MASSACHUSETTS INSTITUTE OF TECHNOLOGY

ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 565

January, 1980

A COMPUTER IMPLEMENTATION OF A THEORY OF HUMAN STEREO VISION

W.E.L. Grimson

ABSTRACT: Recently, Marr and Poggio (1979) presented a theory of human stereo vision. An implementation of that theory is presented, and consists of five steps: (1) 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. (2) Zero-crossings in the filtered images are found along horizontal scan lines. (3) 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, Marr and Poggio showed that false targets pose only a simple problem. (4) 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. (5) When a correspondence is achieved, it is stored in a dynamic buffer, called the $2\frac{1}{2}$-dimensional sketch. To support the sufficiency 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. As well, statistical assumptions made by Marr and 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.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research constract N00014-75-C-0643 and in part by National Science Foundation Grant MCS77-07569.

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. This difference in relative positions — the disparity — may be measured and used to estimate depth. The process of stereo vision, in essence, measures this disparity and uses it to compute depth information for surfaces in the scene.

The steps involved in measuring disparity are (Marr and Poggio, 1979): (S1) a particular location on a surface in the scene must be selected from one image; (S2) that same location must be identified in the other image; and (S3) the disparity between the two corresponding image points must be measured. The difficulty of the problem lies in steps (S1) and (S2), that is, in matching the images of the same location — the so-called correspondence problem. For the case of the human stereo system, it can be shown that this matching takes place very early in the analysis of an image, prior to any recognition of what is being viewed, using 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 and 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: (1) 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. (2) 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. (3) 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 and Poggio showed that false targets pose only a simple problem, because of the roughly bandpass nature of the filters. (4) 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. (5) When a correspondence is achieved, it is stored in a dynamic buffer, called the $2\frac{1}{2}$-dimensional sketch (Marr and 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, of course, 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 (c.f. the cooperative algorithm of Marr and Poggio (1976)).

This article describes an implementation of the Marr-Poggio stereo theory, written with particular emphasis on the matching process (Grimson and Marr, 1979). For details of the derivation and justification of the theory, see Marr and 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 defocussed, 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. Thirdly, the theory makes some statistical assumptions; these are compared with the actual statistics found in practice. Next, some points about the theory that were clarified as a result of writing the program are discussed. Finally, the results of running the program on some natural images are shown.

2. Design of the program

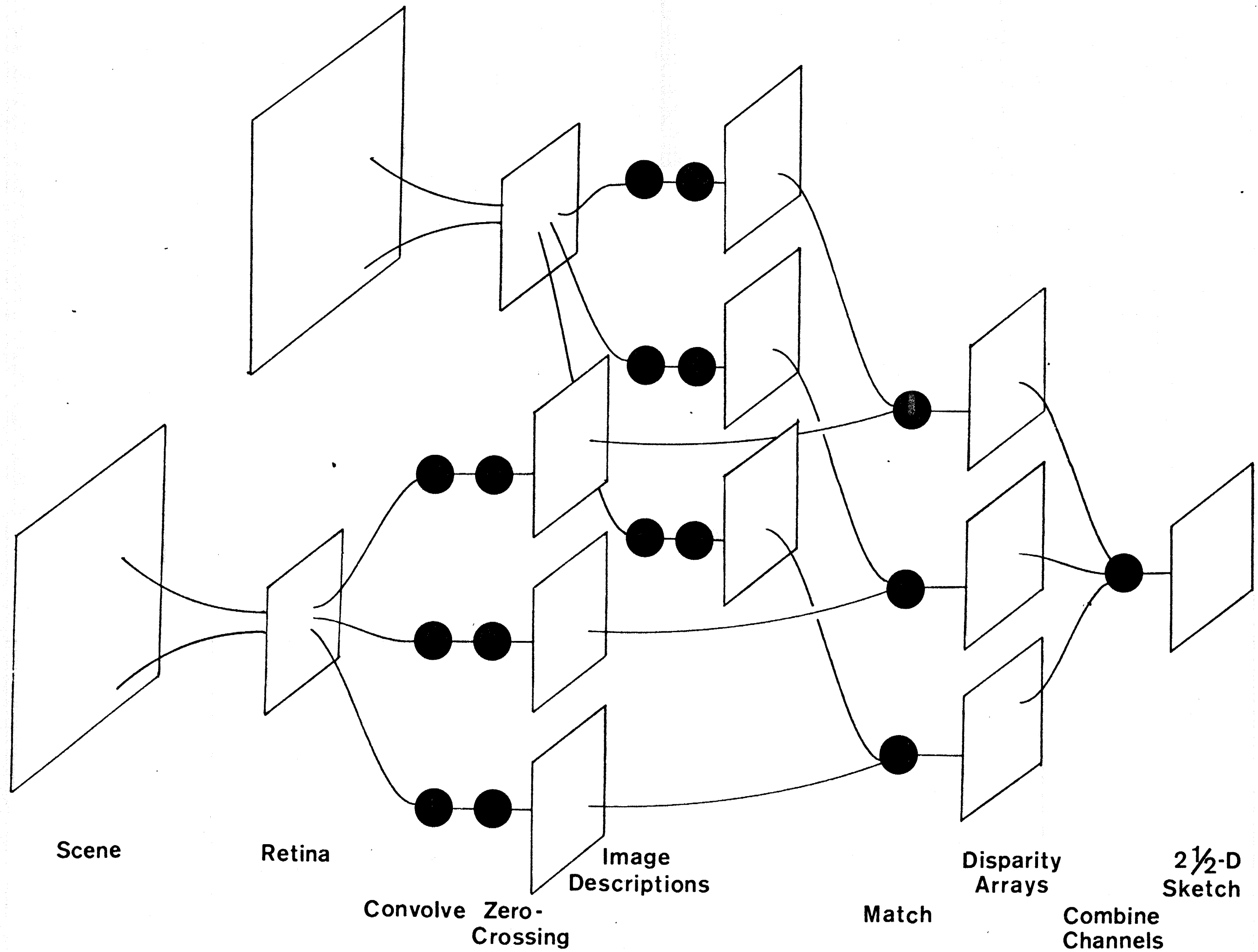
The implementation is divided into five modules, roughly corresponding to the five steps in the summary above. These modules, and the flow of information between them, are illustrated in Figure 1. Each of the components is described in turn.

2.1 Input

There are two aspects of the human stereo system, embedded in the Marr-Poggio theory, which 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 images, representing the entire scene visible to the viewer. This pair of images corresponds to the environment around the

Figure 1. Diagram of the algorithm. The images of the scene are mapped into the images of the retinas, taking into account the eye positions. Each image is convolved with a set of different sized masks and zero-crossings are located for each convolution. For each size mask, the left and right zero-crossing descriptions are matched. These matched descriptions are combined into a single representation. As well, the matches from the larger channels can drive eye vergence movements, causing new retinal images to be created and allowing the smaller channels to come into correspondence.



Scene
Retina
Convolve
Zero-Crossing
Image Descriptions
Match
Disparity Arrays
Combine Channels
2½-D Sketch

visual system, rather than some integral part of the system itself. To create this representation of the scene, natural images were digitized on an Optronix Photoscan System P1000. The sizes of these images are indicated in the legends. Grey-level resolution is 8 bits, providing 256 intensity levels. For the random dot patterns illustrated in this article, the images were constructed by computer, rather than digitized from a photograph.

For a given position of the eyes, relative to the scene, a representation of the images on the two retinas is extracted. The algorithm creates this retinal representation by obtaining a second, smaller pair of images from the images representing the whole scene. The mapping from the scene images into the retinal images accounts for the two factors inherent in the Marr-Poggio theory. First, different sections of the scenes will be mapped to the center (fovea) of the retinal images as the positions of the eyes are varied. Since the matching process will take place on the array 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 which 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. Second, the Marr-Poggio theory also states that the resolution of the earlier stages of the algorithm — the convolution and zero-crossings — scales linearly with eccentricity. The most convenient method for dealing with this fact is to account for the scaling with eccentricity at the level of the extraction of the images. This means that rather than extracting a set of retinal images in a linear manner, we may map the scene into the retinal images by a mapping whose magnification varies with eccentricity. By so doing, the later stages of processing need not explicitly account for the variation with eccentricity. Rather, these processes are considered as operating on a uniform grid. Note that this eccentric mapping is not essential, especially for small images. In most of the cases illustrated in this article, the mapping was not used.

After the completion of this stage, the implementation has created a representation of the images that has accounted for eye position and 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 then necessary to transform them into a form upon which the matcher may operate. Marr and Poggio (1979) argued that the items to be matched in an image must be in one-to-one correspondence with well-defined locations on a physical surface. This led to the use of image predicates which correspond to changes in intensity. Since these intensity changes can occur over a wide range of scales within a natural image, they are detected separately at different scales. This is in agreement with the findings of Campbell and Robson (1968), who showed that visual information is processed in parallel by a number of independent spatial-frequency-tuned channels, and with the findings of Julesz and Miller (1975) and Mayhew and Frisby (1976), who showed that spatial-frequency-tuned channels are used in stereopsis and are independent. Recent work by Wilson and Bergen (1979) and Wilson and Giese (1977) provided evidence for the particular form of these spatial-frequency-tuned operators. Measuring contrast sensitivity to vertical line stimuli, Wilson and his collaborators showed that the image is convolved with an operator which in one dimension may be closely approximated by a difference of two gaussian functions (DOG).

In the original theory (Marr and Poggio, 1979), the proposed masks were oriented bar masks whose cross-section was a difference of two gaussians, as given by the Wilson and Bergen data. 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. However, a number of practical considerations have led Marr and Hildreth (1979) to suggest that the initial operators not be directional operators. The only non-directional linear second derivative operator is the Laplacian. Marr and Hildreth have shown that provided 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. Therefore, theoretically, we can detect intensity changes occuring at all orientations using the single non-oriented Laplacian operator. Thus, Marr and Hildreth propose that intensity changes occuring 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. The operator, together with its fourier transform, is illustrated in Figure 2. The form of the operator is given by:

$$\nabla^2 G(r,\theta) = \left[2 - \frac{r^2}{\sigma^2}\right] exp\{\frac{-r^2}{2\sigma^2}\}.$$

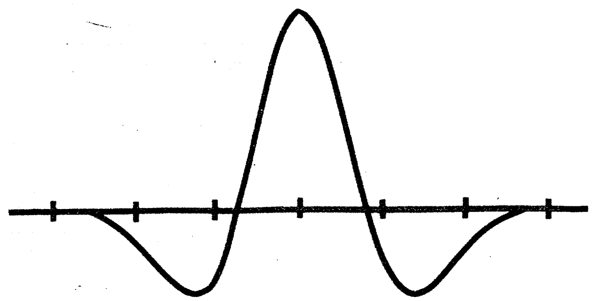
Given the form of the operators, it is only left to determine the size of these masks. To do this, we first note that Marr and Hildreth (1979) showed that the operator $\nabla^2 G$ is a close approximation to the DOG function. Wilson and Bergen's data indicated DOG filters whose sizes — specified by the width w of the filter's central excitatory 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:

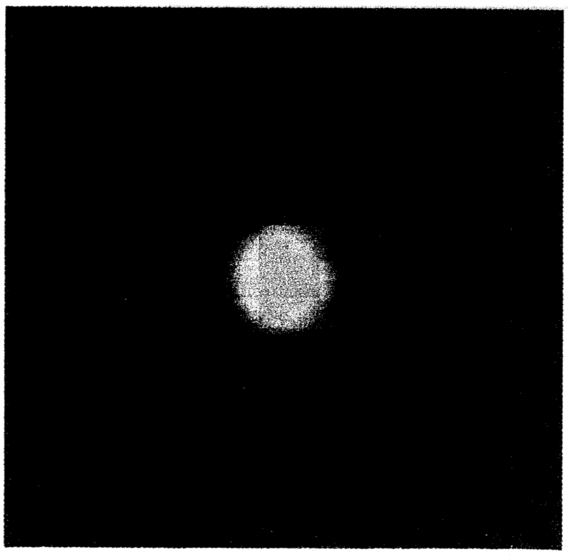
$$\sigma = \frac{w}{2\sqrt{2}}.$$

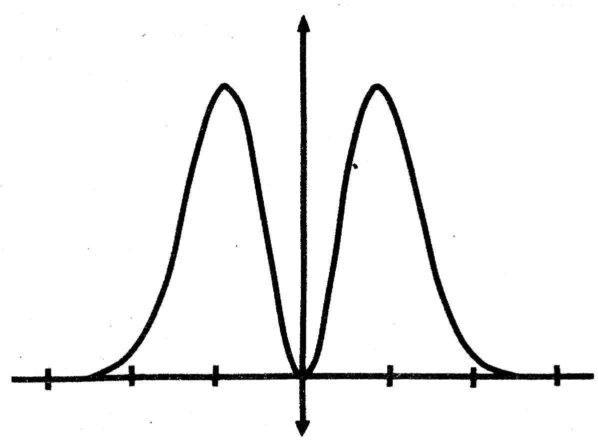
Wilson and Bergen's values were obtained by using oriented line stimuli. To obtain the diameter of the corresponding circularly symmetric center-surround receptive field, the values of w must be multiplied by $\sqrt{2}$. Finally, we want the resolution of the initial images to roughly represent the resolution 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 center-to-center spacing of the cones is 2.0 to 2.3 μm, corresponding to an angular spacing of 25 to 29 arc seconds (O'Brien, 1951). Accounting for the conversion of Wilson and Bergen's data, and using the figure of 27 seconds of arc for the separation of cones in the fovea, one arrives at values of w in the range 9 to 63 image elements, and hence, values of σ in the range 3 to 23 image elements.

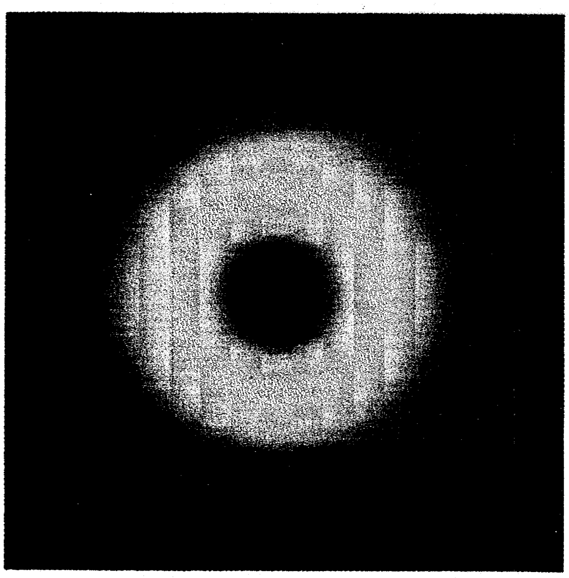
Recently, it has been proposed (Marr, Poggio and Hildreth, 1979) that a further, smaller channel may be present. This channel would have a central excitatory width of $w = 1.5'$, roughly corresponding to 4 image elements.

Figure 2. The operators G'' and $\nabla^2 G$. The top left figure show G'', the second derivative of a one-dimensional guassian distribution. The top right figure shows $\nabla^2 G$, its rotationally symmetric two-dimensional counterpart. The bottom figures show their Fourier transforms.









The present implementation uses four filters, each of which is a radially symmetric difference of gaussians, with w values of 4, 9, 17 and 35 image elements. The coefficients of the filters were represented to a precision of 1 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, 0.68σ.

