MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A. I. Memo 762 January, 1984

Computational Experiments with a Feature Based Stereo Algorithm

W. Eric L. Grimson

Abstract: Computational models of the human stereo system can provide insight into general information processing constraints that apply to any stereo system, either artificial or biological. In 1977, Marr and Poggio proposed one such computational model, that was characterized as matching certain feature points in difference-of-Gaussian filtered images, and using the information obtained by matching coarser resolution representations to restrict the search space for matching finer resolution representations. An implementation of the algorithm and its testing on a range of images was reported in 1980. Since then a number of psychophysical experiments have suggested possible refinements to the model and modifications to the algorithm. As well, recent computational experiments applying the algorithm to a variety of natural images, especially aerial photographs, have led to a number of modifications. In this article, we present a version of the Marr-Poggio-Grimson algorithm that embodies these modifications and illustrate its performance on a series of natural images.

Acknowledgements. 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 under Office of Naval Research contracts N00014-80-C-0505 and N00014-82-K-0334.

1. Introduction

The ability of a sensory system to passively sense the three-dimensional structure of its surrounding environment is frequently a necessary precursor to efficient interactions with that environment, both for biological and artificial systems. A common method for performing this sensing is through stereo vision, and in fact, the human stereo system is remarkably adept at this computation, under a wide variety of conditions. Stereo vision can be characterized by three steps: (1) The point in one image corresponding to the projection of a point on a surface is located. (2) The point in the other image corresponding to the projection of the *same* surface point is located. (3) The difference in projection of the corresponding points is used, together with estimates of the parameters of the imaging geometry (which may be determined solely frmo the correspondences), to determine a measure of the distance to the surface point. While all three steps are important to the process, the second stage has usually been considered the critical one. To deal with this *correspondence problem*, and its concomitant problem of avoiding *false targets* in determining the correct correspondence or match, concern has centered on appropriate representations for matching, and on constraints on the matching process that will ensure the correct correspondence is chosen.

While psychophysical evidence concerning the nature of the human stereo system has been accumulating for some time, recently attention also has been focused on computational investigations of the system. One goal of these investigations has been to consider models of the information processing aspects of the system, independent to a large extent of the specifics of the mechanism that performs the computation. While such models are of importance in understanding the processing of the human system, this relative independence of the algorithm used by the human system and its specific implementation in neural units also suggests that such algorithms may have implications for non-biological applications.

In 1977, Marr and Poggio proposed a feature-point based model of aspects of human stereopsis [Marr and Poggio, 1979]. A computer implementation of their algorithm was then developed and tested [Grimson, 1981a, b]. Initially, the implementation was evaluated on standard psychological test images, in particular, random dot stereograms [Julesz, 1960, 1971]. The intent of this investigation was to demonstrate the adequacy of the Marr-Poggio model for such patterns, and to demonstrate the consistency of the model with known aspects of human stereo perception, including situations in which the system fails. The implementation was also tested on a number of natural images, under a variety of illumination conditions and with a variety of different surface materials. Since the original presentation of the Marr-Poggio model, a number of additional psychophysical predictions of the model have been tested, and consequently, several modifications and improvements have been proposed [e.g. Mayhew and Frisby, 1981; Frisby and Mayhew, 1980; Mowforth, Mayhew and Frisby, 1981; Schumer and Julesz, 1982].

While examining the psychophysical aspects of the model is clearly of importance for perceptual modelling, computational experiments with the algorithm can also provide insights into the information processing aspects of the model. Such experiments are also of importance

when considering applications of the algorithm to domains other than modelling of the human system, as are non-biologically based studies of feature point stereo vision systems [e.g. Arnold and Binford, 1980; Baker, 1982; Baker and Binford, 1981; Barnard and Thompson, 1980; Moravec, 1977, 1980; Ohta and Kanade, 1983 (see also the technique of Kass 1983, 1984, which may also be applicable to feature point stereo)]. Following the original testing of the Marr-Poggio-Grimson algorithm, as reported previously [Grimson, 1981a, b, with some modifications proposed in Marr and Poggio, 1980], extensive additional computational experiments with the algorithm have been performed, especially on natural images. These experiments have led to a number of modifications to the original algorithm, as well as elucidating points that require additional attention. While no inference is made as to the relevance of such modifications for the human system, the modified algorithm may serve as a useful step towards an automated artificial stereo system.

In this paper, we will briefly review the original Marr-Poggio model and outline the previously reported implementation and testing of that algorithm. We will then describe some of the open questions concerning that implementation, as well as some of the modifications suggested by other models [e.g. Mayhew and Frisby, 1981]. A revised algorithm will then be presented. Finally, we will illustrate the performance of the modified algorithm by applying it to a series of natural images. Many of the examples presented are aerial stereo photographs, in part because automated cartography is one of the traditional areas of application of computer stereo algorithms. We also consider an example of a robotics application, and investigate the accuracy of the algorithm in reconstructing the distance to objects in the scene, given measurements for the parameters of the imaging geometry.

2. The Marr-Poggio Stereo Model

In this section, we present a brief review of the original Marr-Poggio model [Marr and Poggio, 1979], its original implementation [Grimson, 1981a, b] and suggested modifications based on psychophysical and computational studies [e.g. Mayhew and Frisby, 1981]. Readers interested in more comprehensive treatments are directed to the original articles.

2.1. The Model

The algorithm proposed by Marr and Poggio for solving the stereo correspondence problem can be described as a feature-point based matching system, using a coarse to fine control strategy to limit the search space of possible matches. As originally proposed [Marr and Poggio, 1979], the algorithm consisted of the following steps.

