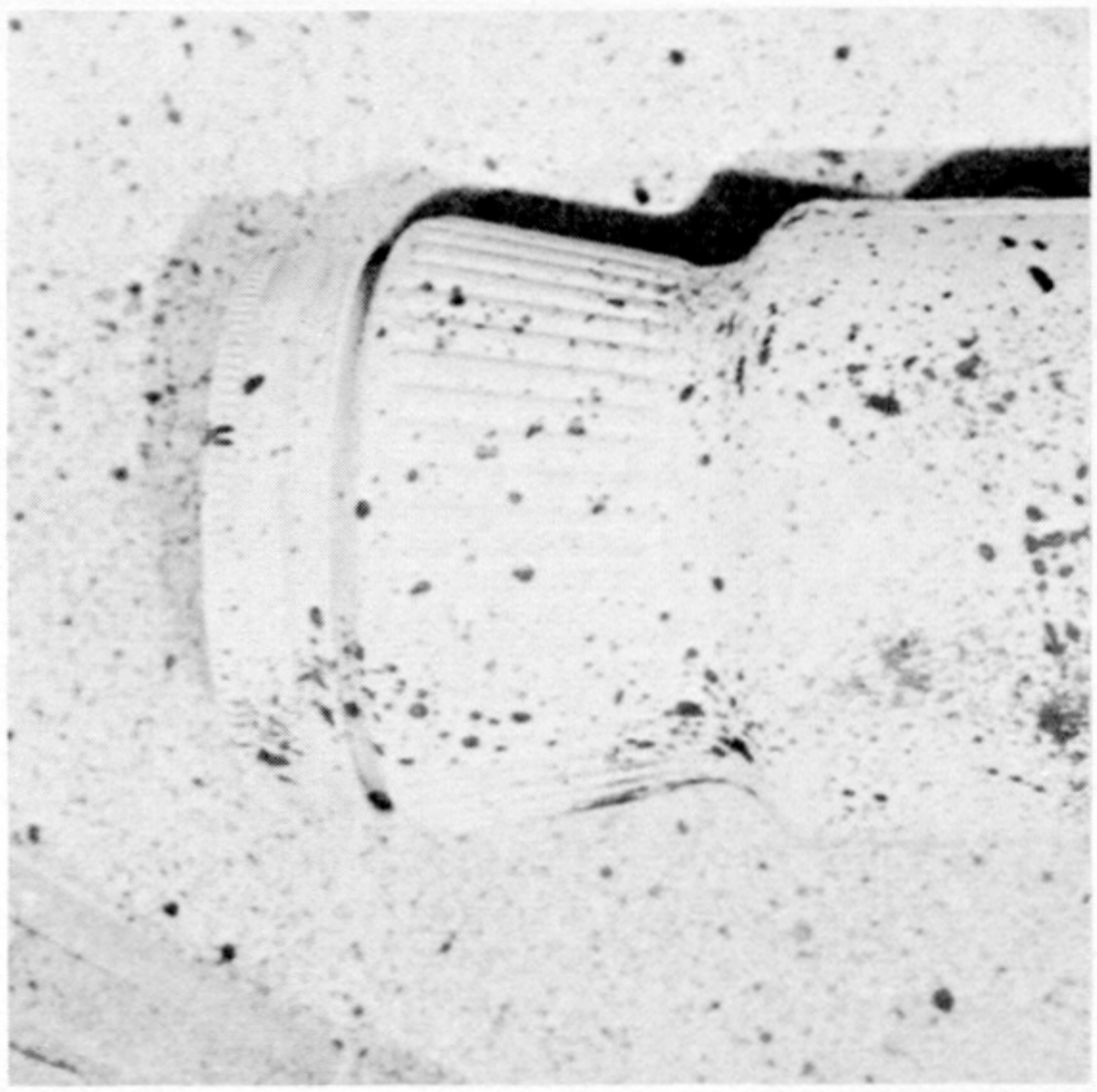
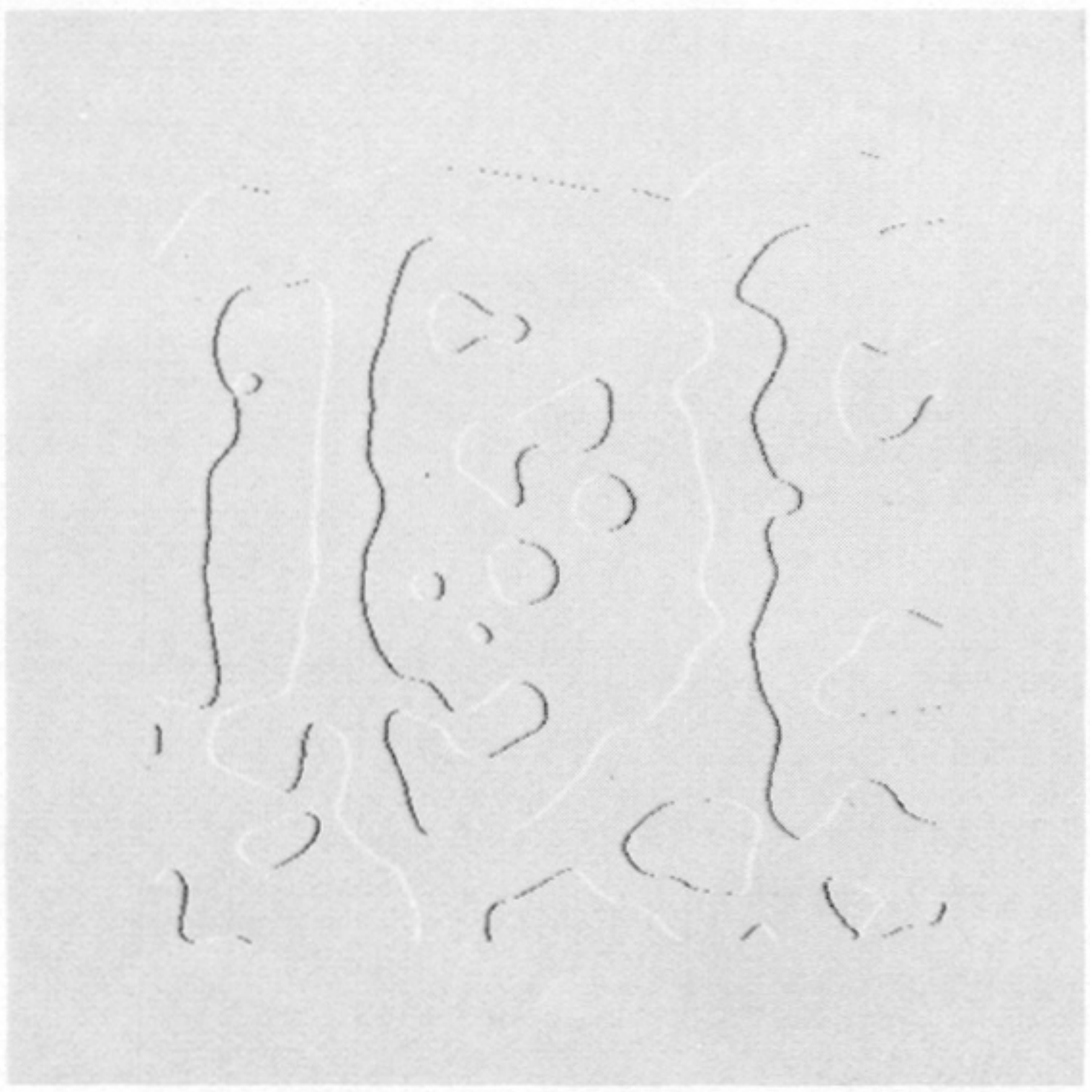


FIGURE 4. Examples of convolutions with $\nabla^2 G$. A random dot pattern is indicated in (*a*). Below are examples of the convolved image, after application of different sized $\nabla^2 G$ operators, with central panel widths of (*b*) 36, (*c*) 18, (*d*) nine and (*e*) four picture elements. The original image was 320 picture elements on a side.