The actual convolutions were performed on a LISP machine constructed at the MIT Artificial Intelligence Laboratory, using additional hardware specially designed for the purpose (Knight, et al. 1979). Figures 3 and 4 illustrates 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 size mask.

2.3 Detection and description of zero-crossings

According to the Marr-Poggio theory, 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; as a consequence, only zero-crossings are used as input to the matcher.

Since, for the purpose of obtaining disparity information, we may ignore horizontally oriented segments, 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 opposite sign. This gives the position of zero-crossings to within an image element.

In addition to their location, we record the sign of the zero-crossings (whether convolution values change from positive to negative or negative to positive as we move from left to right) and a rough estimate of the local, two-dimensional orientation of pieces of the zero-crossing contour. In the present implementation, the orientation at a point on a zero-crossing segment is computed as the direction of the gradient of the convolu-

Figure 3. Examples of convolutions with $\nabla^2 G$. The top figure shows a natural image. The bottom figures show the convolution of this image with a set of $\nabla^2 G$ operators. The sizes of these operators are $w = 36, 18, 9$ and 4 image elements.

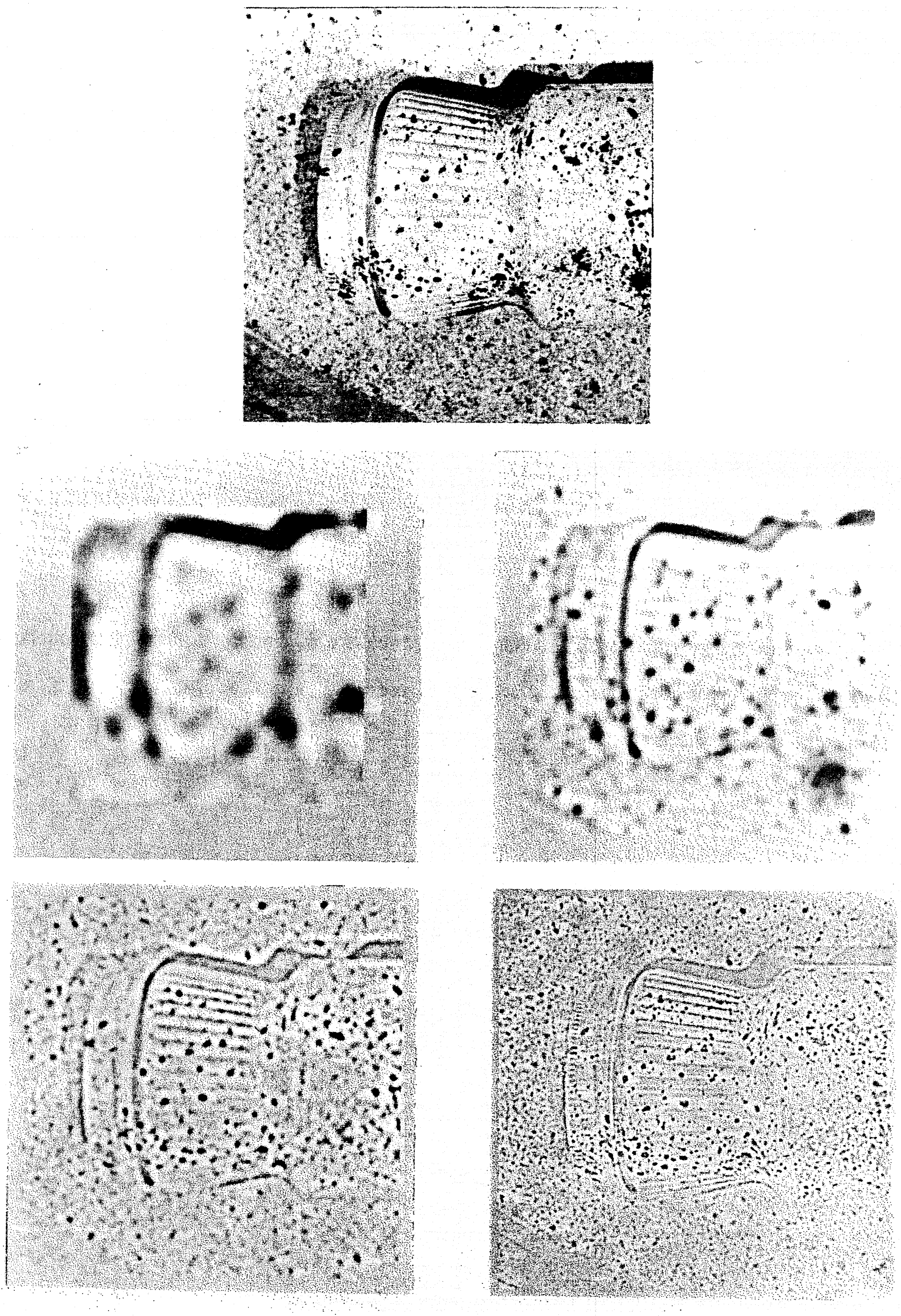
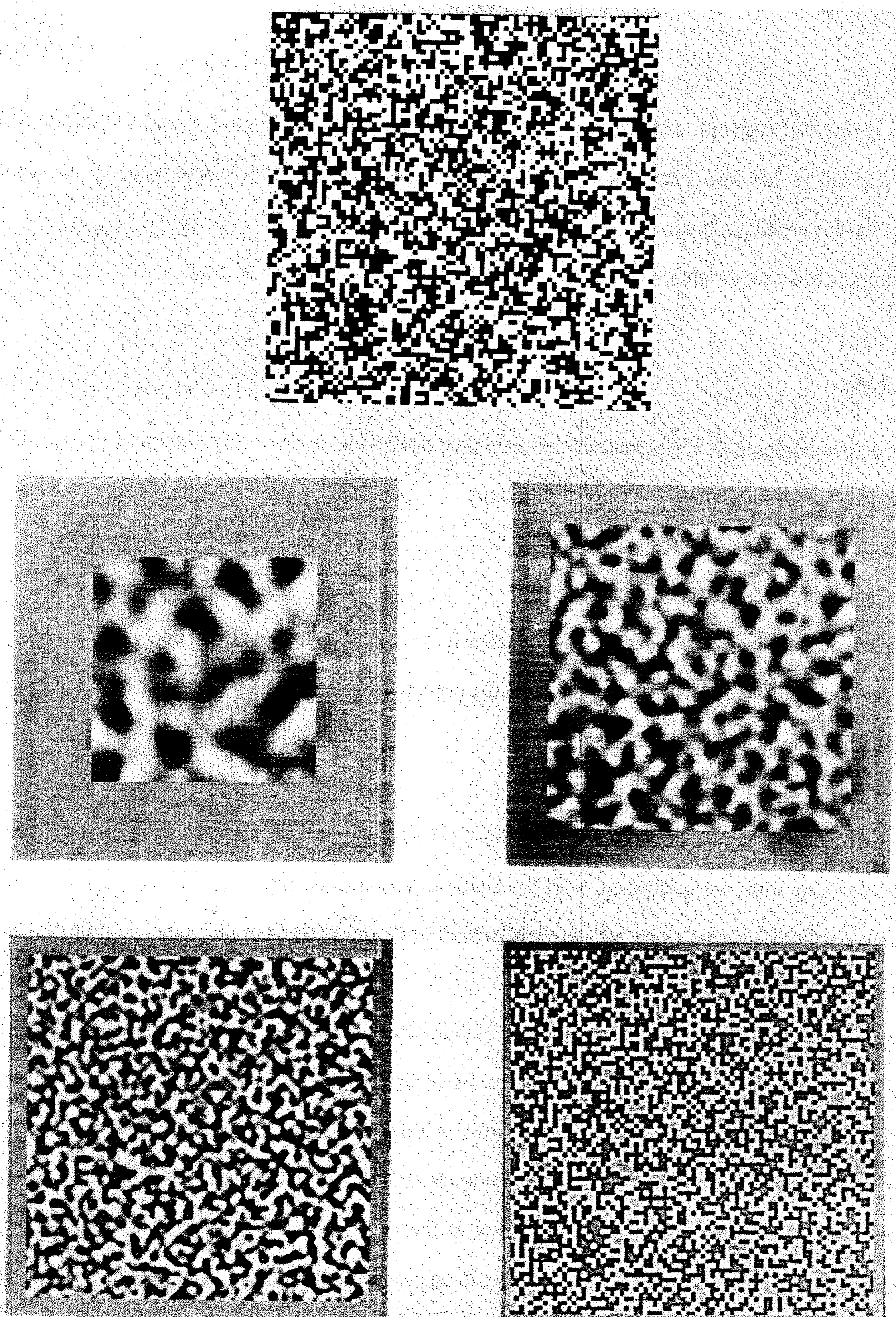


Figure 4. Examples of convolutions with $\nabla^2 G$. The top figure shows a random dot pattern. The bottom figures show the convolution of this image with a set of $\nabla^2 G$ operators. The sizes of these operators are $w = 36, 18, 9$ and 4 image elements.



tion values across that segment, and recorded in increments of 30 degrees. 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, black.

We compute this zero-crossing description for each image and for each size of mask.

2.4 Matching

The matcher implements the second of the matching algorithms described by Marr and Poggio (1979, p.315). For each size of filter, matching consists of 6 steps:

(1) Fix the eye positions.

(2) Locate a zero-crossing in one image.

(3) Divide the region about the corresponding point in the second image into three pools.

(4) Assign a match to the zero-crossing based on the potential matches within the pools.

(5) Disambiguate any ambiguous matches.

(6) Assign the disparity values to a buffer.

These steps may be repeated several times during the fusion of an image. Given a position for the optic axes, 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.

We now expand upon each of the six steps of the matching process. The first step consists of fixing the two eye positions. The alignment between the two zero-crossing descriptions, corresponding to the positions of the optical axes, 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 this 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 optical axes.

Figure 5. Examples of zero-crossing descriptions. The top figure show a natural image. The bottom figures show the zero-crossings obtained from the convolutions of Figure 3. The white lines mark positive zero-crossings and the black lines, negative ones.

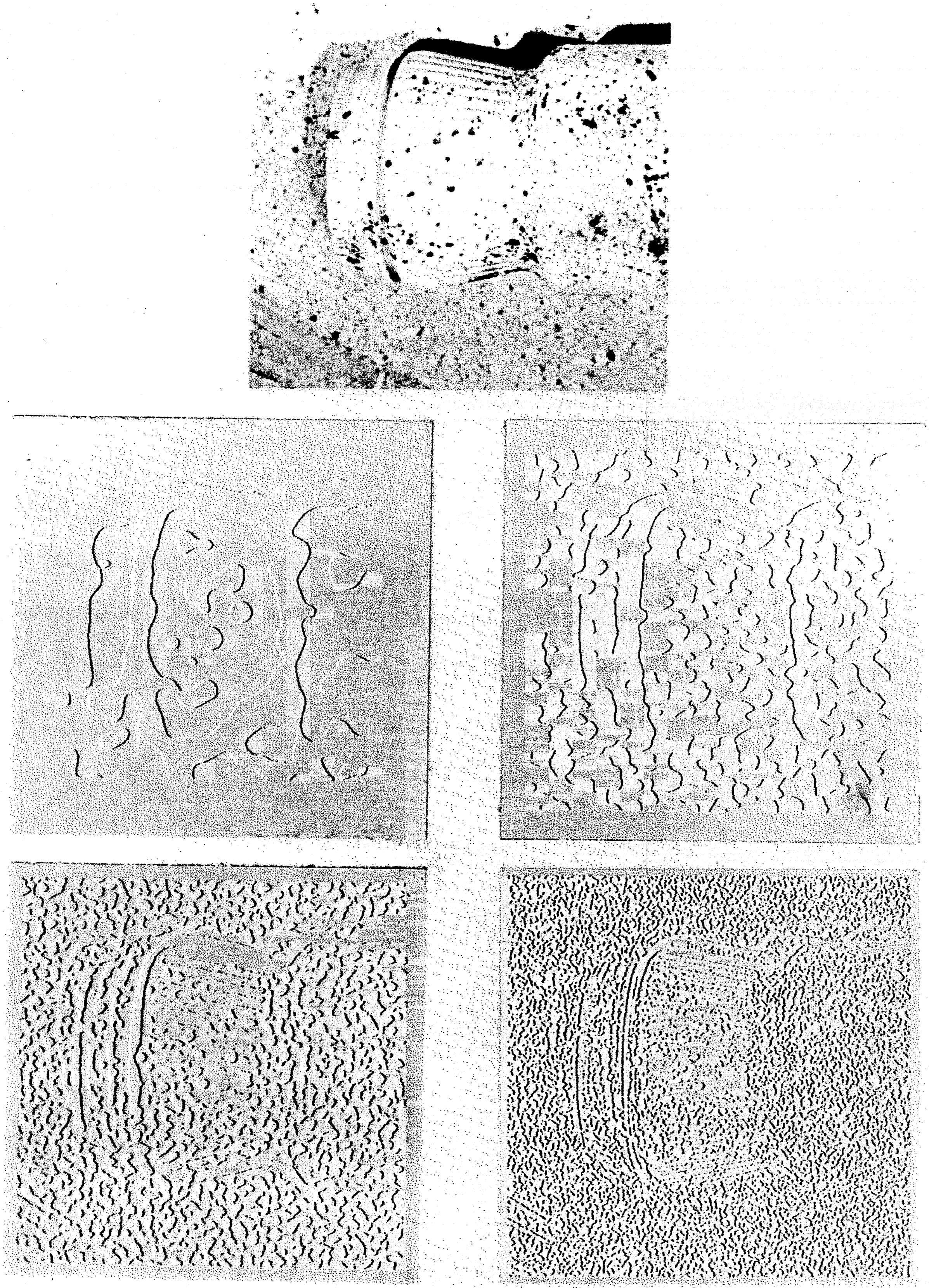
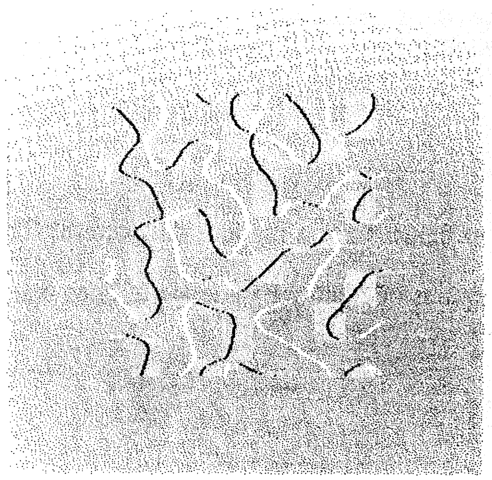
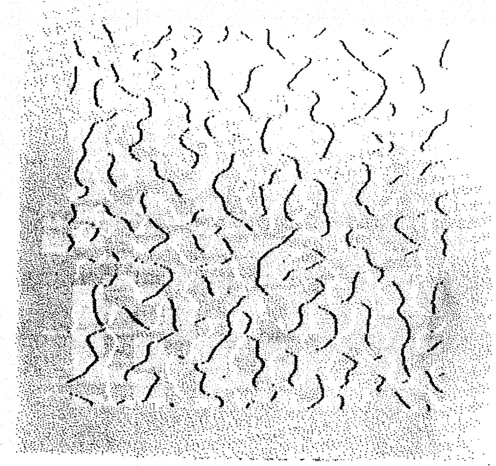
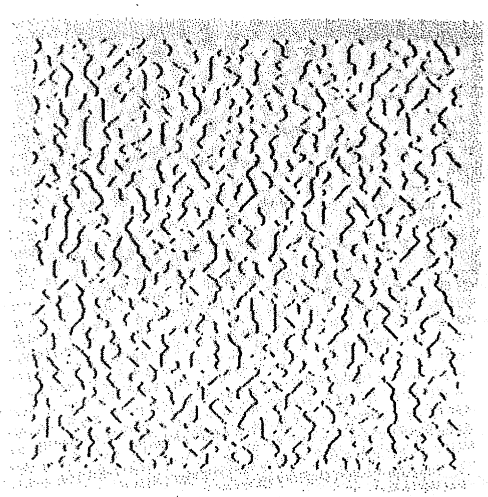
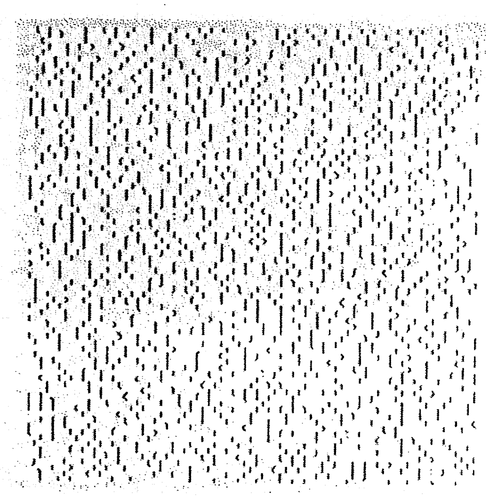


Figure 6. Examples of zero-crossing descriptions. The top figure show a random dot pattern. The bottom figures show the zero-crossings obtained from the convolutions of Figure 4. The white lines mark positive zero-crossings and the black lines, negative ones.