(1) The left and right images are each filtered with oriented second differential operators of four sizes that increase in size with eccentricity (distance from the center of the eye). The cross-section of these operators is approximately the difference of two Gaussian functions with space constants in the ratio 1:1.75. The purpose of this filtering is to allow the detection of significant intensity changes at multiple scales.

(2) Zero-crossings in the filtered images are located by scanning along lines lying perpendicular to the orientation of the original differential operator. These zero-crossings

mark the locations of significant changes in the original intensity function, at different scales. Positions of the ends of lines and edges are also located.

(3) For each operator size and orientation, matching takes place between zero-crossing segments or terminations of the same contrast sign in the two images, for a range of disparities up to about the width of the operator'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) Disparity information obtained by matching features derived from the larger operators can control vergence eye movements, thus allowing feature from the smaller operators 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 [see also, for example, Moravec, 1980].

(5) When a correspondence is achieved, it is stored in a dynamic buffer, called the $2\frac{1}{2}$-dimensional sketch [Marr, 1978].

2.2. The Original Implementation

The first computer implementation of this model was reported in [Grimson, 1981a] (recently an independent reimplementation of the algorithm has been reported in [Kak, 1983]). The original implementation essentially followed the five steps outlined above, although there were a number of differences. Most of these changes arose from observations made during the process of transferring the model described above to a working algorithm, since the process of explicitly detailing the algorithm illuminated some previously unforeseen difficulties, whose solutions led to modifications to the original model.

The steps in the implementation can be briefly outlined as follows.

(1) **Image Filtering:** The left and right images of a stereo pair are convolved with a series of two-dimensional operators, whose shape is given by the Laplacian of a Gaussian :

$$\nabla^2 G(x,y) = \left[\frac{x^2+y^2}{\sigma^2} - 2\right] \exp\left\{\frac{-(x^2+y^2)}{2\sigma^2}\right\},$$

or by an approximation to this operator, using a difference of two Gaussian functions [Marr and Hildreth, 1980]. These operators are isotropic with respect to orientation, and hence differ from the directional operators proposed in the model. (A discussion of this point may be found in [Grimson, 1981a, b].) The size of the operator, as well as its spatial frequency characteristics, is determined by the value of the constant σ, which is related to the width of the central negative portion of the operator, w, by the following expression:

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

Figure 1 illustrates the form of these operators.

If each picture element (pixel) is considered equivalent to one photoreceptor in the fovea of the human visual system, then we may use psychophysical data obtained from measurements on the human system [e.g. Wilson and Bergen, 1979] to determine the appropriate sizes of operators.

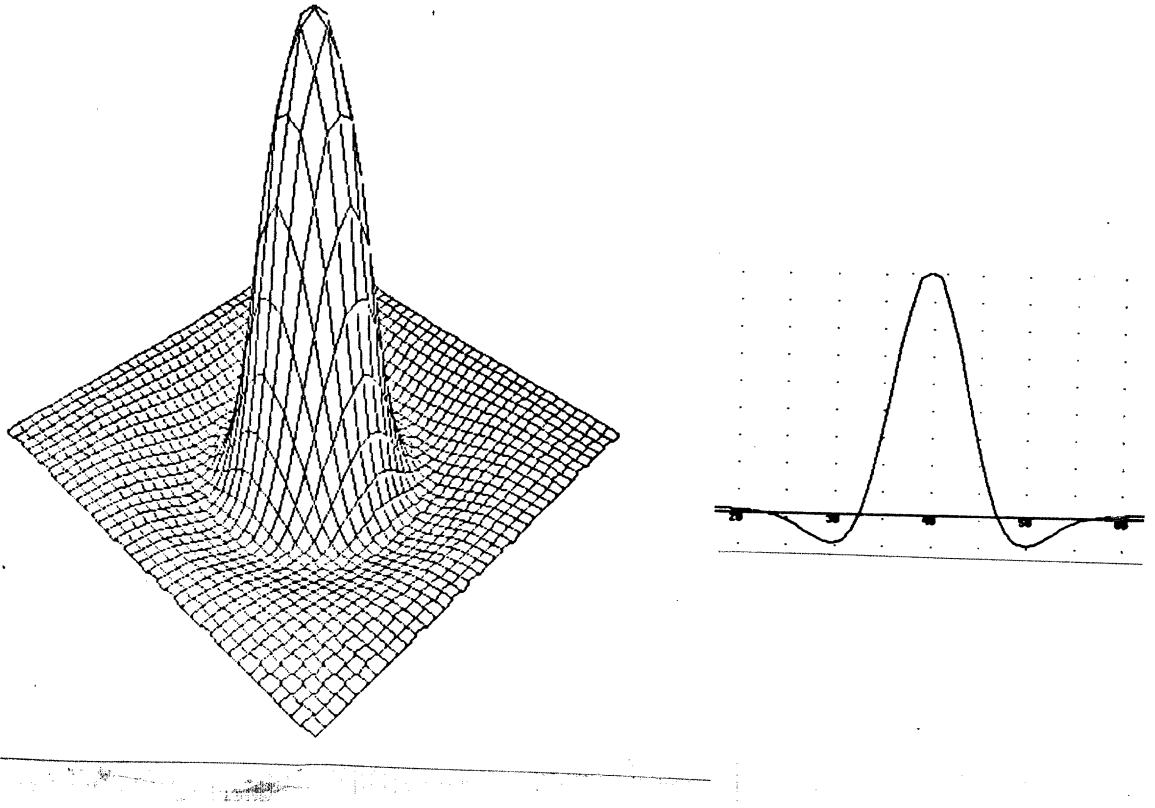

Figure 1. The initial filters. Each image is convolved with a two-dimensional operator whose form is described by a Laplacian of a Gaussian. The size of the operator is determined by the space constant of the Gaussian distribution. Part a shows a perspective plot of a $\nabla^2 G$ filter, part b shows a one-dimensional slice through the center of the filter.