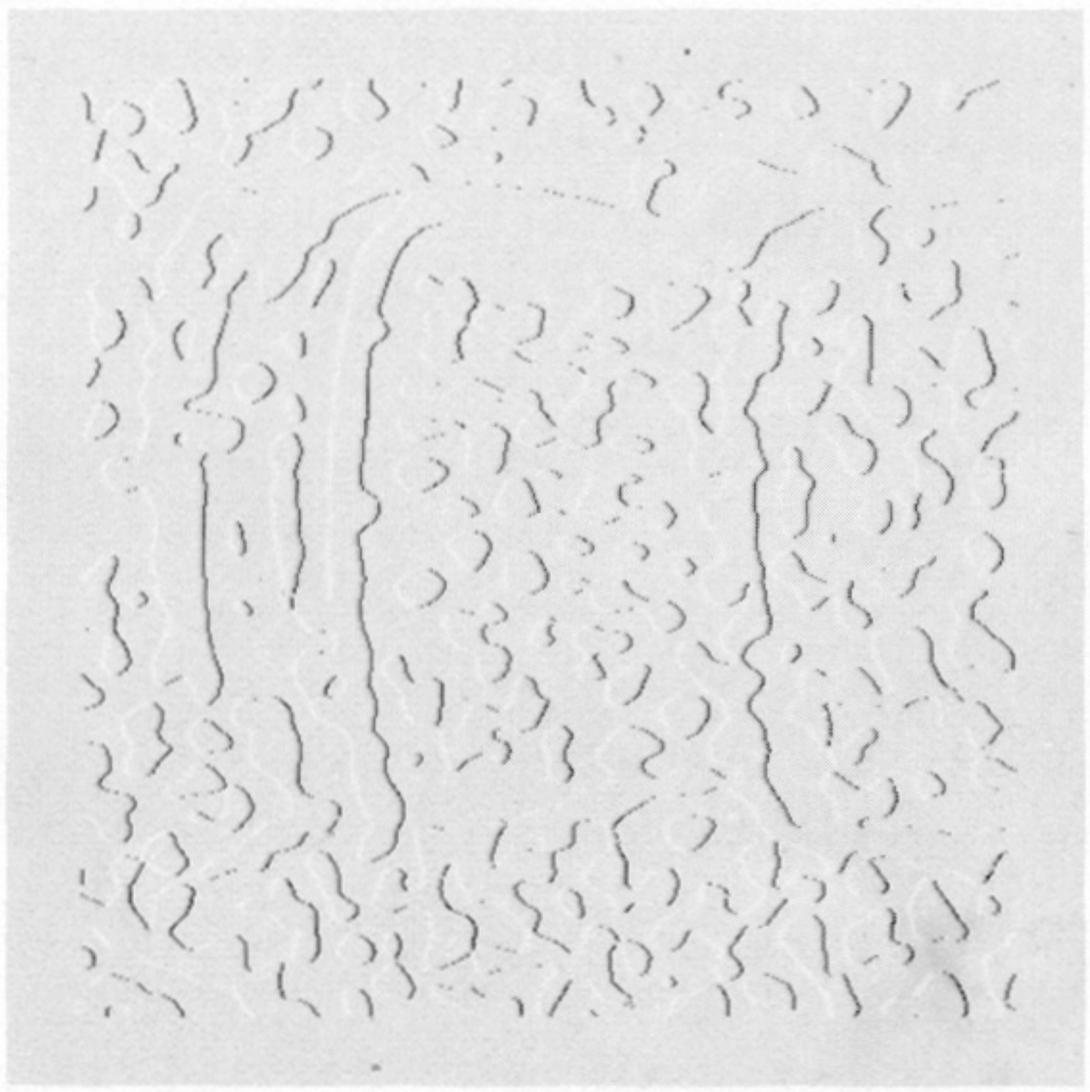
(*a*)



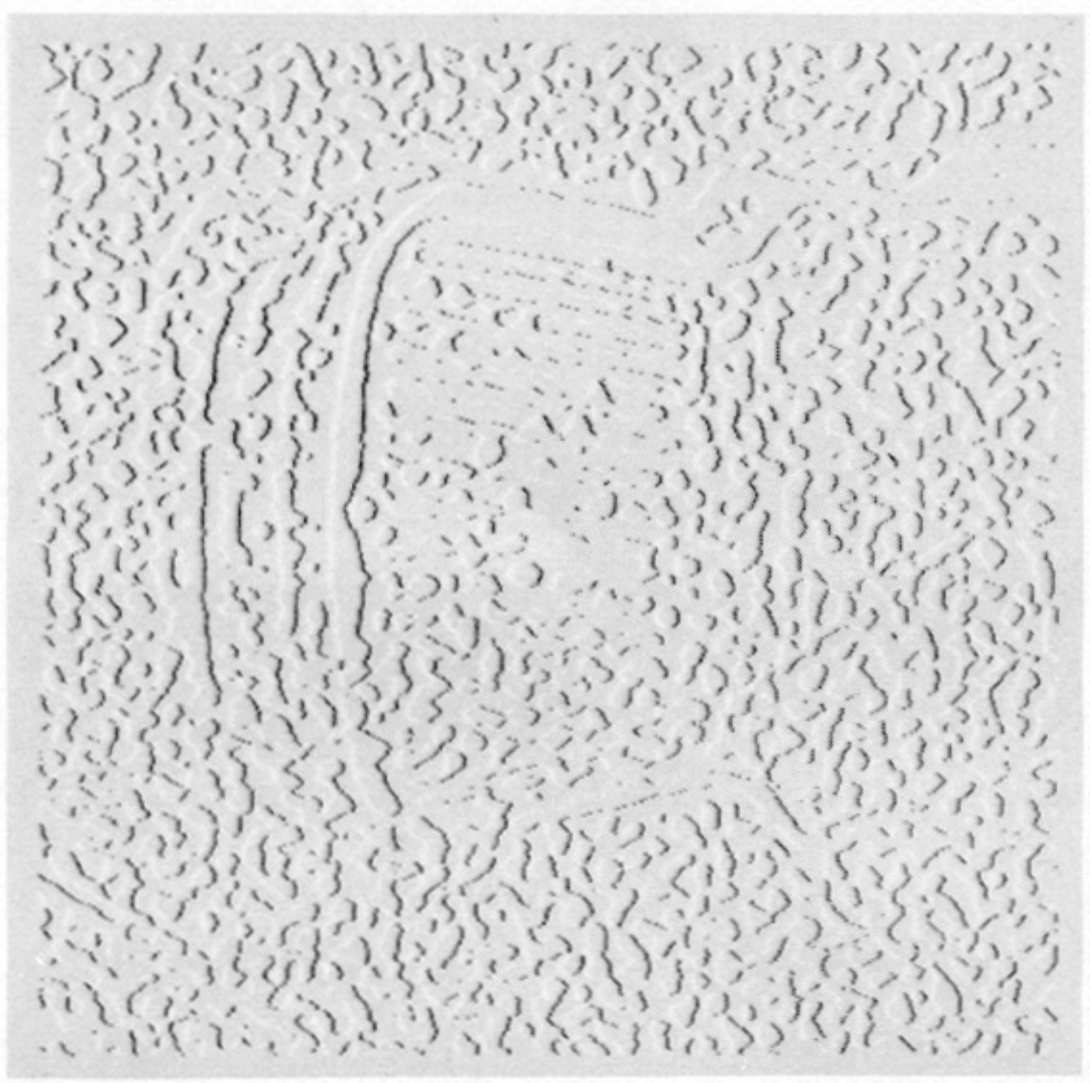
(*b*)