Once the eye positions have been fixed, and the retinal images extracted, the images are convolved with the DOG filters, and the zero-crossing descriptions are extracted from the convolved images. For a zero-crossing description corresponding to a particular mask size, the matching is performed by locating a zero-crossing and executing the following operation. Given the location of a zero-crossing in one image, a horizontal 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 $2w$, where w is the width of the central excitatory region of the corresponding two-dimensional convolution mask.

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

(1) the zero-crossings must come from convolutions with the same size mask.

(2) the zero-crossings must have the same sign.

(3) the zero-crossing segments must have roughly the same orientation.

A match is assigned on the basis of the number of pools containing a matching zero-crossing. If exactly one zero-crossing of the appropriate sign and orientation (within 30 degrees) 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 disparity sign of the unambiguous matches within that neighbourhood. (Disparity sign 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 disparity 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 1. In fact, Marr and Poggio show that the probability of a zero-crossing having at least one candidate match in this case is roughly 0.7. We can perform the following operation in this case. 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.

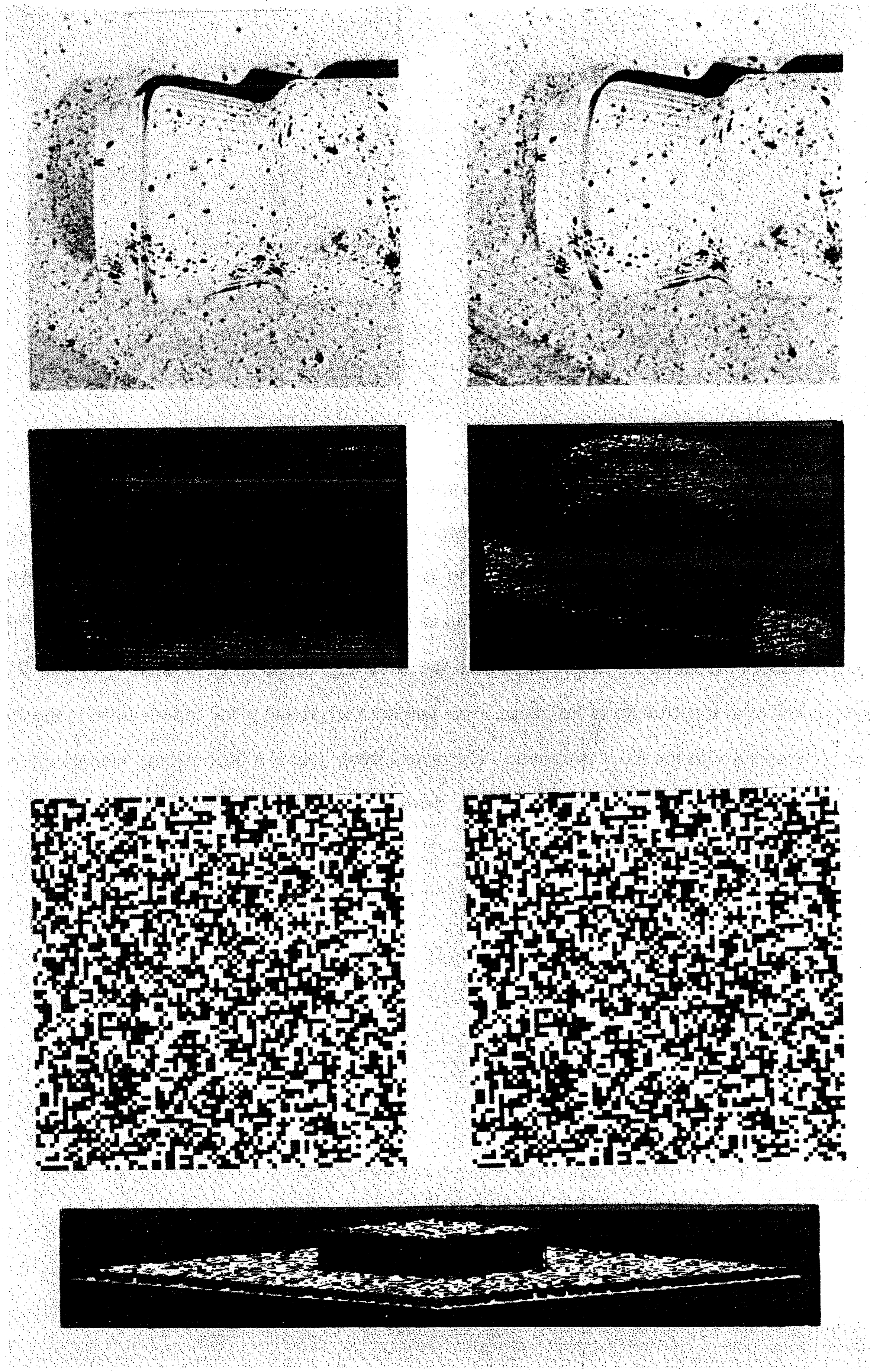
The overall effect of the matching process, as driven from the left image, is to assign disparity values to most of the zero-crossings obtained from the left image. An example of the output appears in Figure 7. In this array, a zero-crossing at position (x, y) with associated disparity d has been placed 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.

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

2.5 Vergence Control

The Marr-Poggio theory states that in order 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 positions of the eyes must

Figure 7. Results of the algorithm. The top stereo pair is an image of a painted coffee jar. The next two figures show 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 left view of the disparity map shows the jar as viewed from the lower edge of the image, and the right view show the jar as viewed 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 is a 50% density random dot pattern. The bottom figure shows the disparity map as viewed orthographically from some distance away. All disparity maps are those obtained from the $w = 4$ channel.



be assigned appropriately 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 which do not have disparity values for the smallest channel, but which do have disparity values for the larger channels. These large channel values are used to provide a refinement to the current eye positions, thereby bringing the smaller channels into range of correspondence. Two 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, or using a local average of the disparity values. In the current implementation, the search for such a region proceeds outwards from the fovea.

It should be noted here that although the use of disparity information from coarser channels to drive eye movements, allowing smaller channels to come into correspondence, is a necessary condition of the Marr-Poggio theory, it is not necessarily the only such condition. In other words, there may be other modules of the visual system which can initiate eye movements, and thereby affect the input to the matching component, by altering the retinal images presented to the matcher. An example of this would be the evidence of Kidd et al. (1979) concerning the ability of texture contours to facilitate stereopsis by initiating eye movements. However, such effects are somewhat orthogonal to the question of the sufficiency 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 degrees of disparity can be held fused in the fovea.)

We shall outline two different possibilities for the combination of the different channels. The method currently used in the 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 it reflects 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 images of the entire scene. 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, in order to store this information into a buffer, explicit information about the positions of the eyes is required. It will be argued in the discussion that this is probably inappropriate as a model of the human system. However, for the purposes of demonstrating the effectiveness of the matching module, such a representation is sufficient.

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 which is within range of fusion. Vergence movements are possible in order 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, those disparities being restricted to positions in the image lying along zero-crossings segments.

2.7 Summary of the process

The complete algorithm, as currently implemented, uses four mask sizes. Initially, the two views of the scene are mapped into a pair of retinal images. These images are convolved with each mask. The zero-crossings and their orientation are computed, for each channel and each view. The initial alignments of the eyes determine the registration of the images. The matching of the descriptions from each channel is performed for this alignment. Any points with either ambiguous matchings or with 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, determined by the statistical analysis of Marr and Poggio (1979), are allowed to remain. All other points are removed. The values which are kept are stored into a buffer. At this stage, vergence movements may take place, using 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 which do not have disparity values at any level of 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 mask 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 masks. 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

A standard tool in the examination of human stereo perception is the random dot stereogram (Julesz, 1960, 1971). This is a pair of stereo images where each image, when viewed monocularly, consists only of randomly distributed dots, yet when viewed stereoscopically, may be fused to yield patterns separated in depth. Such patterns are a useful tool for analysing the stereo component of the human visual system, since there are no visual cues other than the stereoscopic ones. We can test the sufficiency of the algorithm by comparing human perception with the performance of the algorithm on such patterns. As well, since random dot stereograms have well demarked disparity values, it is easy to assess the correctness of the algorithm's performance on such patterns.

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

The first pattern consisted of a central square separated in depth from a second plane. 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 two minutes 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 which were assigned a disparity was 11847. Of these 11847, 11830 were disparity values which were exactly correct, and an

TABLE OF MATCHES						
pattern	density	total	exact	one pixel	wrong	%wrong
square	50%	11847	11830	14	3	.03
square	25%	9661	9632	22	7	.07
square	10%	5286	5264	20	2	.04
square	5%	3500	3498	0	2	.06
wedding	50%	11162	11095	61	6	.06
hnoise1	50%	2270	1909	346	15	.7
hnoise2	50%	8683	6621	1868	194	2.
hnoise3	50%	63	28	24	11	17.
hnoise4	50%	8543	5194	2864	485	6.
uncorr1	50%	9545	9091	263	191	2.
uncorr2	50%	4343	4120	143	80	2.
uncorr3	50%	134	127	2	5	4.
uncorrd	50%	6753	6325	271	157	2.

Table 1.

additional 14 deviated by one image element from the correct value. Approximately 0.03% of the matched points, or roughly 3 points in 10000 were incorrectly matched.

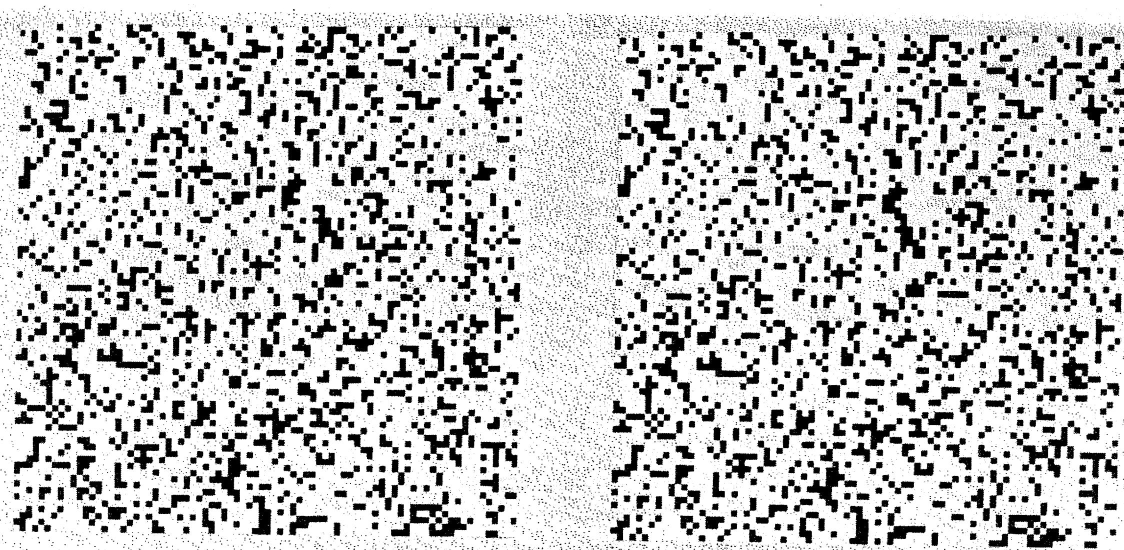
A similar test was run on patterns with a dot density of 25%, 10% and 5%. The results are illustrated in Figure 8.

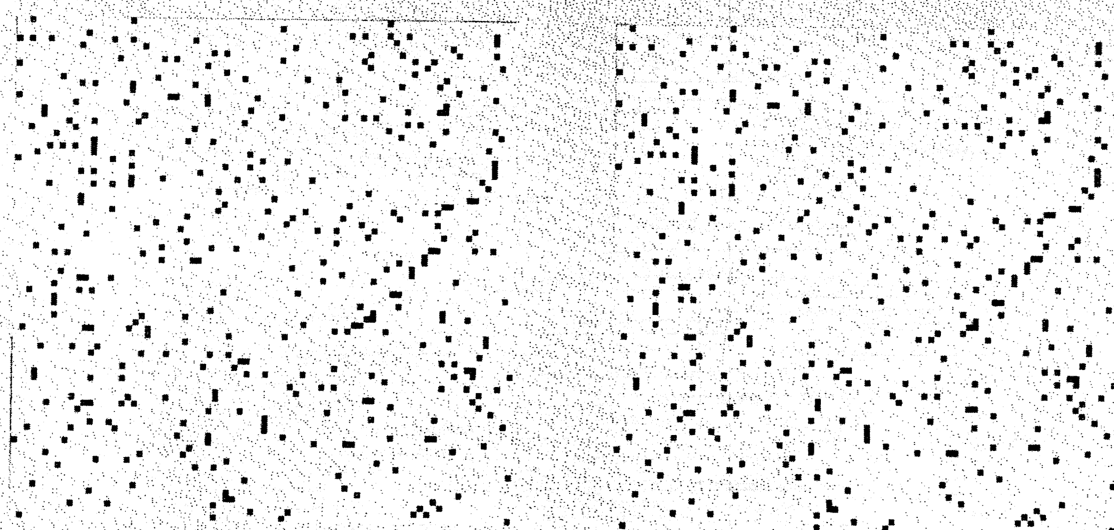
For each of these cases, the number of incorrectly matched points was extremely low. Those points which were assigned incorrect disparities all occured at the border between the two planes, that is, along the discontinuity in disparity.