This led us to implement $\nabla^2 G$ operators with widths of $w = 9, 18, 36$ and 72 picture elements (pixels) each. It has also been argued on computational grounds [Marr, Poggio and Hildreth, 1979] and on vernier acuity grounds [Crick, et al. 1980] that an additional smaller operator corresponding roughly to a width of $w = 4$ may also be present in the human system. The coefficients of the operators were represented to a precision of 1 part in 2048. Coefficients of magnitude less than $\frac{1}{2048}$'th of the maximum value of the operator were set to zero. Thus, the truncation radius of the operator (the point at which all further operator values were treated as zero) was approximately $1.8w$.

(2) **Symbolic Features:** In the original Marr-Poggio theory, the elements to be matched between images were (i) zero-crossings whose orientations are not horizontal, and (ii) terminations. It has since been demonstrated [Nishihara and Poggio, 1982] that aspects of human stereo perception previously believed to imply the need for terminations may be explained strictly on the basis of zero-crossings. Thus, terminations are not included in the implementation reported here. It is assumed that the images have been brought into vertical registration, so that the epipolar lines are horizontal. Thus, zero-crossings in the convolved images are found by scanning along horizontal lines, seeking pairs of adjacent elements of opposite sign, or triplets of adjacent elements,

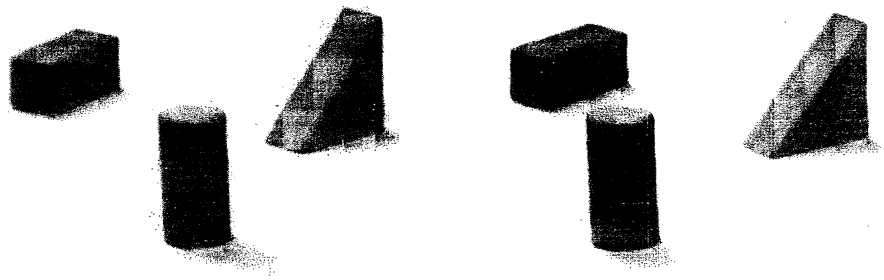

Figure 2. An example of a stereo pair taken in a laboratory setting.

the middle of which is zero, the other two containing convolution values of opposite sign. The positions of the zero-crossings are thus recorded to within an image element. In addition to their location, two other attributes of the zero-crossings were recorded: (1) contrast sign (whether the convolution values change from positive to negative, or negative to positive, as we move from left to right along the scan line) and (2) a rough estimate of the local orientation in the filtered image of segments of the zero-crossing contour. In the original implementation, the orientation of a point on a zero-crossing contour was computed as the direction of the gradient of the convolution values across that segment, and was recorded in increments of 30 degrees.

Examples of the convolutions and zero-crossings for a series of operators are illustrated in Figures 2, 3, and 4.

We note that while the positions of the zero-crossings are specified to within a pixel, it may be possible to perform subpixel localization. Hildreth [1980] (see also [Crick, et al., 1980] has demonstrated that in the case of an isolated zero-crossing, a simple linear interpolation between convolution values serves to localize the zero-crossing to subpixel precision [see also, MacVicar-Whelan and Binford, 1981]. It has been observed in computational experiments that strong isolated zero-crossings, such as those corresponding to occluding boundaries or shadows, for example, can be reliable matched to subpixel precision. In the presence of texture or other confounding photometric effects, however, the accuracy of the subpixel localization decreases, and is probably not effective. This raises an interesting question about human stereo acuity. It suggests that for stimuli with isolated zero-crossings, (for example, line drawings), stereo acuity could lie within the subpixel range [Howard, 1919; Woodburne, 1934; Berry, 1948; Tyler, 1977], but for textured stimuli, (for example, random dot stereograms), stereo acuity might be expected to decrease.

(3) **Matching:** Given a set of zero-crossing representations at different scales for each of the images, the matching process proceeded in a coarse to fine iterative manner. The idea [first used by Moravec, 1977, 1980] is to use a sparse representation of the images, with a coarse spatial sampling,