(*c*)



(*d*)

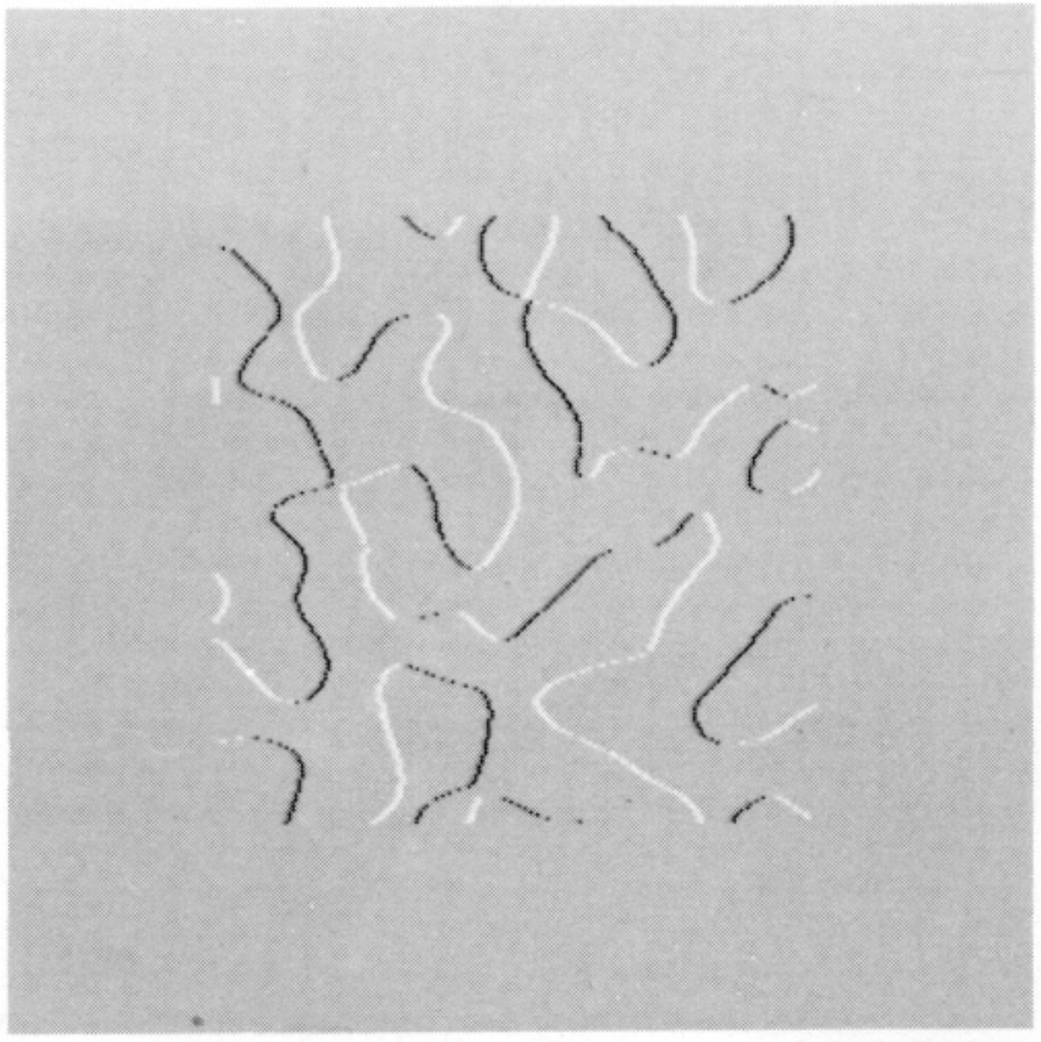


(*e*)

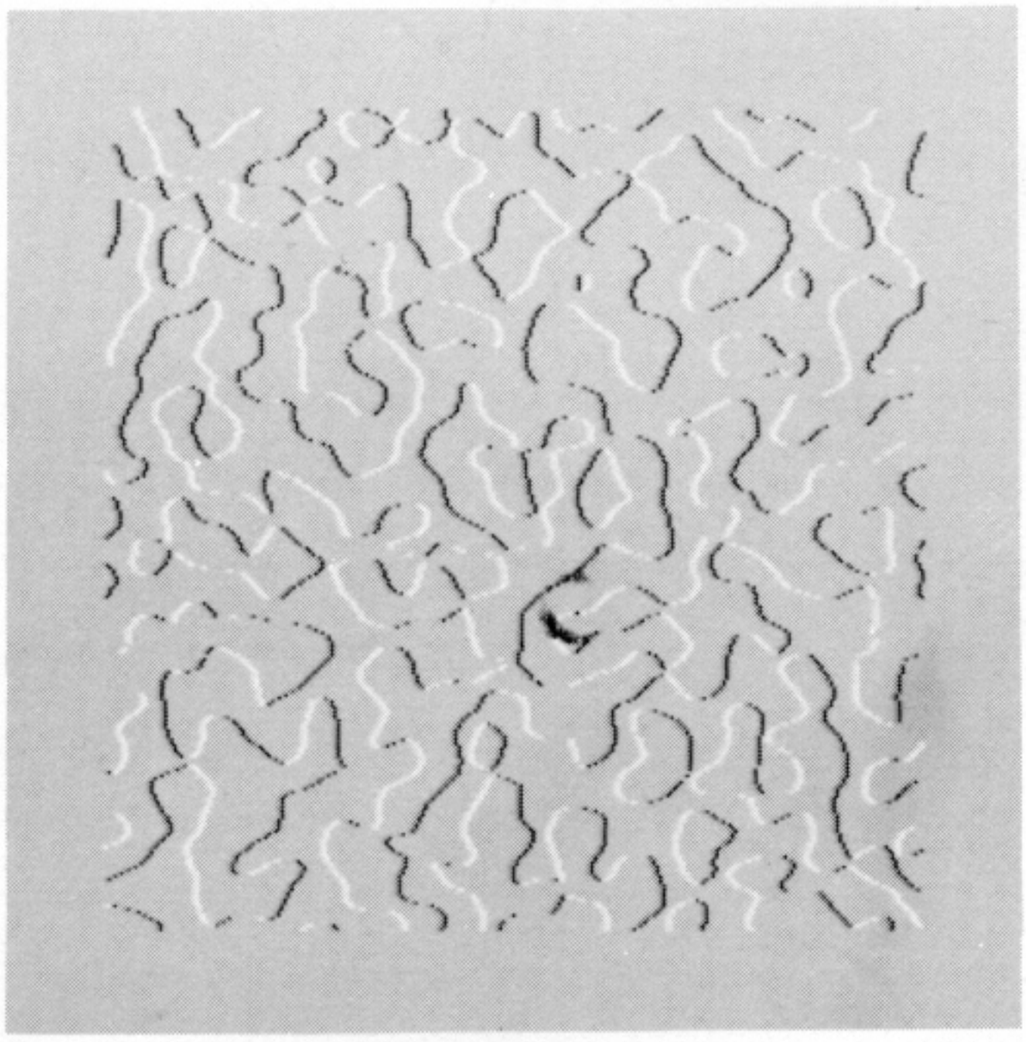
FIGURE 5. Examples of zero crossings. A natural image is indicated in (*a*). Below are examples of the zero crossings, obtained from different-sized $\nabla^2 G$ operators, with central panel widths of (*a*) 36, (*b*) 18, (*c*) nine and (*d*) four picture elements. The positive zero crossings are shown as white, the negative ones as black.