A more complex random dot pattern consisted of a wedding cake, built from four different planar layers, each separated by 8 image elements, or 2 dot widths. This is illustrated in Figure 9.

In this case, the number of zero-crossing points assigned a disparity was 11162. Of these points, 11095 were assigned a disparity value which 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 occured at the boundaries between the planes. A second complex pattern is illustrated in Figure 9.

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





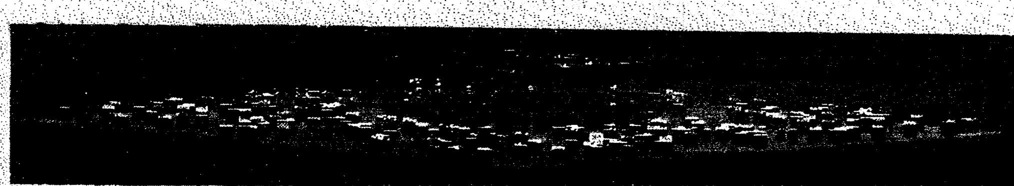
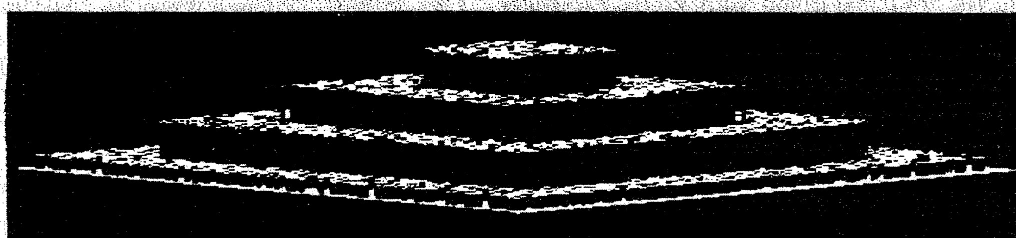


Figure 9. The top stereo pair is a 50% density wedding cake, composed of four planar levels. The disparity map is shown below it. The bottom stereo pair is a 50% spiral. The disparity map is shown below it, in a manner similar to Figure 7. Both disparity maps are obtained from the $w = 4$ channel.



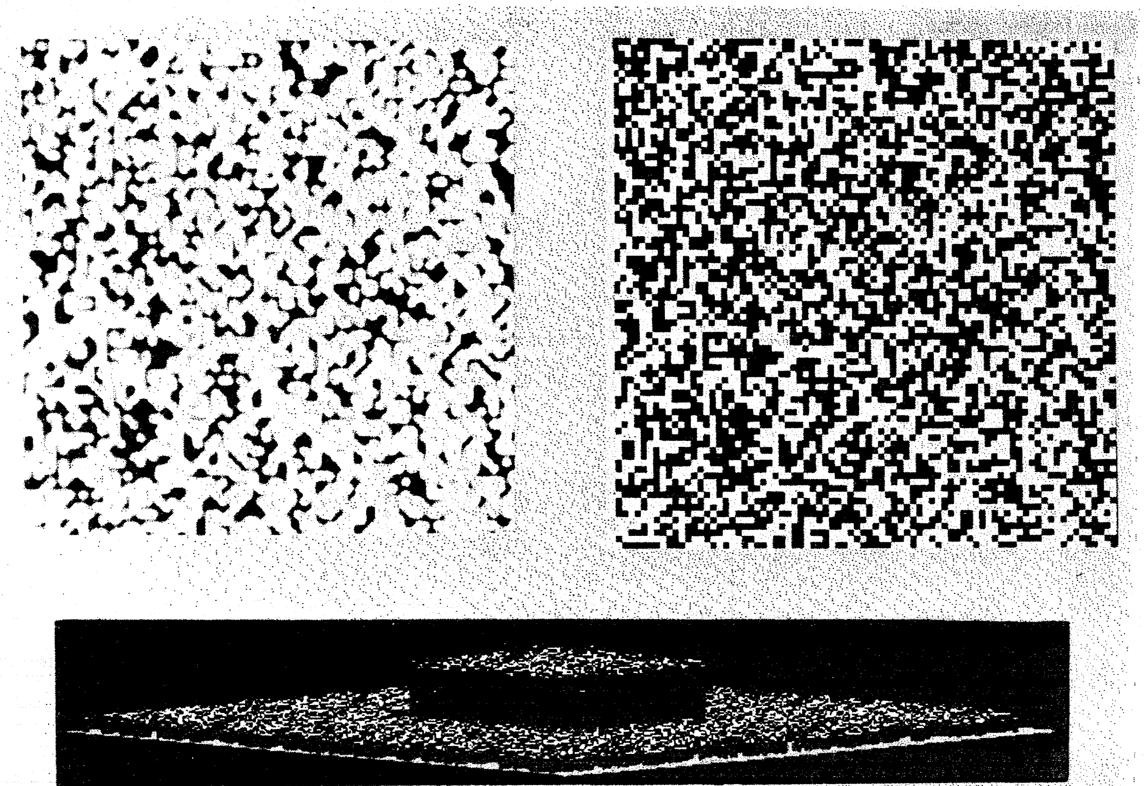
The object is a spiral with a range of continuously varying disparities.

There are a number of special cases of random dot patterns which 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 between the performance of the algorithm, and human perception is given.

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 mask. This is illustrated in Figure 10. The disparity values obtained in this case were not as exact as in the case of no 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%. However, the qualitative performance of the algorithm is still that of two planes separated in depth. It is interesting to note that 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.

Julesz and 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 disjoint from the spectrum of the pattern, it can be demonstrated that 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 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. In the case shown by HNOISE1 and HNOISE2 in Table 1, the noise was added such that the maximum magnitude of the noise was equal to the maximum magnitude of the original image. HNOISE1 illustrates the performance of the smallest channel. HNOISE2 illustrates the performance of the next larger channel. It can be seen that for this case, some fusion is still possible in the smallest channel, although it is patchy. The

Figure 10. The top stereo pair is a 50% density pattern in which the left image has been blurred. The disparity map 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 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 shown is that obtained by the $w = 9$ channel.



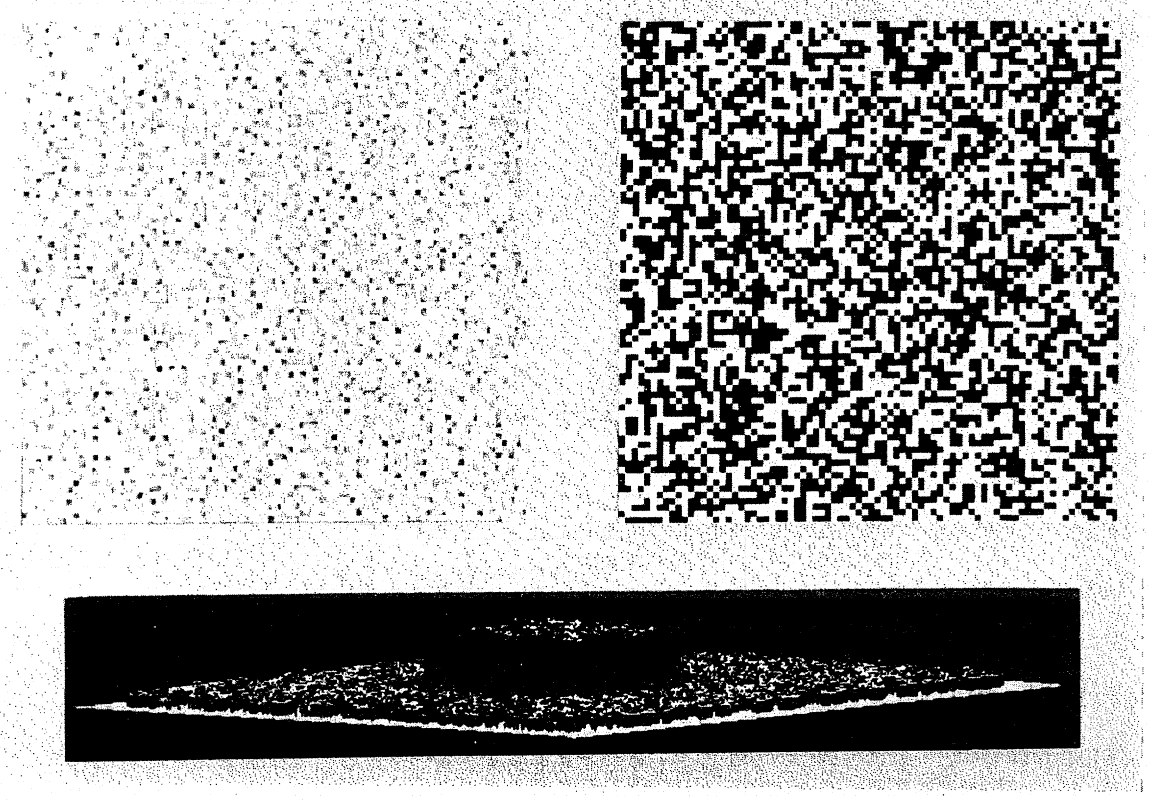
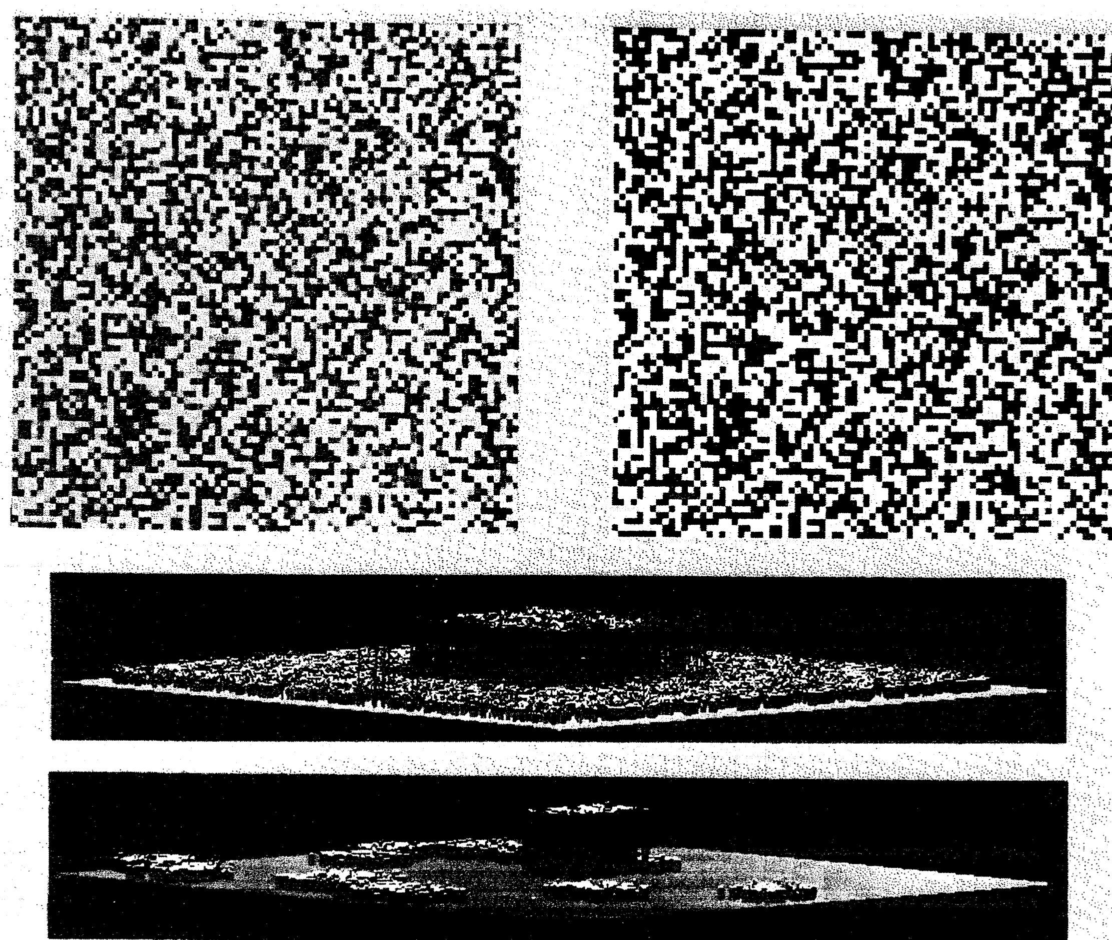
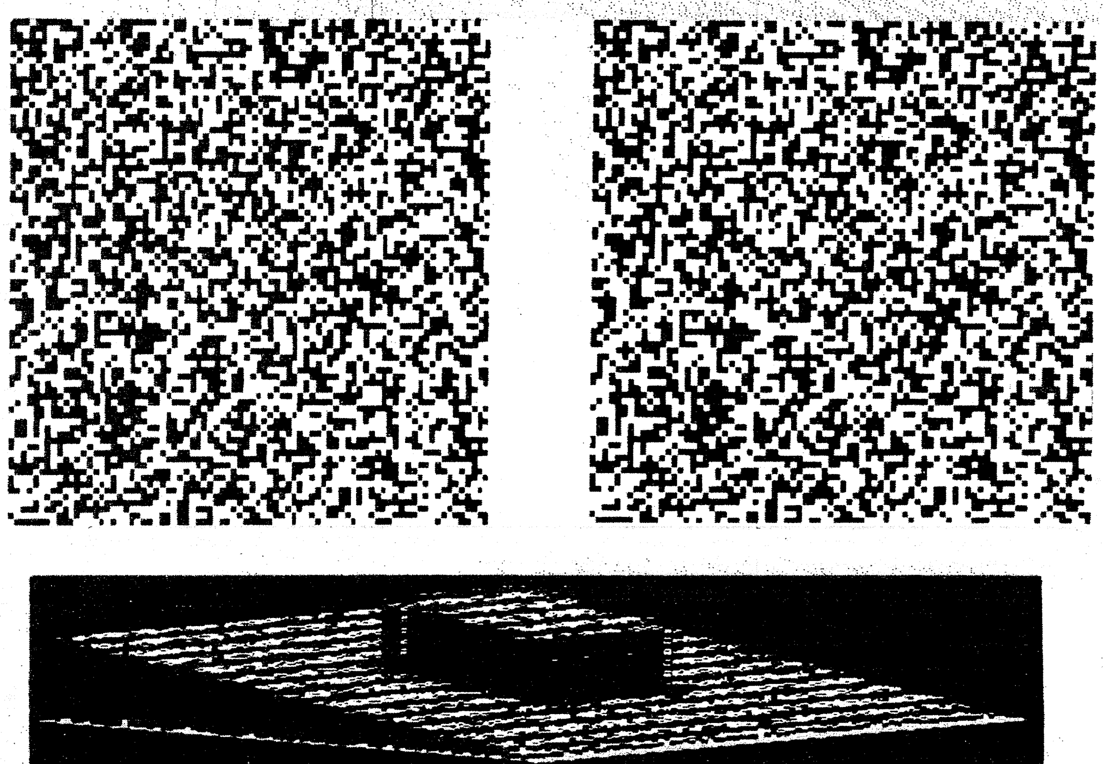


Figure 11. The top stereo pair is a 50% density pattern. The left image has ahd high pass filtered noise added to it so that the maximum magnitude of the noise is half the maximum magnitude of the image. The top disparity map is that obtained from the $w = 9$ channel, while the next disparity map is 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 is a 50% density pattern in which the left image has been compressed in the horizontal direction. The disparity map 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.





next larger channel also obtains fusion. In both cases, the accuracy of the disparity values is reduced from the normal case. This is to be expected, since the introduction of noise tends to displace the positions of the zero-crossings. In the case shown by HNOISE3 and HNOISE4 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 (HNOISE3). Yet matching in the next larger channel is only marginally affected (HNOISE4).