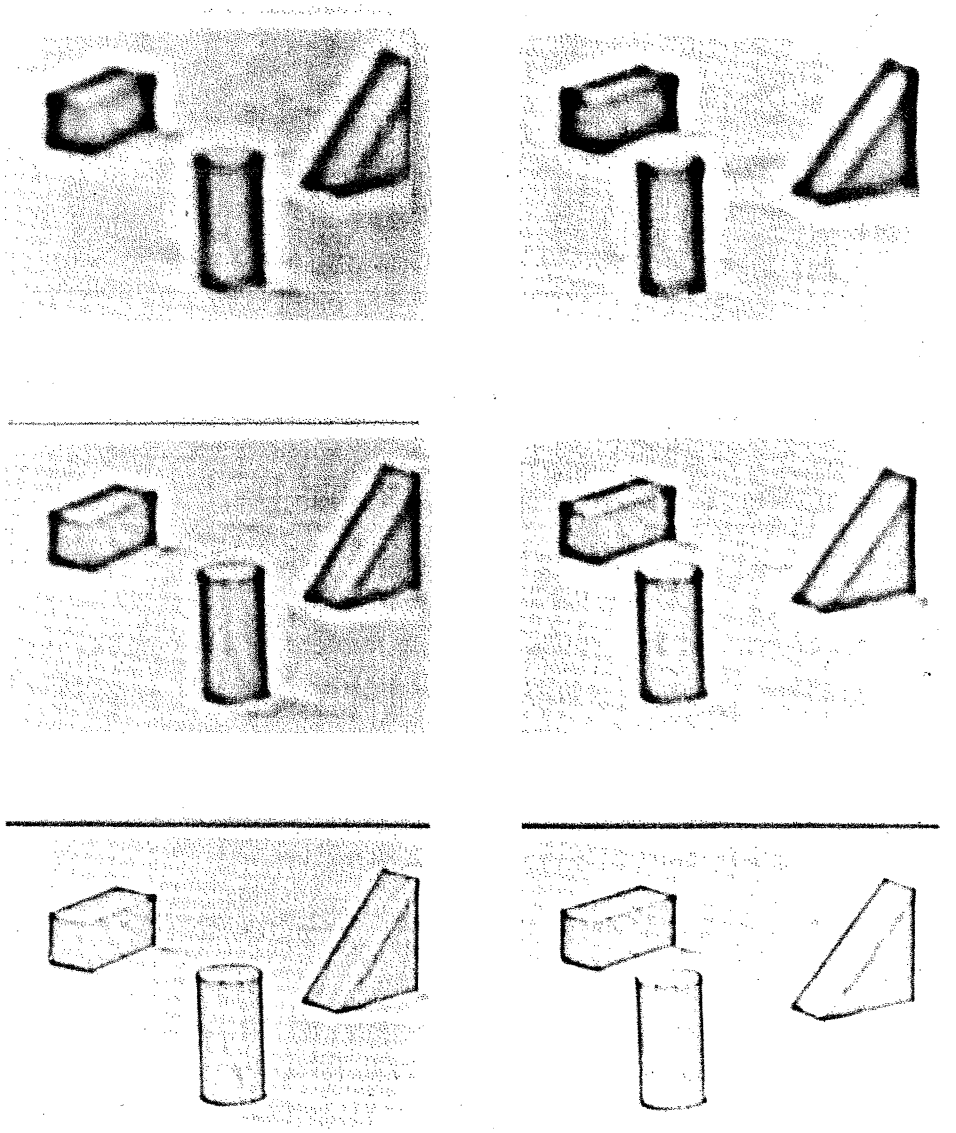

Figure 3. Convolutions of the blocks images.

for the initial matching of points. The reduced density of points greatly reduces the search space and makes matching easier, at the expense of reduced resolution. This initial match can then be used to constrain the matching of finer detailed representations, again reducing the search space of the matching process, while allowing finer detail disparity information to be obtained. Thus, the matching is guided by a flow of information from coarse representations to finer ones.

(3.1) **Feature Point Matching:** Consider first the zero-crossing representations obtained from the coarsest filters (with central width w_c), and suppose that we are given some estimate d_i of the disparity in a region of the image (which we may initially assume to be some arbitrary value). For a zero-crossing in one image (say the left) at position (x, y), the search for a matching zero-crossing

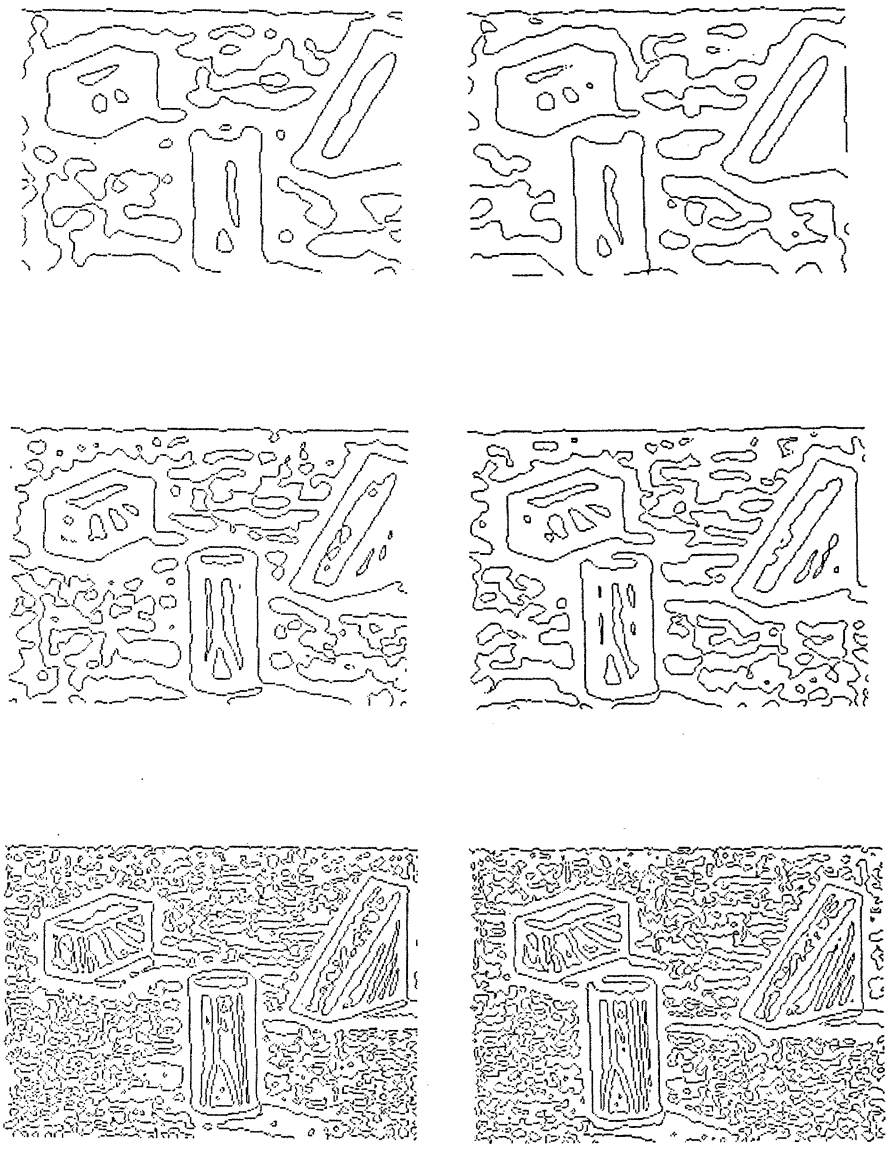

Figure 4. Zero-crossings of the blocks images.

in the right image is constrained to the region

$$\{(x', y) \mid x + d_i - w_c \leq x' \leq x + d_i + w_c\}.$$

(Note that the search takes place along the same horizontal scan line, thereby assuming that the images have been registered so as to yield horizontal epipolar lines.) This $\pm w_c$ range in the right image is divided into three pools, two larger convergent and divergent regions, and a smaller one lying centrally between them. For each pool, matching zero-crossings in the left and right filtered images must have (1) the same contrast sign, and (2) roughly the same orientation.