(a)

(b)



(c)



(d)

(e)

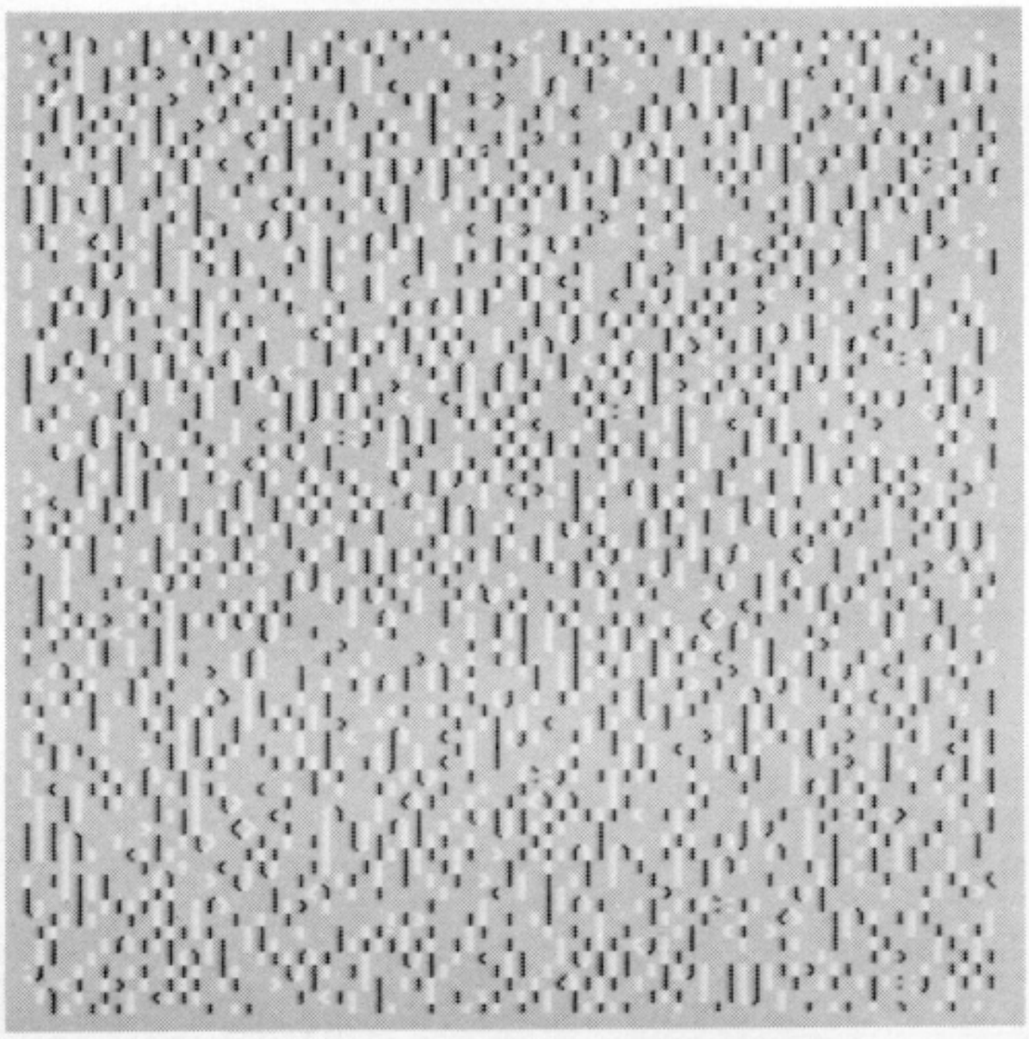


FIGURE 6. Examples of zero crossings. A random dot pattern is indicated in (*a*). Below are examples of the zero crossings, obtained from different-sized $\nabla^2 G$ operators, with central panel widths of (*b*) 36, (*c*) 18, (*d*) nine and (*e*) four picture elements. The positive zero crossings are shown as white, the negative ones as black.