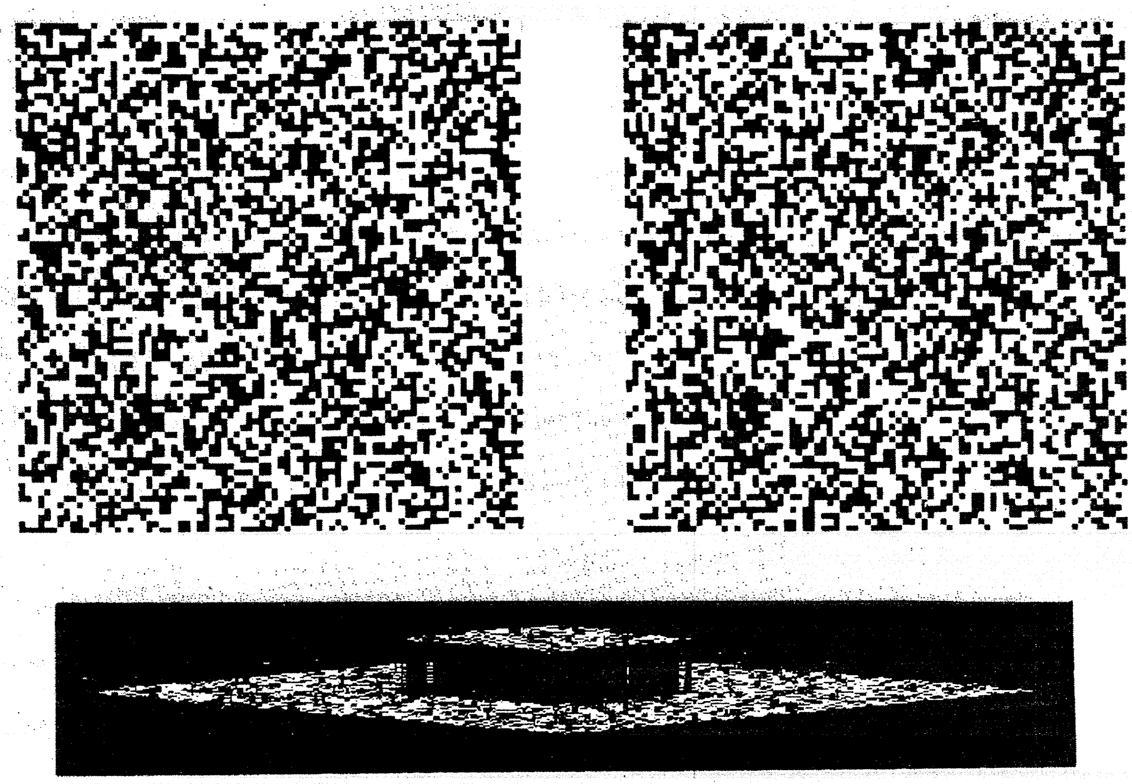
The implementation was also tested on the case of adding low pass filtered noise to a random dot pattern, with results similar to that of adding high pass filtered noise. Here, the larger channels are unable to obtain a good matching, while the smaller channels are relatively unaffected.

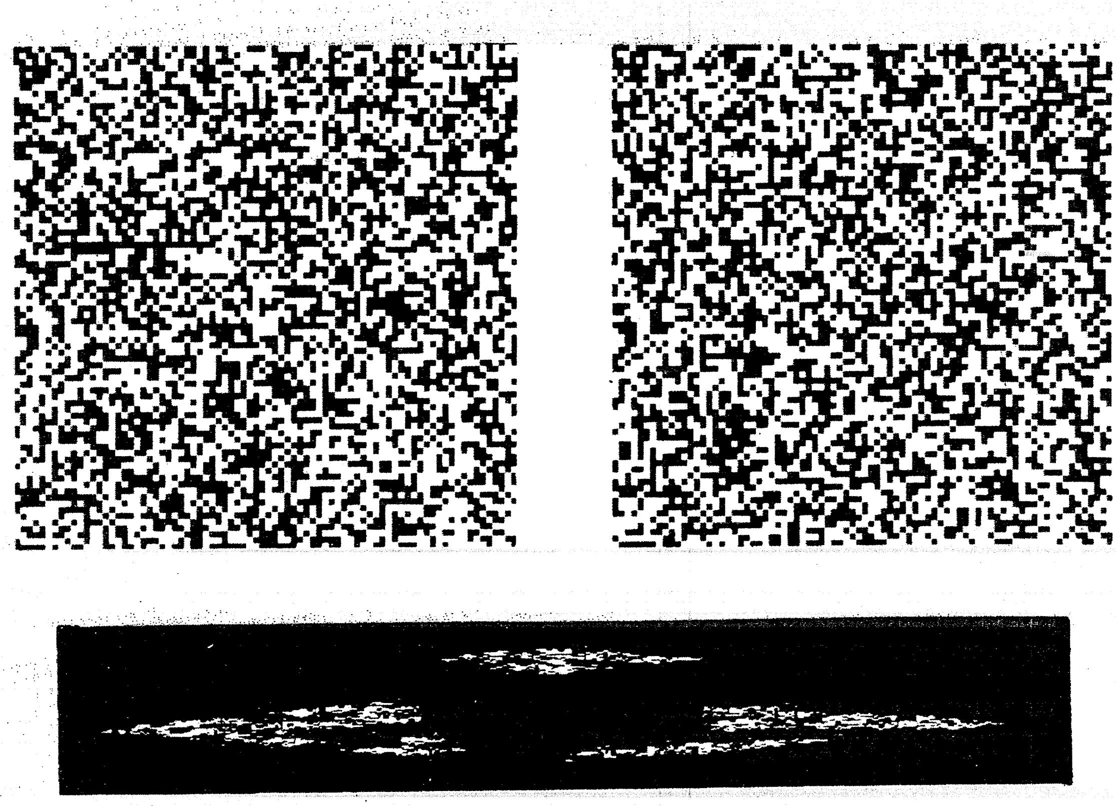
If one of the images of a random dot pattern is compressed in the horizontal direction, the human stereo system is still able to achieve fusion (Julesz 1971, p.213). The algorithm was tested on this case, and the results are shown in Figure 11. It can be seen that the program still obtains a reasonably 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, the cases of 10%, 20% and 30% were tried, and are illustrated in Figure 12. For the 10% case, (table entry Uncorr1) 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 Uncorr2), the number of matched points decreased again, although the percentage of those which were incorrectly matched remained about the same. Finally, when the percentage of decorrelated dots was increased to 30% (table entry Uncorr3), the algorithm found virtually no section of the image which could be fused.

The failure of the algorithm to match the 30% decorrelated pattern is caused by the component of the algorithm which checks that each region of the image is within range of correspondence. Recall that in order

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





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 image within range of fusion, the Marr-Poggio theory requires that the percentage of unmatched points is less than some threshold. This threshold is approximately 0.3, according to the statistical analysis of Marr and Poggio (1979). For the case of the pattern with 30% decorrelation, on the average, each region of the image will have roughly 30% of its zero-crossings different and hence the algorithm decides that the region is out of range of correspondence. Hence, no disparitites are accepted for this region.

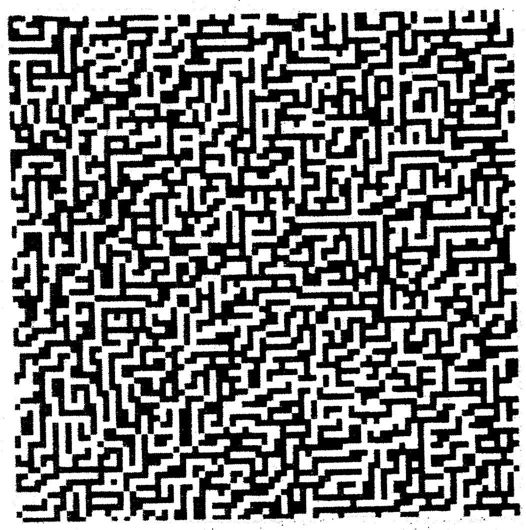
For the algorithm, the computational reason for the failure to process patterns with 30% decorrelation is that it could not distinguish a correctly matched region of such a pattern from a region which was out of range of correspondence, but had a random set of matches for many of the points in the region. It is interesting to note that many human subjects observe a similar behavior; that is, some kind of fusion for up to 20% decorrelation, although the fusion becomes increasingly weaker, and virtually no fusion for patterns with 30% decorrelation.

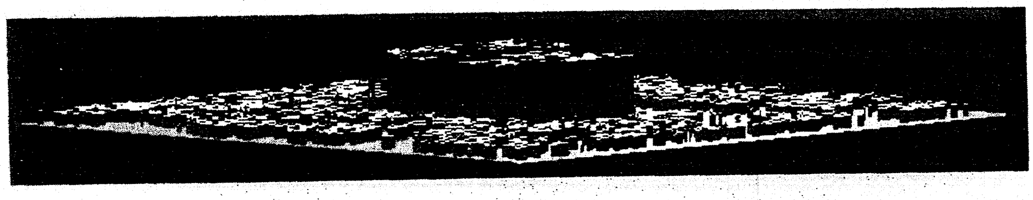
One can also decorrelate the pattern by breaking up all white triplets along one set of diagonals, and all black triplets along the other set of diagonals (Julesz 1971, p.87). The table entry Uncorrd indicates the matching statistics for this case. Again, it can be seen that the program still obtains a good match, as do human observers. The performance of the algorithm is illustrated in Figure 13.

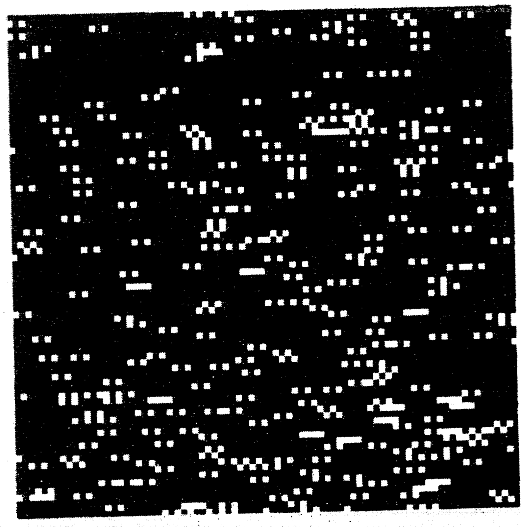
4. Statistics

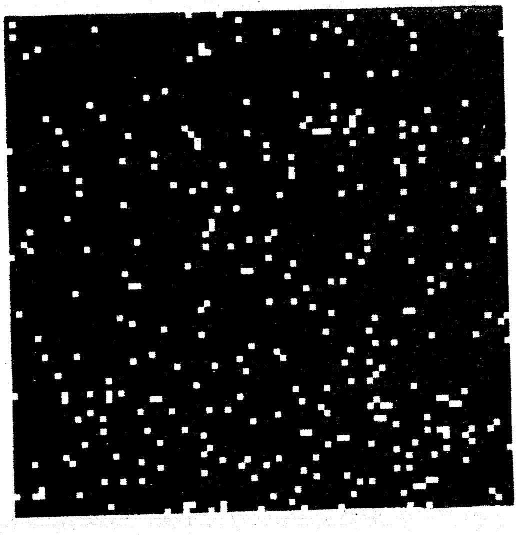
A number of parameters are important for the theory, which makes assumptions about them, and they have been measured on random dot images. The worst cases occur for patterns with a density of 50%, and

Figure 13. The top stereo pair is a 50% density pattern in which the left 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 is shown below. The bottom stereo pair 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 is shown below, and consists of two superimposed planes.









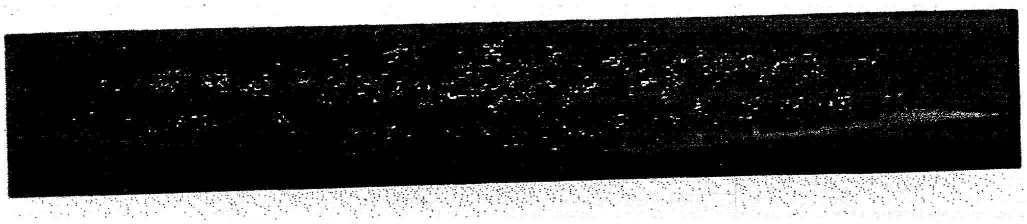


TABLE OF STATISTICS				
parameter	expected worst case-behavior	large channel w = 35	medium channel w = 17	small channel w = 9
average distance between zero-crossings of same sign	2 w	1.51 w	1.88 w	1.87 w
probability of candidates in at most one pool	> .50	.77	.75	.69
probability of candidates in two pools	< .45	.21	.25	.31
probability of candidates in all three pools	< .05	.02	.01	.01
given a candidate near zero, probability of no other candidates	> .9	.88	.85	.87

Table 2.

for such patterns the worst case values encountered for the parameters have the values shown in Table 2. The theoretical worst case bounds used by Marr and Poggio appear for comparison.

5. Comments and Discussion

Implementing a computational theory offers us the opportunity of testing its adequacy. In this case, 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 defocussing of, compression of, and the introduction of various kinds of masking noise to one image of a random dot stereo pair.

The process of implementing the theory also led to the following observations and refinements of the theory.

(1) There are a number of questions concerning the form of the $2\frac{1}{2}$-D sketch. The first critical question concerns whether the sketch reflects the initial or the retinal images. In the first case, the coordinates of the sketch would be directly related to the coordinates of the images of the entire scene. However, since disparity information about the scene is extracted from several eye positions, in order 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. For the computer implementation, this is possible, but for a model of the human visual system, it seems unlikely that such information is 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 would be retinocentric, reflecting the current positions of the eyes. This seems to be the most natural representation.

The second question concerns the use of a fovea. Different sections of the images are analyzed 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. Thus, a refinement to the implementation, as outlined above, would be to use a representation that is retinocentric, and which represents disparities with decreasing resolution as eccentricity increases.

For the cases illustrated in this article, the $2\frac{1}{2}$-D sketch was created by storing fine resolution disparity values into a scene-centered representation. 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, with good results.

(2) The neighbourhood over which a search for a matching zero-crossing is conducted is broken into three pools. 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 a 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 $\frac{\pm w}{2}$, corresponding to the divergent or convergent pools.

Computer experiments show that either scheme will work. In the case of a single disparity value for each pool, the disparities assigned by the smallest channel are within an image element of those obtained using 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.

(3) 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 O. 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 the matching process were only driven from one image, for example, matches were made from the right image to the left. In this case, the implementation would not be able to account for the Braddick perception, 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, and an unambiguous match from either side accepted. In this case, although every zero-crossing in the right image would have an ambiguous match, the implementation would obtain a unique match for each zero-crossing in the left image. The implementation was designed to account for matching from either image.

Braddick's case has been tested on the implementation, and the results are shown in figure 13. It can be seen that the results of the implementation are that of two transparent planes.