A match is assigned on the basis of the responses of the pools. If exactly one zero-crossing of the appropriate sign and orientation (within 30 degrees) is found within a pool, its location

is transmitted to the matcher. If two candidate zero-crossings are found within one pool (a very unlikely event [see, for example, Grimson, 1981b]), the matcher is notified and no attempt is made to assign a match for the point in question. If the matcher finds a single zero-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 neighborhood about the point in question and recording the sign of the disparity of the unambiguous matches within that neighborhood. (The sign of the disparity refers to the sign of the pool from which the match comes: divergent, convergent or zero.) If the ambiguous point has a potential match of the same sign as the dominant type within the neighborhood, then that is chosen as the match. Otherwise, the match at that point is left ambiguous.

(3.2) **Continuity:** It is possible that the region under consideration does not lie within the $\pm w_c$ disparity range examined by the matcher. This is detected and handled by the following operation. If the region does lie within the disparity range $\pm w_c$, then excluding the case of occluded points, every zero-crossing in the region will have at least one candidate match in the other filtered image. On the other hand, if the region lies beyond the disparity range $\pm w_c$, then the probability of a given zero-crossing having at least one candidate match will be roughly 0.7 [Marr and Poggio, 1979; Grimson, 1981a, b]. Thus, by counting the percentage of zero-crossings within a region that have at least one match, and thresholding based on the probabilities stated above, disparities will be accepted only in regions lying within the current disparity range. This constraint is based on the continuity assumption [Marr and Poggio, 1979] that surfaces generally vary in a smooth manner relative to the viewer.

(3.3) **Control Strategy:** Finally, once this matching has been performed for the coarsest filter, the sparse disparities obtained can be used to realign the images, and the process can be repeated at the next finer scale. Since the density of zero-crossings increases as the size of the filter is decreased, this coarse to fine control strategy allows the matching of very dense zero-crossing descriptions with greatly reduced false target problems, by using coarser resolution matching to drive the alignment process.

(3.4) **Vertical Disparity:** While the matching as described above only searches for corresponding zero-crossing points along the same horizontal scan lines, the control strategy of the algorithm can easily be modified to handle small amounts of vertical disparity. First, note that due to the size of the $\nabla^2 G$ filters, the coarser level zero-crossing representations are less sensitive to local vertical disparity than the finer level ones. Now suppose that the matching has been performed for the coarsest filter and that the horizontal and vertical disparity in a region of the image is roughly given by d and v respectively. When proceeding to a finer filter, the search for matching zero-crossings is initially centered about this disparity. If, however, the density of zero-crossing points that can be matched at this level is small, it is likely that the horizontal disparity is nearly correct, but that the vertical alignment is in error. Thus, reapplying the matching process with

the same horizontal alignment, d, but with small variations (on the order of several lines) in the vertical alignment, $v \pm \epsilon$, will lead to a correct alignment of the images, and hence to a greater density of zero-crossings being assigned valid disparity values.

2.3. Testing of the Original Implementation

As reported in [Grimson, 1980, 1981], this implementation of the Marr-Poggio algorithm has been tested on a variety of images. Much of the original testing was performed on random dot stereograms, for two reasons. First, because the stereograms are synthetically created, it is possible quantitatively to compare the disparities computed by the algorithm with the physically correct disparities. Second, because random dot stereograms are a standard psychological method for examining attributes of the human stereo system, the performance of the algorithm on such test cases could be compared to human perception, providing a means of examining the adequacy of the underlying model. Examples of the testing included two-planar stereograms of varying densities, more complex figures such as a wedding cake and a spiral staircase, stereograms in which one or both images had been blurred, stereograms with added spatial frequency filtered noise, stereograms in which one of the images had been decorrelated by different amounts, and stereograms in which one of the images had been compressed. It was found that on the standard random dot stereograms, the matching algorithm performed very well, usually with an error rate of less than one part in a thousand. On noisy or decorrelated stereograms, the error rate was normally on the order of one percent, while the density of points to which a disparity was assigned decreased (and in the limit vanished).

The implementation was also tested on a number of natural images, using a variety of illumination geometries and with objects of differing photometric characteristics. Examples included a speckled coffee jar, a basketball game, an outdoor metallic sculpture, and a portion of the Martian surface. For these natural images, a quantitative evaluation was more difficult to obtain, precisely because the imaging geometry was not controlled, but it was observed that the qualitative performance of the algorithm was still good.

2.4. Discussion

While the initial testing of the algorithm did serve to support the adequacy of the Marr-Poggio algorithm as a model of aspects of the human stereo system, and while the overall performance of the matching algorithm was very good, a number of weak points in the algorithm were illuminated during this testing.

2.4.1. Continuity constraints

It was observed that most of the actual matching errors occurred along discontinuties in depth, for example at occluding boundaries between two objects. This follows from the use of matching statistics over a region as a means of distinguishing correct matches from random ones. Theoretically, this test is based on the observation that surfaces are generally smooth relative to the observer, and hence disparity will generally also be smooth. While the theoretical observation is sound, the implementation of it by means of a statistical measure over a region of the image has some difficulties.