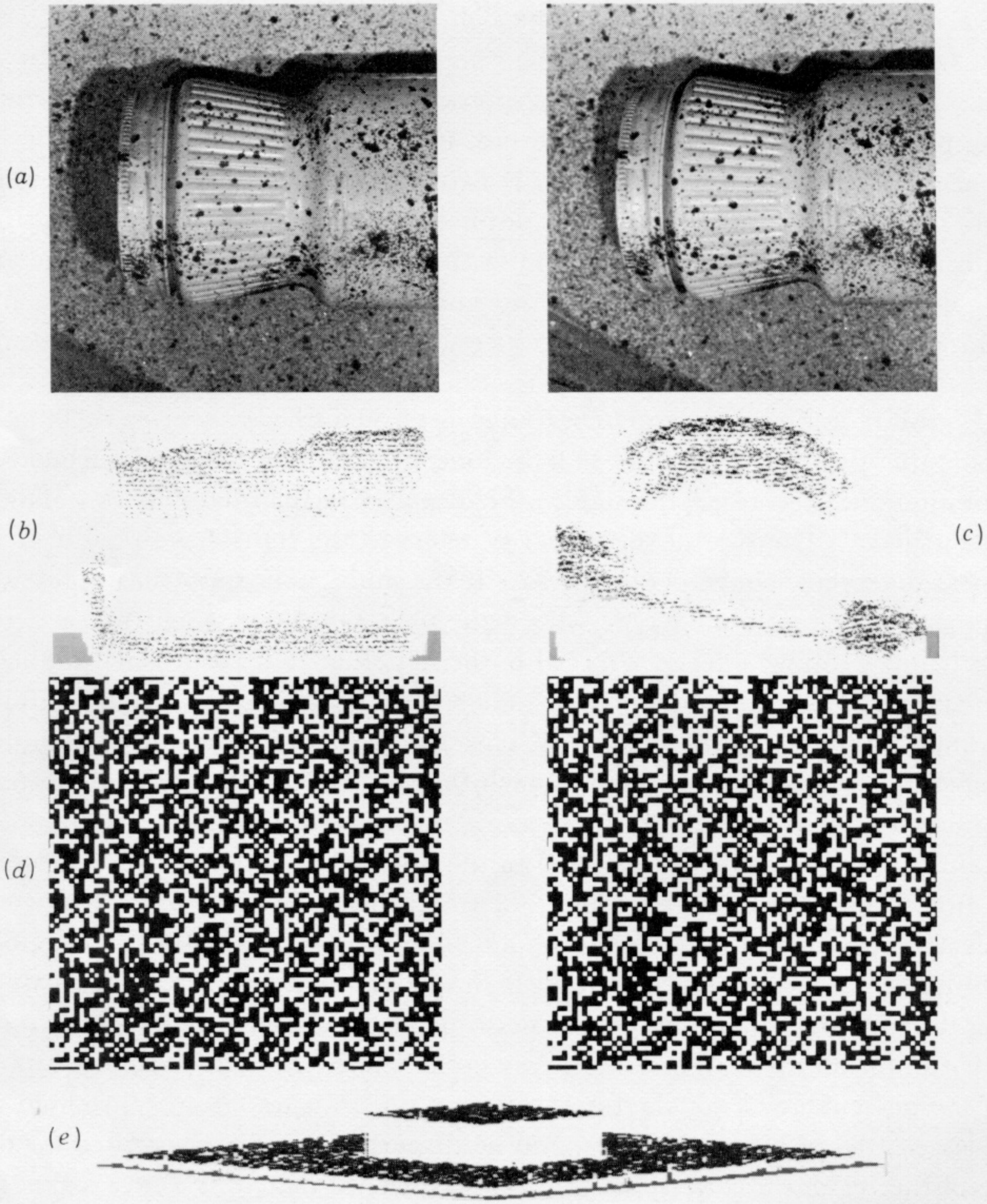
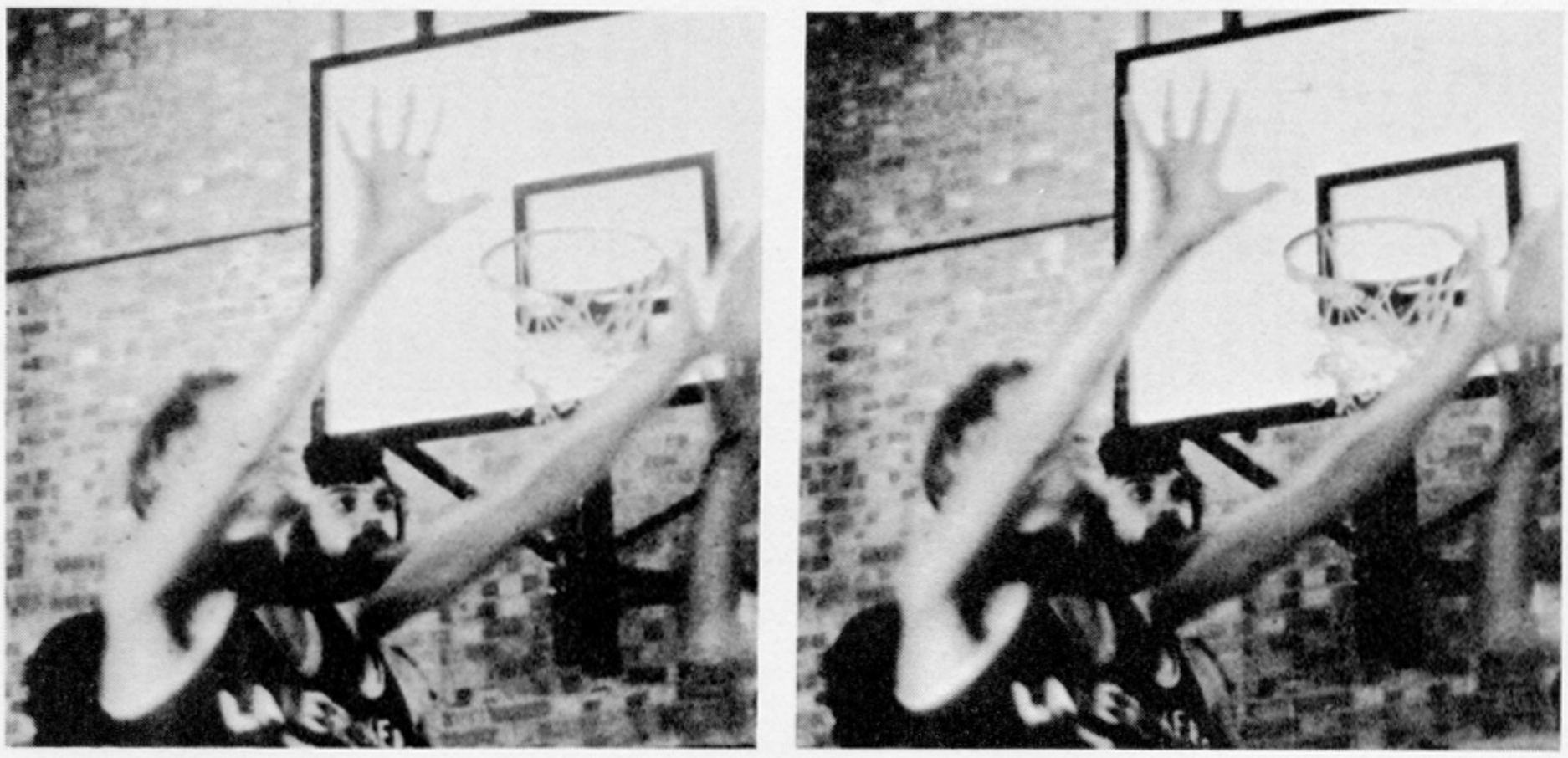
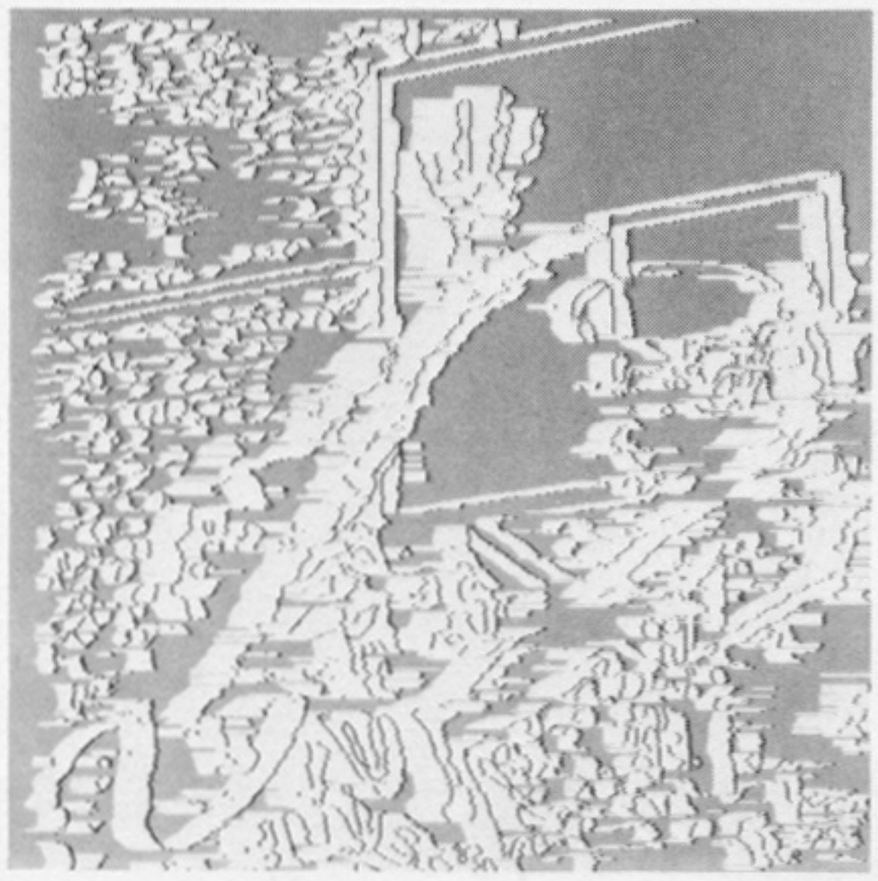


FIGURE 7. Results of the algorithm. The top stereo pair (*a*) is an image of a painted coffee jar. (*b*), (*c*). Two orthographic views of the disparity map. The disparities are displayed as $[x, y, c-ad(x,y)]$, where c is a constant and $d(x,y)$ is the difference in the location of a zero crossing in the right and left images. For purposes of illustration, a has been adjusted to enhance the features of the disparity map. The jar is viewed in (*b*) from the lower edge of the image and in (*c*) from the left edge of the image. Note that the background plane appears tilted in the disparity map. This agrees with the fused perception. The second stereo pair (*d*) is a 50% density random dot pattern. (*e*) The disparity map as viewed orthographically from some distance away. All disparity maps are those obtained from the $w = 4$ channel.