(4) 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 which is present in one image, but not 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 which 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 which 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.

(5) It is interesting to comment on the effect of depth discontinuities for the different sized masks. 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 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.

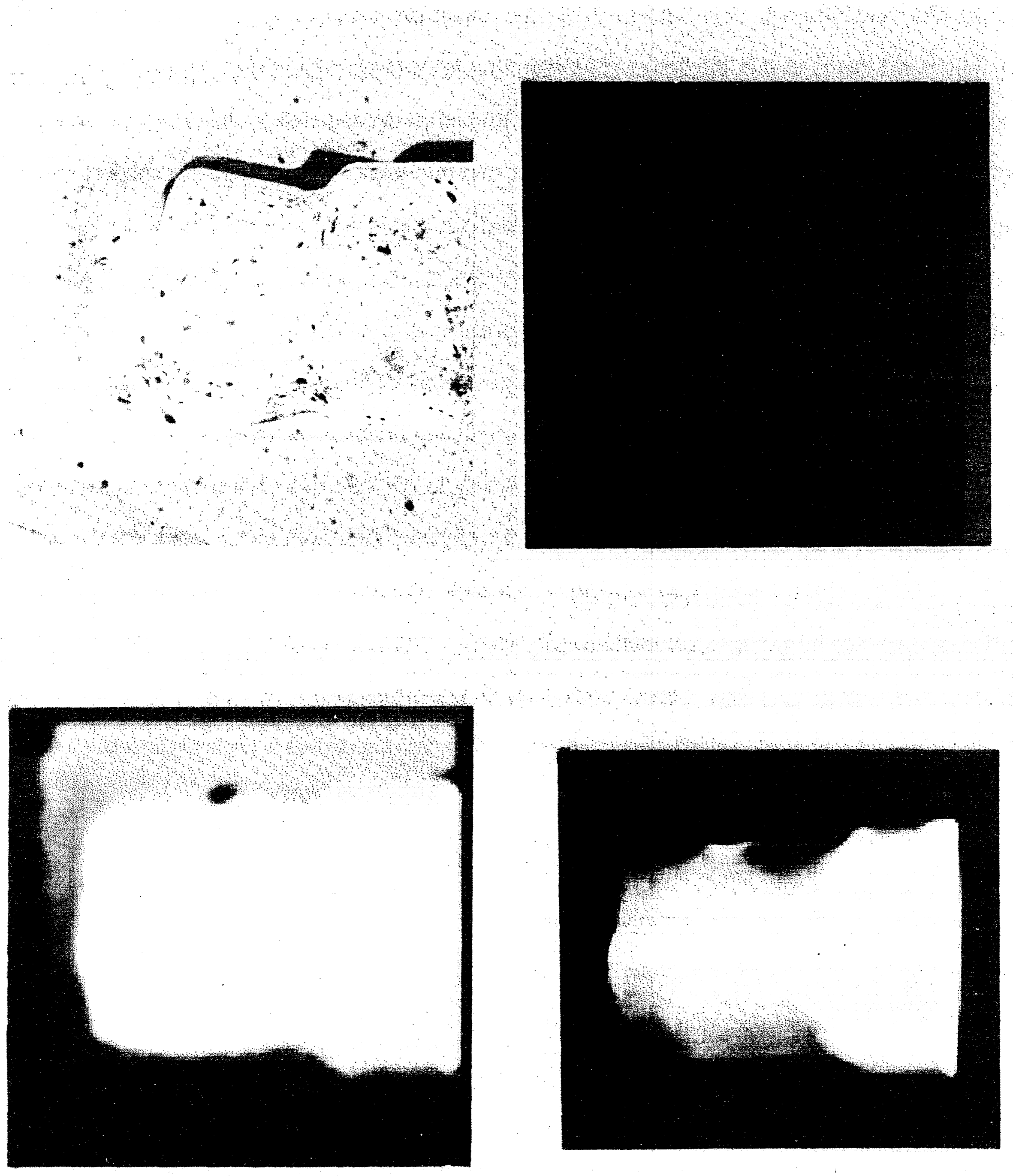
For the smaller mask this does not occur, as the zero-crossing contours tend to outline individual dots or connected groups of dots. Thus the disparities assigned are such that the dots belong to one plane or the other and the final disparity map is one of two separated planes.

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, in preparation). 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.

(6) One consequence of the Marr-Poggio theory is that explicit disparity values will be obtained only along the zero-crossing contours. It may be desirable to create a more complete reconstruction of the shapes of the objects in the scene, by filling in disparity values between the zero-crossing contours. Some work has been done in this direction (Grimson, in preparation) and an example is shown in Figure 14.

(7) 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 and

Figure 14. Example of filling in the disparity map. The top left figure is the initial image. The top right figure shows the disparity map associated with the image, where the disparity is represented by the intensity of the point. The bottom figures show the filled in map, again using intensity to represent disparity. In the left figure, the full range of disparity is shown, indicating the slant of the background plane, and the extreme difference in disparity between the jar and the background. In the right figure, the intensities have been adjusted to enhance the disparities of the jar, indicating the general shape of the interpolated surface.

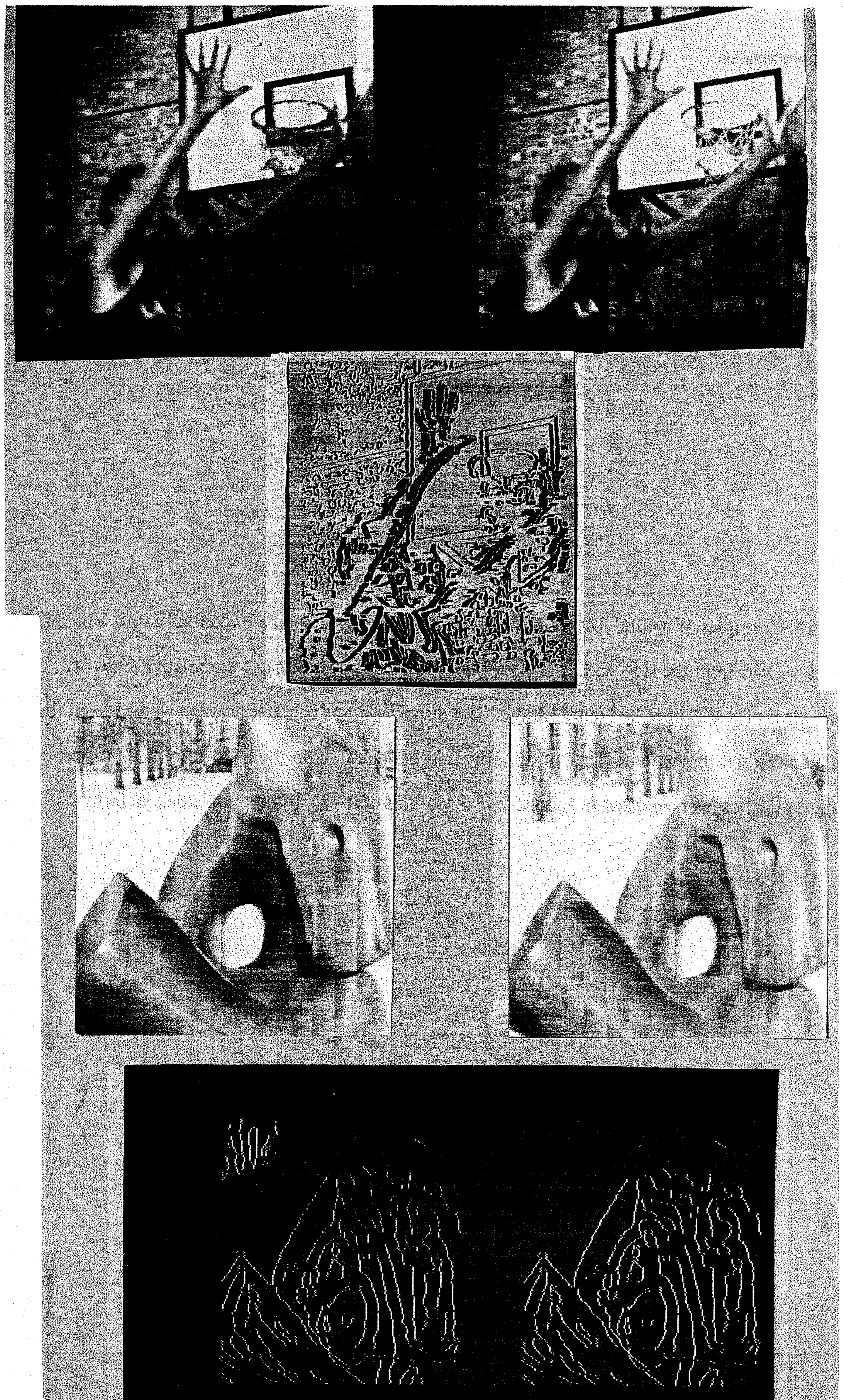


Poggio, 1979, Marr and Hildreth, 1980, Ullman, 1979). The constraints so derived are critical in the formation of the computational theory, and in the design of an algorithm for solving the problem. An interesting question to raise is whether the algorithm explicitly checks that the constraints imposed 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 case of 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 explicity 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 which 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, for matching in this direction. 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 above would indicate that this is not the case. 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.

(8) It is worth observing the distinction between the performance of the implementation on random dot patterns and the performance of the implementation on natural images. Some examples are shown in Figure 15. The main point is that on the whole, the performance is quite acceptable for random dot patterns. However, the implementation can occasionally fail in the case of natural images. The question is whether this reflects a basic inadequacy in the theory and its implementation, or whether there are other aspects of the visual process interacting with stereo which have not been included in this implementation.

This can be approached in two ways: (1) Is the assumption of modularity incorrect? In other words, is there something wrong with the matching module as developed by Marr and Poggio, and as implemented here. (2) Are there other modules, not considered here, which may affect the input or the output of the

Figure 15. Examples of natural images. The top stereo pair is a scene of a basketball game. The disparity map below is viewed from the side, so that the width of the black bars indicates the relative disparity. The bottom stereo pair is of a sculpture by Henry Moore. The disparity maps below it are also viewed from the side. The left map illustrates the extreme range of disparity between the trees in the background and the sculpture itself. The right map has been adjusted to enhance the disparities of the sculpture, indicating its form.



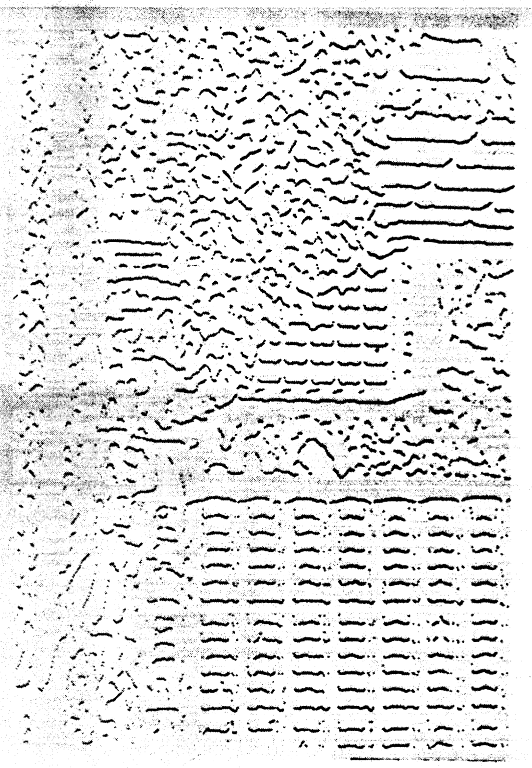
matching module?

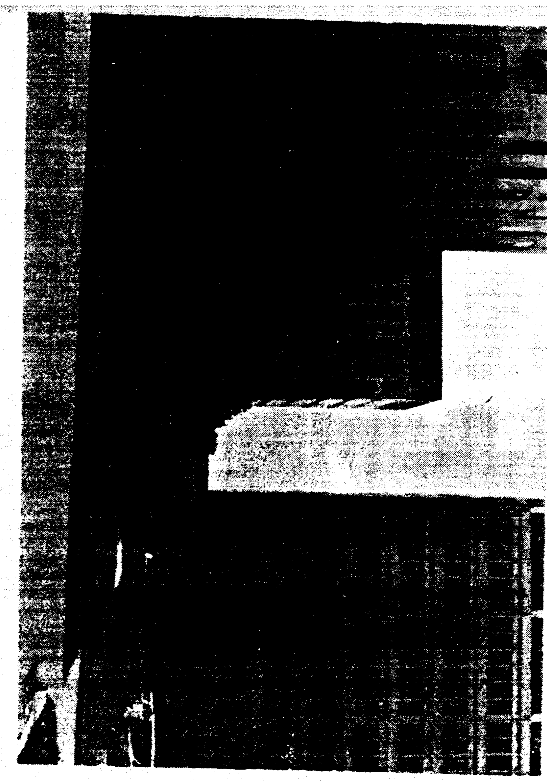
The results of testing the implementation on the broad range of images, indicated in previous sections, seems to indicate that the matching module is acceptable as an independent one. In particular, the agreement between the performance of the algorithm and that of human observers on the many random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

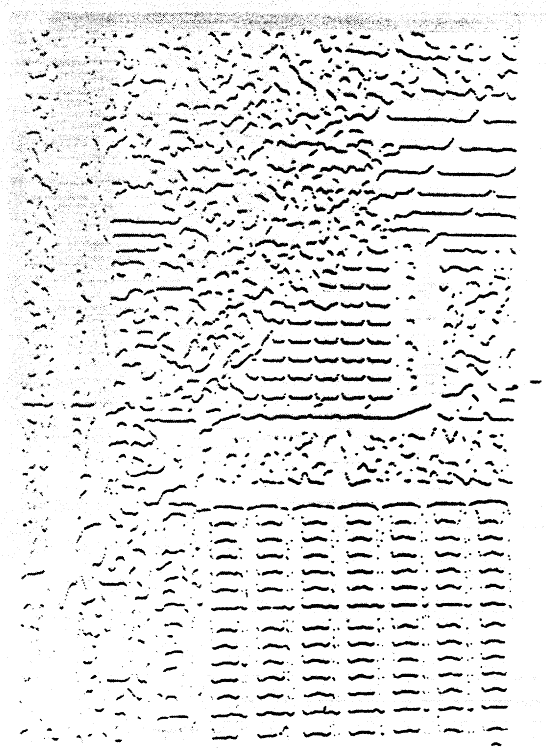
When we turn to natural images, it is reasonable to expect that other visual modules may affect the input to the matcher and that they may alter the output of the matcher. This is not to suggest that the matcher is incorrect, only that the effects of other modules must be taken into account in order to explain the complete human perception. For example, the evidence of Kidd, Frisby and Mayhew (1979) concerning the ability of texture boundaries to drive eye vergence 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 matcher itself needs to be modified.