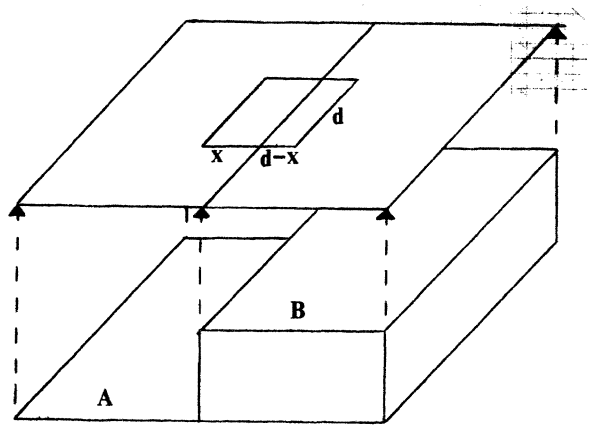

Figure 5. The problem of the continuity constraint near object boundaries.

This is most easily illustrated by the following example. Suppose the region over which the matching statistics are measured is a square of side d (while this is the easiest to implement, it is not critical and the following argument holds for other shapes as well). Further suppose that the stereogram consists of two planar surfaces with a sharp break in disparity between them. Let the density of zero-crossings be ρ and presume that the region is positioned such that $\frac{x}{d}$ percent of the region covers surface A and that $1 - \frac{x}{d}$ percent covers surface B (see Figure 5). Finally, assume that the fixation of the eyes is currently positioned on surface B, so that the portion of the region covering the surface A is out of range of the matching process. If ϵ is the threshold for accepting the matches in a region as being within the range of the matcher, (for the analysis of Marr and Poggio [1979, p317] $0.7 < \epsilon \leq 1.0$), then the question to consider is for what values of x the percentage of matched points in the region will exceed ϵ.

In theory, the number of matched points in the surface B region is expected to be $\rho d(d-x)$, and the number of matched points in the surface A region is expected to be $0.7\rho xd$. Thus, the percentage of matched points is given by

$$\frac{\rho d(d-x) + 0.7\rho xd}{\rho d(d-x) + \rho xd} = 1 - 0.3\frac{x}{d}.$$

The values of x for which this percentage exceeds ϵ is given by

$$x \leq \frac{1-\epsilon}{0.3}d.$$

The most conservative threshold would be $\epsilon = 1$, in which case $x = 0$ and the only position of the region for which the disparity values are accepted as correct is that in which the region is entirely positioned over surface B. While this would work on perfect data, in practice it is likely to be overly conservative, causing a large reduction in the percentage of zero-crossings to which a disparity is assigned, although the error rate should be virtually zero. One difficulty with real data is that even for regions of the image whose disparities are completely within range of the matcher, the zero-crossing points may not all have matches. For example, geometric distortion in the sensors, perspective distortions in the imaging geometry, noise in the irradiance values and local

photometric effects all can cause slight variations in the zero-crossings that may result in a small number of unmatched points. Rather than discard all the disparity information in a region because a single zero-crossing point does not have an assigned match, we would like to preserve such information, by using a less conservative threshold. Consider, however, the compromise case of $\epsilon = 0.85$. In this case, the constraints on the positioning of the region are given by $0 \leq x \leq 0.5d$, and in this case, any (incorrect) disparity values lying within $0.5d$ pixels of the edge of surface B will be accepted as correct. This is observed in examples of the testing of the algorithm, and while the number of such errors is small, it is unavoidable within the context of this type of statistical check. This problem will be very apparent in the case of thin elongated surfaces suspended above a background, where the widths of the surfaces are less than the diameter of the statistics region, for example, in an aerial stereo image of a highway interchange.

One means of overcoming this problem is to observe that while it is difficult to ensure that a region of the image corresponds strictly to a single surface, edges (or zero-crossings) in a filtered image will generally correspond to a single surface, since they usually reflect changes in the surface topography or the surface photometry. Thus, rather than imposing a condition of disparity continuity over an area of the image, one could instead require a continuity of disparity along a contour in the filtered image. This is essentially the *figural continuity constraint* of Mayhew and Frisby [1981], and has been suggested in a slightly different form in Arnold and Binford [1980]. Thus, we need to derive a contour based analog to the regional continuity check used in the original Marr-Poggio implementation.

Once the feature points have been matched, it can be observed that the collection of all matched points is composed to two distinct sets. In regions of the image where the zero-crossing representations lie within matching range of the current image alignment, the matched feature points tend to form extended contours. Elsewhere, the matched feature points tend to lie in scattered small segments. The goal of the figural continuity constraint is to distinguish between these two situations.

We now derive an explicit form for the constraint. We know, by applying Rice's theorem [Grimson, 1981b, p. 78], that the expected distance between zero-crossings of the DOG filter of the same contrast sign is given by

$$s = \frac{5.29w}{2\sqrt{2}}.$$

Then given uncorrelated left and right zero-crossing descriptions, the probability of no match at a particular disparity is

$$1 - \frac{1}{s},$$

and if p denotes the horizontal width of a matching pool, and v denotes its vertical extent, the probability of no match within a pool of dimensions $p \times v$ is

$$\left(1 - \frac{1}{s}\right)^{pv},$$

and hence the probability of a match in this pool is

$$\rho = 1 - \left(1 - \frac{1}{s}\right)^{pv}.$$

Now we consider the probability of randomly matching segments of a contour. Given a contour segment of length k in one image, we want to determine the probability that m of those k points has a match within the corresponding pool in the other image, when the two images are uncorrelated. Clearly, this is given by

$$P_{k,m} = \sum_{i=0}^{k-m} \binom{k}{i} \rho^{k-i}(1-\rho)^i. \tag{1}$$

Thus, given some threshold, ϵ, on the expected error rate, such that $0 \leq \epsilon < 1$, we can determine constraints on the length of a matched zero-crossing contour that will be accepted as corresponding to a correct match. That is, given a threshold ϵ, and a value for the number of unmatched gaps in the contour, $k - m$, we can find the minimum length k of a contour such that $P_{k,m} < \epsilon$. In particular, we let

$$\ell_j = \min\{k \mid P_{k,k-j} < \epsilon\}$$

denote the threshold on the length of matched contour required to satisfy the figural continuity constraint, for some number of gaps. Note that this is a function of the expected error threshold ϵ, as well as the horizontal pool size p, the vertical pool size v, and the mask size w.