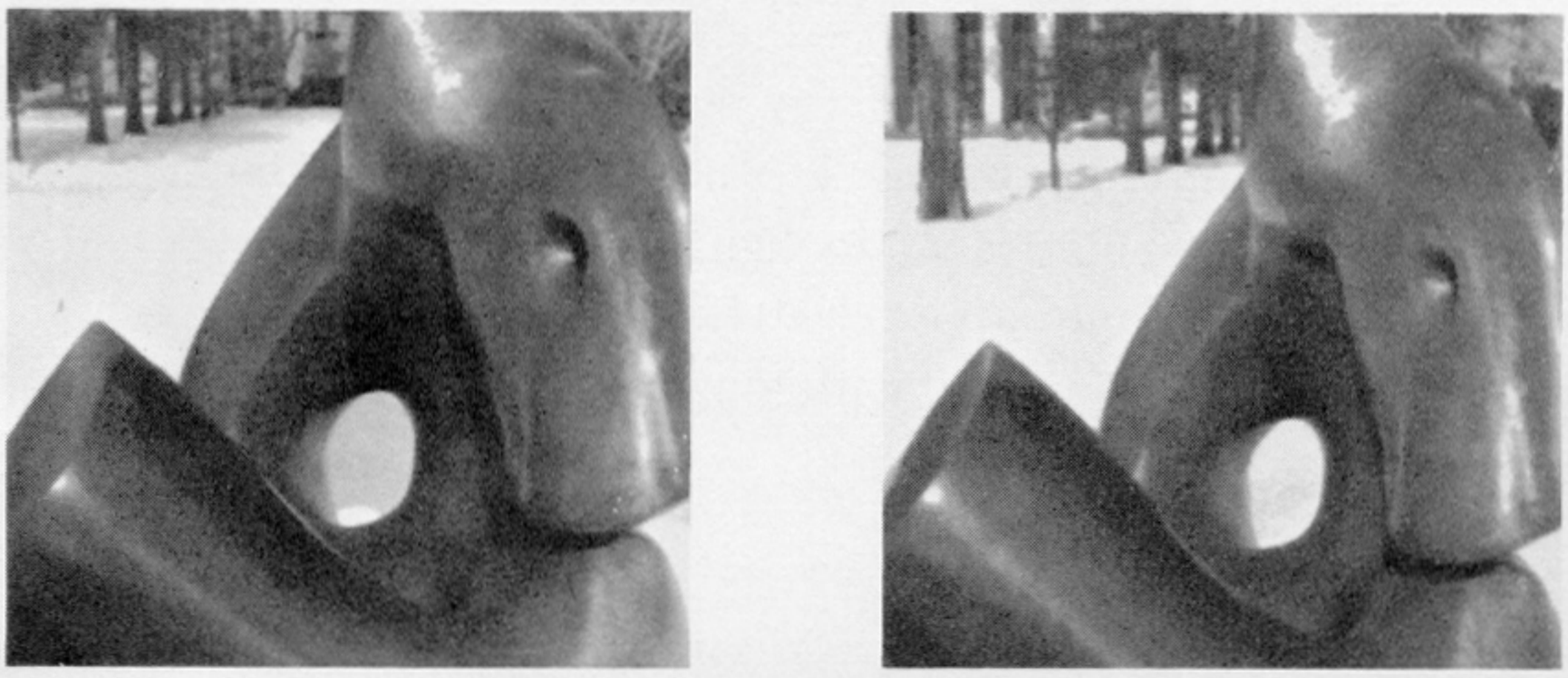
(a)



(b)



(c)



(d)

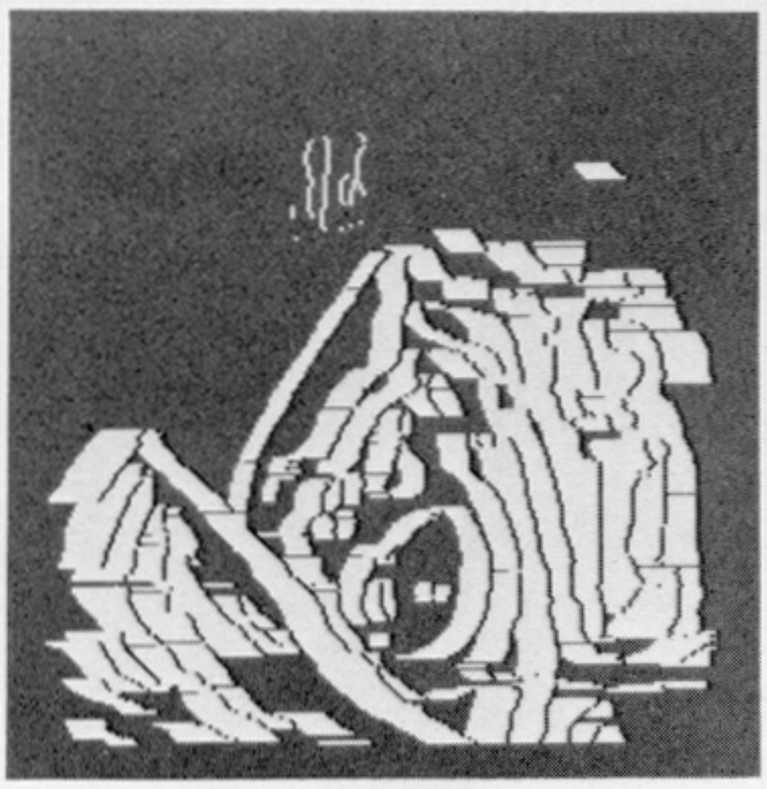


FIGURE 14. Examples of natural images. The top stereo pair (*a*) is a scene of a basketball game. The disparity map (*b*) is represented in such a manner that the width of the white bars, terminated by a black dot, corresponds to the disparity of the point. It can be seen that the disparity values are all qualitatively correct with the arm of the foremost player emerging from the background of the basket and the wall. The images were 480 pixels on a side. The bottom stereo pair (*c*) is a scene of sculpture by Henry Moore. The disparity array (*d*) is represented as in the top stereo pair. It can be seen that the disparity values obtained by the program roughly correspond to the shape of the surface. The images were 320 pixels on a side.