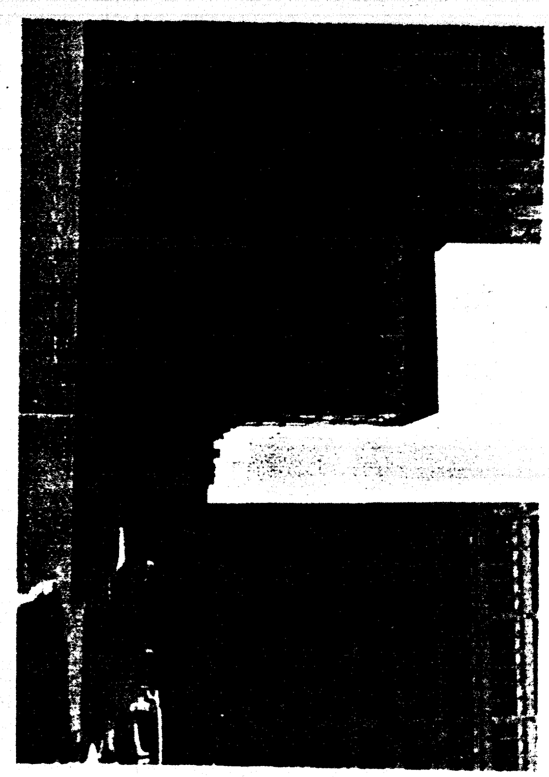
Interestingly, the performance of the implementation supports this point. The implementation, which is considered a distinct module, also performs very well on random dot patterns, where there is no possibility of interaction with other visual processes. For many natural images, this is still true. However, 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 16 . Here, the regular pattern of the windows provides a strong false targets problem. In running the implementation, the following behavior was observed. If the optical axes were aligned at the level of the building, the zero-crossings corresponding to the windows were all assigned a correct disparity. If, however, the optical axes were aligned at the level 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 is the implementation wrong? Curiously, if one fuses the zero-crossing descriptions of the convolved images without eye movements, human observers have the same problem: if the eyes are fixated at the level of the building, the windows are correctly matched; if the eyes are fixated at the level of the trees,

Figure 16. The false targest problem. The top figures are a stereo pair of a group of buildings. The bottom figures show 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.









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 matching and writes it into the $2\frac{1}{2}$-D sketch. However, it is probably the case that some later processing module, which examines the contents of the $2\frac{1}{2}$-D sketch, is capable of altering the contents stored there, based on more global information than is available to the matching component of the stereo process.

Thus, I would suggest that future refinements to the Marr-Poggio theory must account for the interactions of other aspects of visual information processing on the input and output of the matching module. Some initial work has already been done in this direction (Grimson, in preparation).

6. Acknowledgements

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.

7. References

Braddick, O. 1978 Multiple matching in stereopsis. (unpublished MIT report).

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

Grimson, W.E.L. A refinement of a computational theory of human stereo vision *in preparation*.

Grimson, W.E.L. and Marr, D. 1979 A computer implementation of a theory of human stereo vision. *Proceedings: Image Understanding Workshop* 41-47.

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: The University of Chicago Press.

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

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

Knight, T.F., Moon, D.A., Holloway, J., and Steele, G.L. 1979 CADR *MIT Artificial Intelligence Laboratory Memo* 528.

Marr, D. and Hildreth, E. 1980 Theory of edge detection. *Proc. R. Soc. Lond.* (in the press).

Marr, D. and 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. and Poggio, T. 1976 Cooperative computation of stereo disparity. *Science, N.Y.* 194, 283-287.

Marr, D. and Poggio, T. 1979 A computational theory of human stereo vision. *Proc. R. Soc. Lond.* B. 204, 301-328.

Marr, D., Poggio, T. and Hildreth, E. 1979 The smallest channel in early human vision. *JOSA* (submitted for publication).

Mayhew, J.E.W. and Frisby, J.P. 1976 Rivalrous texture stereograms. *Nature, Lond.* 264, 53-56.

O'Brien, B. 1951 Vision and resolution in the central retina. *J. Opt. Soc. Am.* 41, 882-894.

Ullman, S. 1979 *The interpretation of visual motion* Cambridge: MIT Press.

Wilson, H.R. and Bergen, J.R. 1979 A four mechanism model for spatial vision. *Vision Res.* (in the press).

Wilson, H.R. and Giese, S.C. 1977 Threshold visibility of frequency gradient patterns. *Vision Res.* 17, 1177-1190.

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 image within range of fusion, the Marr-Poggio theory requires that the percentage of unmatched points is less than some threshold. This threshold is approximately 0.3, according to the statistical analysis of Marr and Poggio (1979). For the case of the pattern with 30% decorrelation, on the average, each region of the image will have roughly 30% of its zero-crossings different and hence the algorithm decides that the region is out of range of correspondence. Hence, no disparitites are accepted for this region.

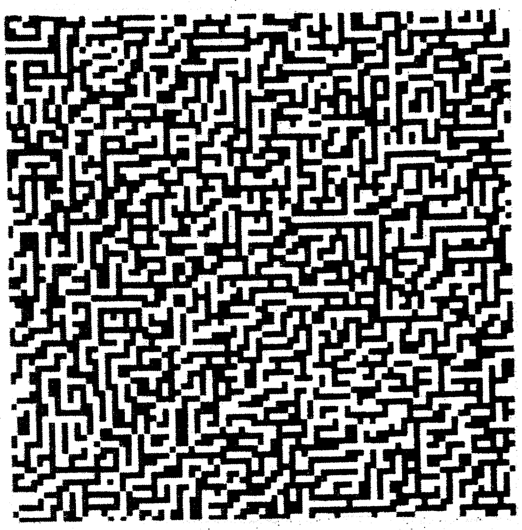
For the algorithm, the computational reason for the failure to process patterns with 30% decorrelation is that it could not distinguish a correctly matched region of such a pattern from a region which was out of range of correspondence, but had a random set of matches for many of the points in the region. It is interesting to note that many human subjects observe a similar behavior; that is, some kind of fusion for up to 20% decorrelation, although the fusion becomes increasingly weaker, and virtually no fusion for patterns with 30% decorrelation.

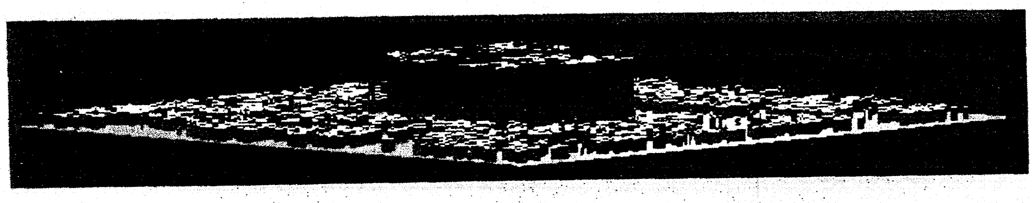
One can also decorrelate the pattern by breaking up all white triplets along one set of diagonals, and all black triplets along the other set of diagonals (Julesz 1971, p.87). The table entry Uncorrd indicates the matching statistics for this case. Again, it can be seen that the program still obtains a good match, as do human observers. The performance of the algorithm is illustrated in Figure 13.

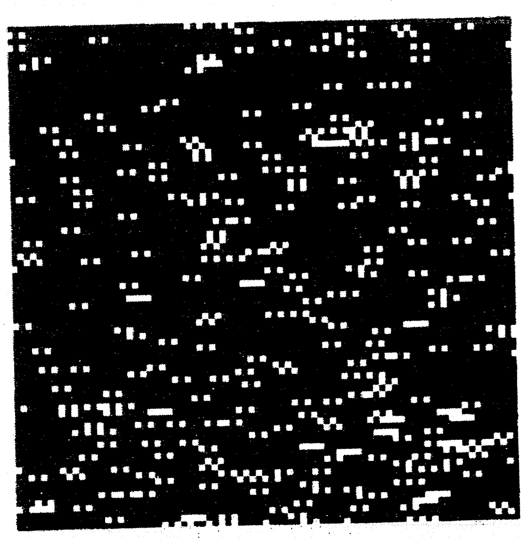
4. Statistics

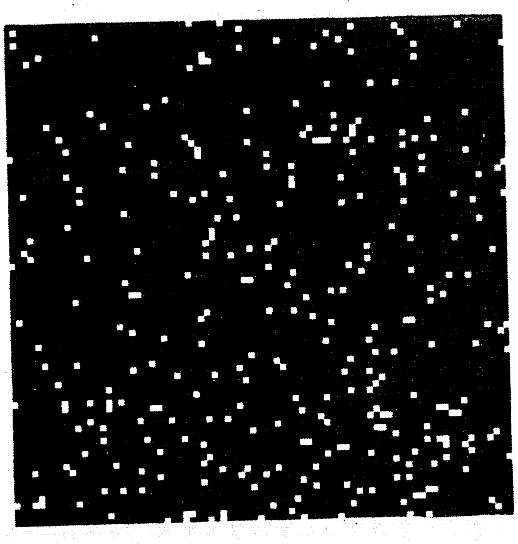
A number of parameters are important for the theory, which makes assumptions about them, and they have been measured on random dot images. The worst cases occur for patterns with a density of 50%, and

Figure 13. The top stereo pair is a 50% density pattern in which the left 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 is shown below. The bottom stereo pair 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 is shown below, and consists of two superimposed planes.









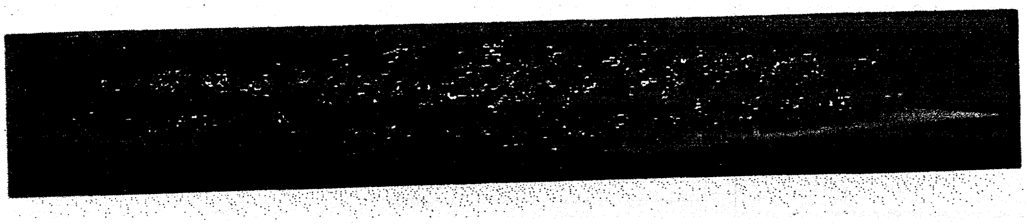


TABLE OF STATISTICS				
parameter	expected worst case-behavior	large channel w = 35	medium channel w = 17	small channel w = 9
average distance between zero-crossings of same sign	2 w	1.51 w	1.88 w	1.87 w
probability of candidates in at most one pool	> .50	.77	.75	.69
probability of candidates in two pools	< .45	.21	.25	.31
probability of candidates in all three pools	< .05	.02	.01	.01
given a candidate near zero, probability of no other candidates	> .9	.88	.85	.87

Table 2.

for such patterns the worst case values encountered for the parameters have the values shown in Table 2. The theoretical worst case bounds used by Marr and Poggio appear for comparison.

5. Comments and Discussion

Implementing a computational theory offers us the opportunity of testing its adequacy. In this case, 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 defocussing of, compression of, and the introduction of various kinds of masking noise to one image of a random dot stereo pair.

The process of implementing the theory also led to the following observations and refinements of the theory.

(1) There are a number of questions concerning the form of the $2\frac{1}{2}$-D sketch. The first critical question concerns whether the sketch reflects the initial or the retinal images. In the first case, the coordinates of the sketch would be directly related to the coordinates of the images of the entire scene. However, since disparity information about the scene is extracted from several eye positions, in order 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. For the computer implementation, this is possible, but for a model of the human visual system, it seems unlikely that such information is 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 would be retinocentric, reflecting the current positions of the eyes. This seems to be the most natural representation.

The second question concerns the use of a fovea. Different sections of the images are analyzed 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. Thus, a refinement to the implementation, as outlined above, would be to use a representation that is retinocentric, and which represents disparities with decreasing resolution as eccentricity increases.

For the cases illustrated in this article, the $2\frac{1}{2}$-D sketch was created by storing fine resolution disparity values into a scene-centered representation. 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, with good results.

(2) The neighbourhood over which a search for a matching zero-crossing is conducted is broken into three pools. 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 a 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 $\frac{\pm w}{2}$, corresponding to the divergent or convergent pools.

Computer experiments show that either scheme will work. In the case of a single disparity value for each pool, the disparities assigned by the smallest channel are within an image element of those obtained using 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.

(3) 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 O. 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 the matching process were only driven from one image, for example, matches were made from the right image to the left. In this case, the implementation would not be able to account for the Braddick perception, 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, and an unambiguous match from either side accepted. In this case, although every zero-crossing in the right image would have an ambiguous match, the implementation would obtain a unique match for each zero-crossing in the left image. The implementation was designed to account for matching from either image.

Braddick's case has been tested on the implementation, and the results are shown in figure 13. It can be seen that the results of the implementation are that of two transparent planes.

(4) 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 which is present in one image, but not 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 which 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 which 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.

(5) It is interesting to comment on the effect of depth discontinuities for the different sized masks. 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 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.

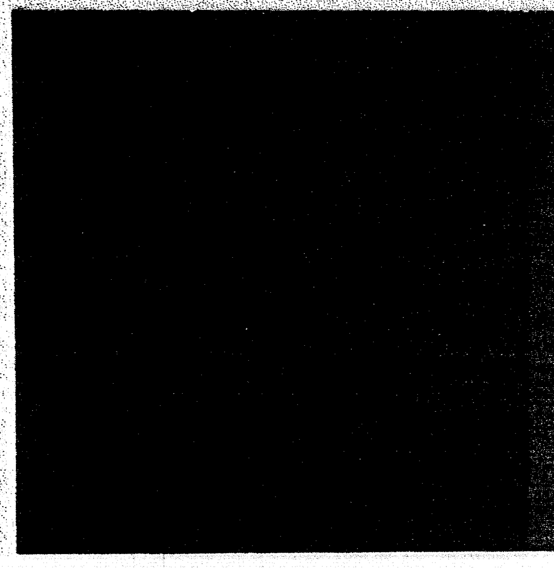
For the smaller mask this does not occur, as the zero-crossing contours tend to outline individual dots or connected groups of dots. Thus the disparities assigned are such that the dots belong to one plane or the other and the final disparity map is one of two separated planes.

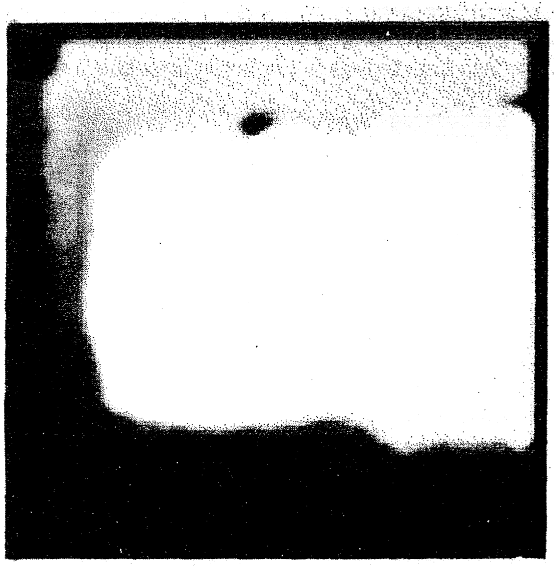
To achieve perfect results from stereo, it is probably necessary to include in the $2\frac{1}{2}$-dimensional sketch a way of dealing competently with discontinuities. Some initial work has already been done in this direction (Grimson, in preparation). Interestingly, when one looks at a 5% random-dot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

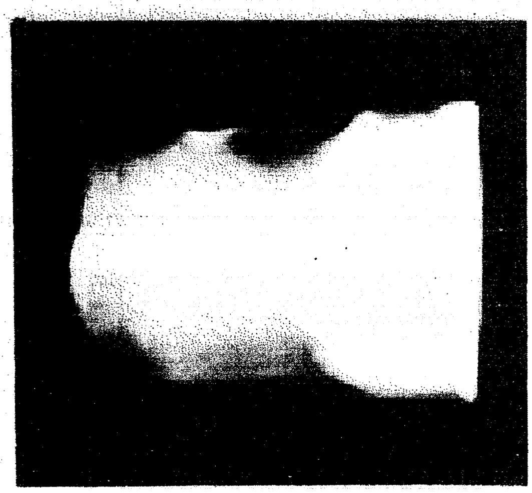
(6) One consequence of the Marr-Poggio theory is that explicit disparity values will be obtained only along the zero-crossing contours. It may be desirable to create a more complete reconstruction of the shapes of the objects in the scene, by filling in disparity values between the zero-crossing contours. Some work has been done in this direction (Grimson, in preparation) and an example is shown in Figure 14.