Thus we have derived a specific form for the figural continuity constraint, namely that the length of contour that must be matched, as a function of the error threshold, as well as the parameters listed above is given by the values of ℓ_j.

2.4.2. Vertical disparity

One of the implicit assumptions of the Marr-Poggio algorithm is that the geometry of the two sensors yields horizontal epipolar lines. While it is possible to rectify the images to remove gross geometric distortions caused by factors such as cyclotorsion and camera tilt, there are likely to be local distortions of the epipolar geometry, due to geometric distortions in the sensor, or perspective effects. Furthermore, the discrete nature of the zero-crossing representation may cause small variations (on the order of a pixel) in the positions of the zero-crossings. These factors suggest that although large scale effects on the epipolar geometry can be handled by some type of image rectification, there may still be small scale variations on the epipolar geometry that must be handled by the matching algorithm.

In light of this discussion, it is interesting to note recent evidence concerning the effect of vertical disparities on the human stereo system. It has been observed psychophysically [Duwaer and van den Brink, 1981a, 1981b] that while up to a degree of vertical disparity can be tolerated by the human stereo system, almost all of this is handled by invoking an eye movement to align the images. In the absence of eye movements [Nielsen and Poggio, 1983], only about 2–4 minutes of vertical disparity can be tolerated. One interpretation of these results is that the stereo matching mechanism is capable of performing the correspondence process only if the images have been nearly rectified, and that grosser distortions of the epipolar geometry are corrected for by changing the alignment of the eyes.

Interestingly, the original implementation of the Marr-Poggio algorithm essentially incorporated this effect in the following manner. Initially, the vertical disparity was assumed to be zero (although if monocular cues were incorporated into the system, it would be possible to precompute a less arbitrary vertical alignment of the images [Marr and Poggio, 1980]), and the matching was performed at the coarsest resolution. Because of the large size of the filter, the effects of vertical disparity in the images is less likely to affect the performance of the matcher. Suppose we consider some region of the image, and use the disparity information computed by the coarse filter to align the images. If the finer filtered images cannot be matched (or can be only very sparsely matched), this can be taken as an indication that the images have been correctly aligned to remove any horizontal disparity, but that a small amount of vertical disparity may be present. Thus, by applying small alignment corrections in the vertical direction, the images can be brought into alignment, thereby increasing the density of computed disparity values. This behavior was observed in computational experiments on a number of natural images.

Although the performance of the Marr-Poggio-Grimson implementation was qualitatively consistent with the psychophysical data, the use of a stringent epipolar matching geometry was probably too strict. In other words, while it is feasible to use gross alignments of the images to account for large scale geometric effects, a strict epipolar matching strategy may be too sensitive to small local distortions in the zero-crossing descriptions, either due to geometric or perspective effects, due to noise in the early processing, or due to discretization effects. As a consequence, it is suggested that the matching of zero-crossings be relaxed slightly. (Note that in the original Marr-Poggio algorithm, the use of oriented filters suggests that vertical disparity effects would be more tolerable.) For example, suppose there is a zero-crossing at some point (x, y) in the left image. The initial Marr-Poggio implementation would search for a corresponding zero-crossing in the region

$$\{(x', y) \mid x + d - w \leq x' \leq x + d + w\}$$

in the right image. Instead, we propose to search for a corresponding zero-crossing in the region

$$\{(x', y') \mid x + d - w \leq x' \leq x + d + w; \quad y - \epsilon \leq y' \leq y + \epsilon\}$$

where ϵ is on the order of 1 or 2 scan lines. Note that while this will make the matcher less sensitive to small distortions or noise, it will also reduce the accuracy of the matching process, since a single zero-crossing point in one image could potentially be matched to all the points on a zero-crossing segment lying within this window in the second image, yielding a small range of disparity values, rather than a single one. The effect will become more noticeable as the orientation of the zero-crossing segment approaches horizontal.

We also note, while discussing vertical disparity, that several authors have recently proposed using measured vertical disparities to obtain the additional camera parameters needed to convert disparity directly into distance [Mayhew, 1982; Longuet-Higgins, 1982; Mayhew and Longuet-Higgins, 1982; Prazdny, 1982, 1983]. While the algorithm described here does not use the vertical isparity information is this manner, it is possible to augment the algorithm to do so.

2.4.3. Control strategies and search spaces

Finding the correspondence between points in the two images can be considered as a problem of searching a space of possible correspondences for the correct solution. In considering this type of formulation, two separate issues must be considered.

1. Restricting the set of possible alternatives. The key point is to improve the reliability of the computation, by attempting to ensure no false positives, and as few false negatives as possible, i.e. no incorrect matches, and as few cases of no answer as possible.
2. Strategies for efficiently searching the space of alternatives to find the correct one.

We wish to separate these two issues, since while they are related, techniques used to reduce the space of possible correspondences need not be inextricably tied to particular strategies for searching for those correspondences.