(*a*)

(*b*)

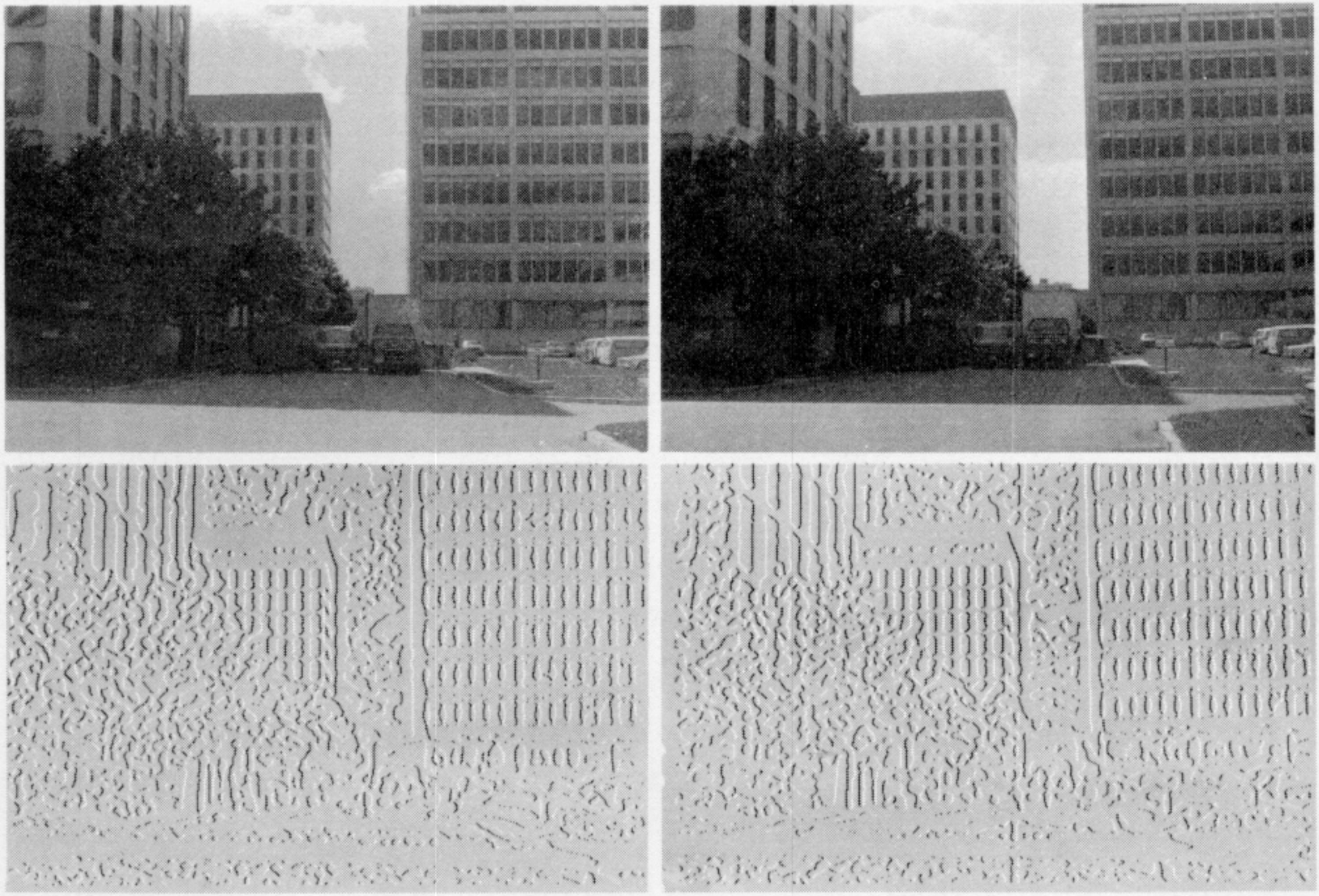
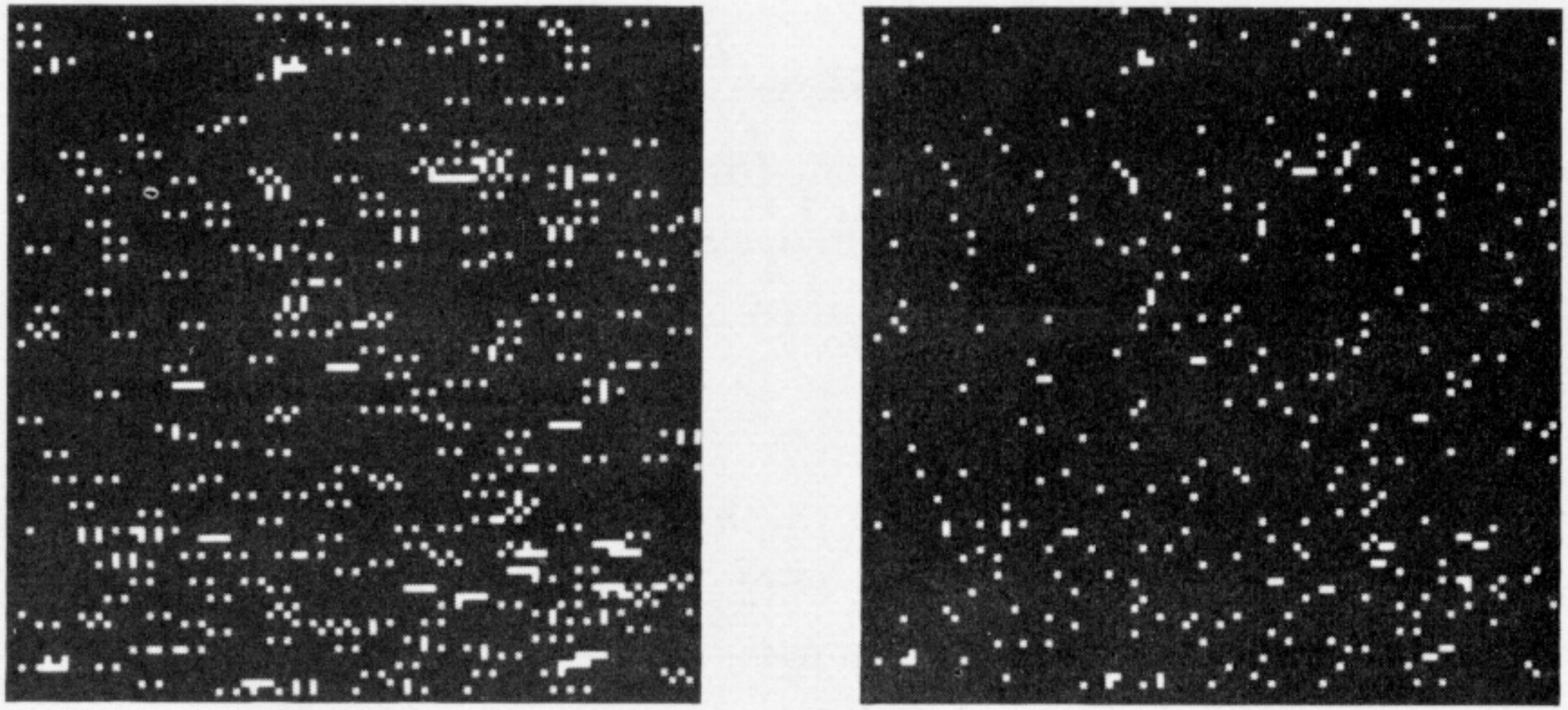


FIGURE 18. The false targets problem. (*a*) A stereo pair of a group of buildings. (*b*) The zero-crossing descriptions of these images. The regular pattern of the windows of the rear building causes difficulties for the matcher. If the alignment of the eyes corresponds to fixating at the level of the building, the algorithm matches the zero crossings corresponding to the windows correctly. If the alignment of the eyes corresponds to fixating at the level of the trees in front of the building, the algorithm matches the zero crossings corresponding to the windows incorrectly. Experiments indicate that under similar conditions humans have a similar perception.