(7) 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 and

Figure 14. Example of filling in the disparity map. The top left figure is the initial image. The top right figure shows the disparity map associated with the image, where the disparity is represented by the intensity of the point. The bottom figures show the filled in map, again using intensity to represent disparity. In the left figure, the full range of disparity is shown, indicating the slant of the background plane, and the extreme difference in disparity between the jar and the background. In the right figure, the intensities have been adjusted to enhance the disparities of the jar, indicating the general shape of the interpolated surface.





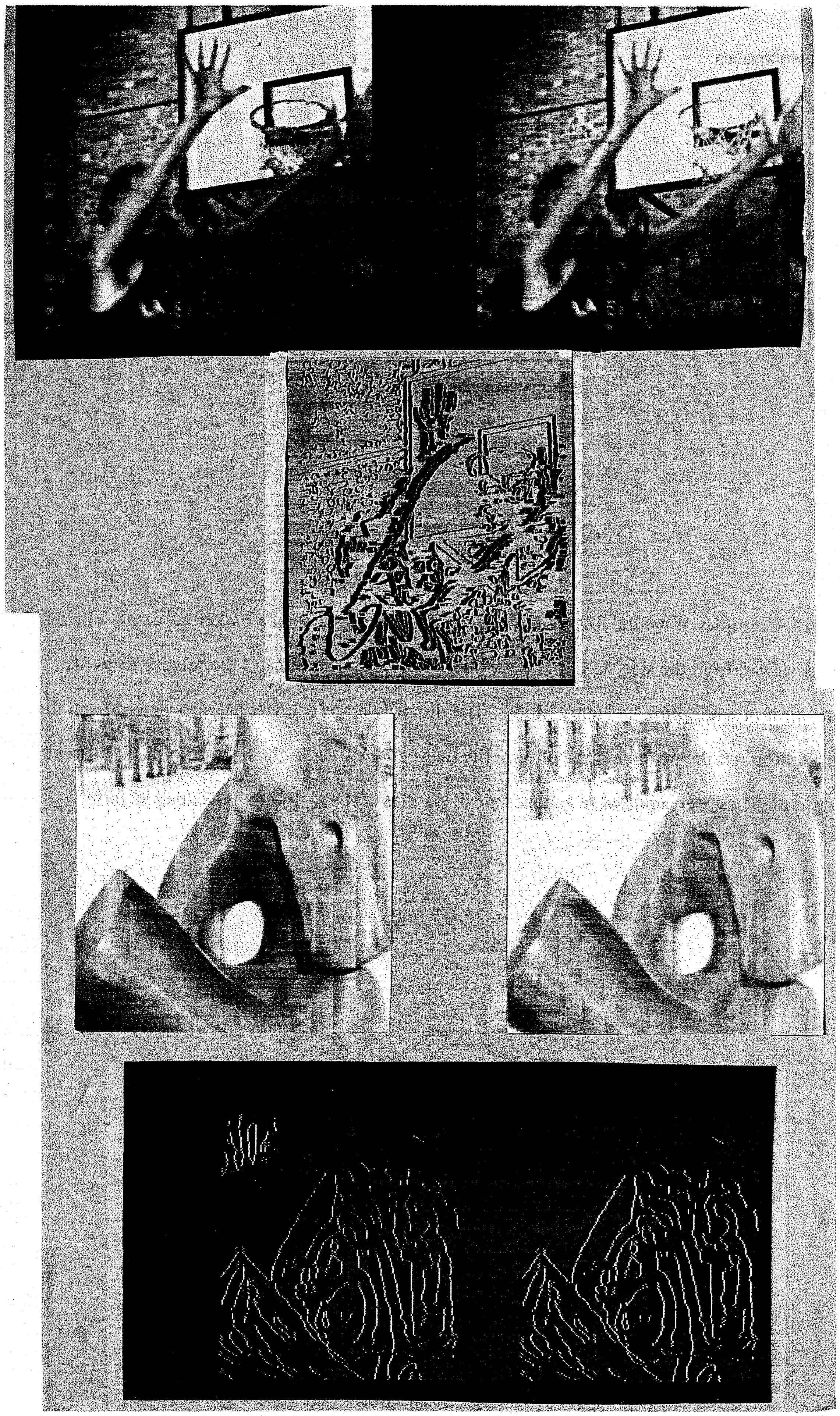


Poggio, 1979, Marr and Hildreth, 1980, Ullman, 1979). The constraints so derived are critical in the formation of the computational theory, and in the design of an algorithm for solving the problem. An interesting question to raise is whether the algorithm explicitly checks that the constraints imposed 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 case of 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 explicity 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 which 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, for matching in this direction. 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 above would indicate that this is not the case. 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.

(8) It is worth observing the distinction between the performance of the implementation on random dot patterns and the performance of the implementation on natural images. Some examples are shown in Figure 15. The main point is that on the whole, the performance is quite acceptable for random dot patterns. However, the implementation can occasionally fail in the case of natural images. The question is whether this reflects a basic inadequacy in the theory and its implementation, or whether there are other aspects of the visual process interacting with stereo which have not been included in this implementation.

This can be approached in two ways: (1) Is the assumption of modularity incorrect? In other words, is there something wrong with the matching module as developed by Marr and Poggio, and as implemented here. (2) Are there other modules, not considered here, which may affect the input or the output of the

Figure 15. Examples of natural images. The top stereo pair is a scene of a basketball game. The disparity map below is viewed from the side, so that the width of the black bars indicates the relative disparity. The bottom stereo pair is of a sculpture by Henry Moore. The disparity maps below it are also viewed from the side. The left map illustrates the extreme range of disparity between the trees in the background and the sculpture itself. The right map has been adjusted to enhance the disparities of the sculpture, indicating its form.



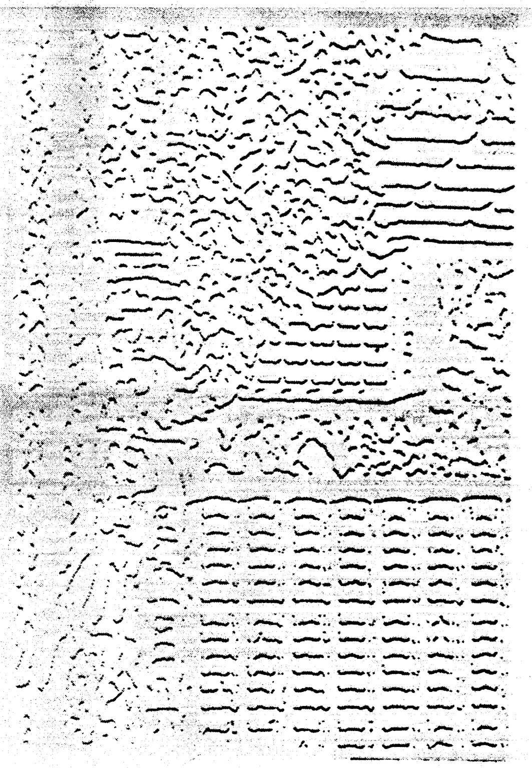
matching module?

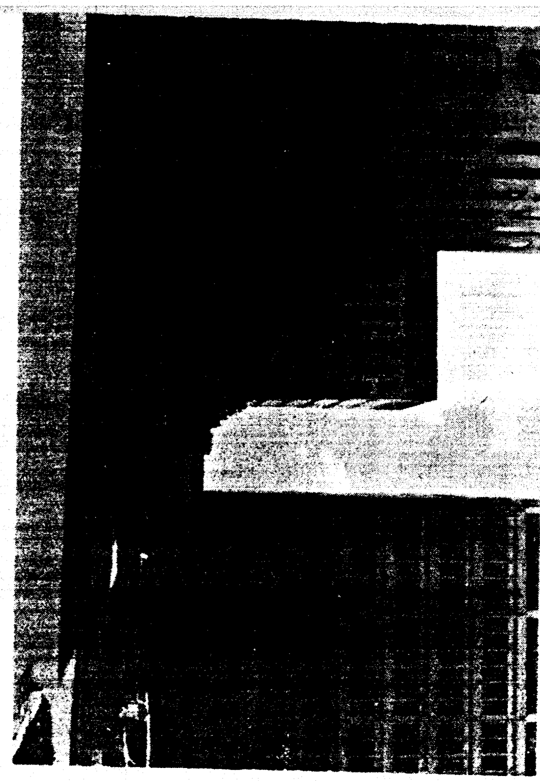
The results of testing the implementation on the broad range of images, indicated in previous sections, seems to indicate that the matching module is acceptable as an independent one. In particular, the agreement between the performance of the algorithm and that of human observers on the many random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

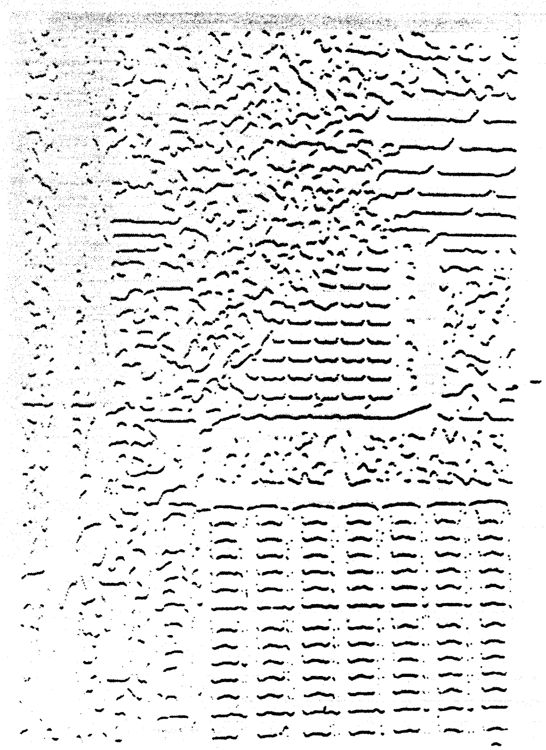
When we turn to natural images, it is reasonable to expect that other visual modules may affect the input to the matcher and that they may alter the output of the matcher. This is not to suggest that the matcher is incorrect, only that the effects of other modules must be taken into account in order to explain the complete human perception. For example, the evidence of Kidd, Frisby and Mayhew (1979) concerning the ability of texture boundaries to drive eye vergence 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 matcher itself needs to be modified.

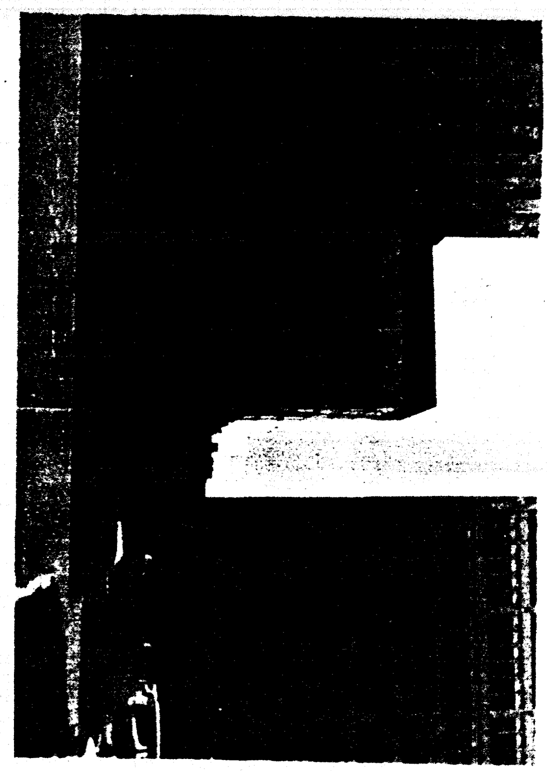
Interestingly, the performance of the implementation supports this point. The implementation, which is considered a distinct module, also performs very well on random dot patterns, where there is no possibility of interaction with other visual processes. For many natural images, this is still true. However, 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 16 . Here, the regular pattern of the windows provides a strong false targets problem. In running the implementation, the following behavior was observed. If the optical axes were aligned at the level of the building, the zero-crossings corresponding to the windows were all assigned a correct disparity. If, however, the optical axes were aligned at the level 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 is the implementation wrong? Curiously, if one fuses the zero-crossing descriptions of the convolved images without eye movements, human observers have the same problem: if the eyes are fixated at the level of the building, the windows are correctly matched; if the eyes are fixated at the level of the trees,

Figure 16. The false targest problem. The top figures are a stereo pair of a group of buildings. The bottom figures show 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.









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 matching and writes it into the $2\frac{1}{2}$-D sketch. However, it is probably the case that some later processing module, which examines the contents of the $2\frac{1}{2}$-D sketch, is capable of altering the contents stored there, based on more global information than is available to the matching component of the stereo process.

Thus, I would suggest that future refinements to the Marr-Poggio theory must account for the interactions of other aspects of visual information processing on the input and output of the matching module. Some initial work has already been done in this direction (Grimson, in preparation).

6. Acknowledgements

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.

7. References

Braddick, O. 1978 Multiple matching in stereopsis. (unpublished MIT report).

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

Grimson, W.E.L. A refinement of a computational theory of human stereo vision *in preparation.*

Grimson, W.E.L. and Marr, D. 1979 A computer implementation of a theory of human stereo vision. *Proceedings: Image Understanding Workshop* 41-47.

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: The University of Chicago Press.

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

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

Knight, T.F., Moon, D.A., Holloway, J., and Steele, G.L. 1979 CADR *MIT Artificial Intelligence Laboratory Memo* 528.

Marr, D. and Hildreth, E. 1980 Theory of edge detection. *Proc. R. Soc. Lond.* (in the press).

Marr, D. and 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. and Poggio, T. 1976 Cooperative computation of stereo disparity. *Science, N.Y.* 194, 283-287.

Marr, D. and Poggio, T. 1979 A computational theory of human stereo vision. *Proc. R. Soc. Lond.* B. 204, 301-328.

Marr, D., Poggio, T. and Hildreth, E. 1979 The smallest channel in early human vision. *JOSA* (submitted for publication).

Mayhew, J.E.W. and Frisby, J.P. 1976 Rivalrous texture stereograms. *Nature, Lond.* 264, 53-56.

O'Brien, B. 1951 Vision and resolution in the central retina. *J. Opt. Soc. Am.* 41, 882-894.

Ullman, S. 1979 *The interpretation of visual motion* Cambridge: MIT Press.

Wilson, H.R. and Bergen, J.R. 1979 A four mechanism model for spatial vision. *Vision Res.* (in the press).

Wilson, H.R. and Giese, S.C. 1977 Threshold visibility of frequency gradient patterns. *Vision Res.* 17, 1177-1190.