(*a*)



(*b*)

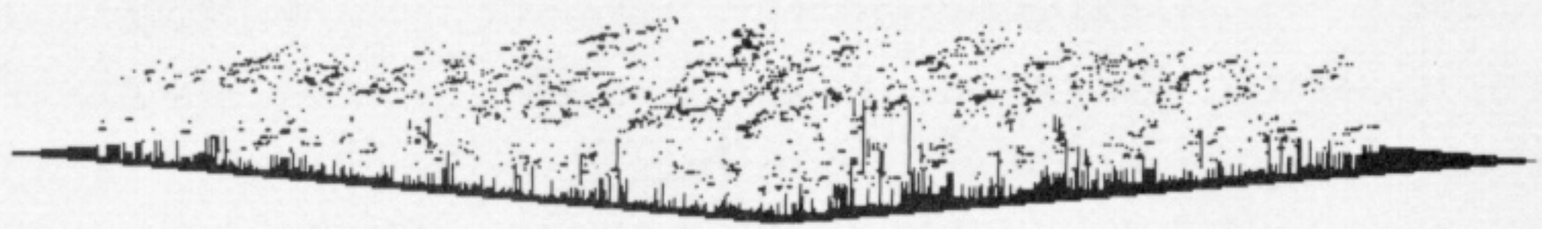
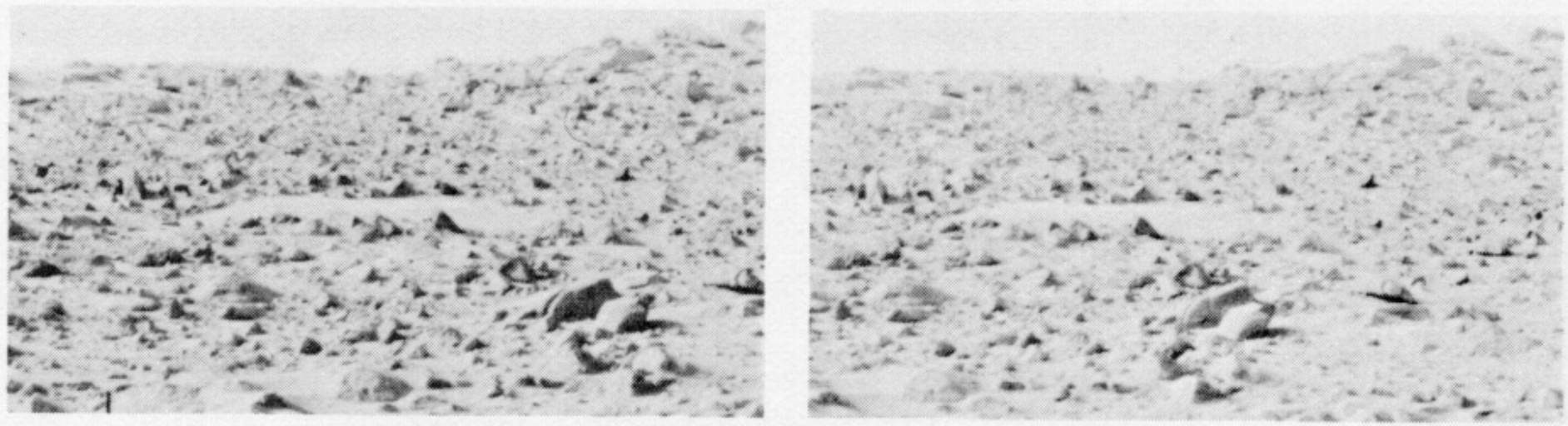


FIGURE 19. Panum's special case. The pair (*a*) is a special case of Panum's limit. The left image is formed by superimposing two slightly displaced copies of the right image. The disparity map (*b*) is shown below, and consists of two superimposed planes.

(*a*)



(*b*)

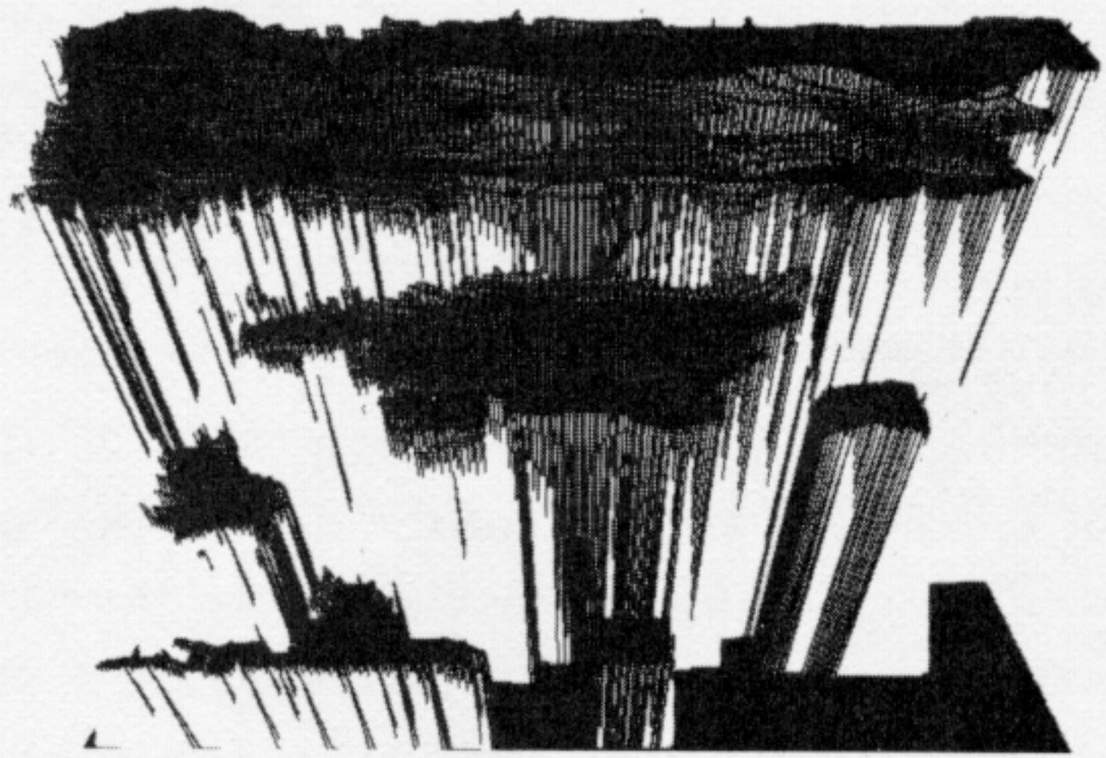


FIGURE 20. The interpolated Martian surface. The top pair of images (*a*) are a stereo pair of the Martian surface. (*b*) The disparity map, which has been interpolated between the known disparity points of the Marr–Poggio algorithm (see Grimson (1980) for details). The height of a point above the reference plane corresponds to the distance from the viewer to the point in the image. The total range of disparity in this image is roughly 200 pixels. Some sections of the foreground were not matched by the algorithm, and have not been interpolated. It is interesting to note that the disparity map contains a series of sharp breaks in disparity, corresponding to occluding hills in the image. These breaks are not evident in the monocular images, yet are clearly visible when the two images are fused.