First, we consider means for reducing the space of alternatives that must be explored in order to find the correct correspondence. Assume that each image is $n \times n$. Then initially each point in one image has n^2 possible matches. As well, there are n^2 points in each image, so a straightforward, British Museum style, search algorithm requires n^4 total comparisons. How can we reduce this?

Feature point systems, while suffering a reduction in the density of computed depth values, can significantly reduce the space of possible correspondences, by attempting to restrict the computation to "distinguishable" points in the images. If the density of feature points is ρ, then the set of possible matches becomes ρn^2 and the number of total comparisons under the British Museum algorithm is $\rho^2 n^4$. Note that in the case of the Marr-Poggio algorithm, ρ varies with the size of the initial filter. In particular, the expected density of zero-crossings is

$$\frac{1}{cw}$$

where

$$c = \frac{2\sqrt{2}}{5.29} = 1.87$$

by the analysis of [Grimson, 1981, p.78]. Thus, the number of possible candidates for a correspondence reduces to

$$\frac{n^2}{cw}$$

and the total number of comparisons involved in the search is

$$\frac{n^4}{c^2 w^2}.$$

The next major constraint that can be applied to the matching process is the epipolar one. If we take a liberal interpretation of this constraint, then a point on line y can be matched only to points on lines v' such that $y - v \leq v' \leq y + v$, for some constant v. In this case, each point has a space of possible matches on the order of

$$\frac{(2v+1)n}{cw}$$

and the total number of comparisons over the whole image is

$$\frac{(2v+1)n^3}{(cw)^3}.$$

The final matching constraint used in the Marr-Poggio algorithm is that of continuity, which is intended to reduce the number of possible matching candidates from order n to 1. Of course, one can clearly construct situations in which the number of matching candidates is not reduced to a unique solution, but in general, as the discussion in the previous section indicated, the continuity constraint can be structured so as to reduce the probability of false matches to virtually zero.

Note that all of the constraints introduced in this discussion have been matching constraints, that is, they have reduced the number of possible matches for a given point. As a consequence, the total size of the search space has also been reduced, but it is important to note that all the discussion to this point has been independent of the particular search strategy to be employed in finding corresponding matches. This distinction between the use of matching constraints to alter the space of possible correspondences, in order to ensure the existence of a unique solution, and the use of efficient techniques for searching the space of solutions to find the correct solution, is important in light of the final constraint of the Marr-Poggio algorithm, the use of multiple resolution representations of the image.

One use of multiple resolution representations is in dealing with false targets. For example, if a fine resolution feature point representation has more than one possible match for a particular point, the correspondence information at a lower resolution representation can be used to resolve this ambiguity. This was one of the main uses of multiple resolution representations in the original Marr-Poggio algorithm. This disambiguation technique was also intertwined with an efficient search algorithm as well, however. In particular, the matching of finer level representations is directly driven from coarser level correspondences (whenever possible). Not only does this provide one means of avoiding false targets, but it is also an extremely efficient method for searching the space of possible matches, as is indicated in the following discussion.

Let w_0 denote the size of the smallest image filter, and assume that we have $k+1$ such filters, each one doubling in size from the previous one. Then, by the discussion above, we know that at the coarsest level, we must search on the order of

$$\frac{n^2}{c2^k w_0} \cdot \frac{(2v+1)n}{c2^k w_0} = \frac{(2v+1)n^3}{c^2 2^{2k} w_0^2}$$

alternatives in order to find correspondences for all the feature points in this level of representation. If the matching process is driven in a coarse-to-fine manner, then at each subsequent level, the image representations are aligned based on previous matching, and for each feature point, we need only search an area of size cw to find the correct match. Thus, in principle, we need only compare

$$\frac{(2v+1)cw}{cw} = (2v+1)$$

points. This implies that at each of the subsequent levels, we must search $2v+1$ comparisons for each of

$$\frac{n^2}{2^i c w_0}$$

feature points. Thus, the total number of comparisons needed is on the order of

$$\frac{(2v+1)n^3}{c^2 w_0^2 2^{2k}} + \sum_{i=0}^{k-1} \frac{n^2(2v+1)}{2^i c w_0}$$

points, or equivalently,

$$\frac{(2v+1)n^2}{cw_0}\left[2 - \frac{1}{2^{k-1}} + \frac{n}{cw_0 2^{2k}}\right]$$

points. This is still $O(n^3)$ but as k increases, we see that the amount of search involved in finding feature point correspondences reduces to the order of the dimensions of the image, i.e. n^2. Thus, one of the advantages of multiple level representations, besides its use in disambiguation of false targets, is its efficiency in finding the correspondences especially in situations, such as the human visual system, in which high resolution information is only required over small portions of the image at any one time. (Compare this estimate of $O(n^2)$ pointwise comparisons with the results of [Ohta and Kanade 83] of $O(n^5)$ primitive computations for a general 3-D search algorithm and $O(n^3)$ primitive computations under certain limiting assumptions.)

It is curious to note as an aside that one could use the above expression to predict the number of levels of representation (or equivalently, the number of $\nabla^2 G$ filters) needed to reduce the search space to $O(n^2)$. If we consider an area spanning 8^o on a side with foveal-level receptor spacing, then a straightforward calculation predicts that 6 filters are necessary to reduce the search space to $O(n^2)$. Interestingly, recent investigations by Wilson [1983] provide evidence for 6 such filters.

If the key consideration is not speed, but rather, high resolution depth information at all points in the image, it is possible to propose an alternative search strategy, while still taking advantage of the disambiguation properties of multiple resolutions representations. Rather than driving the matching process directly from the coarse level information, we can instead use that information only when needed for disambiguation.

As in the original Marr-Poggio algorithm, for any given alignment of the images (fixation of the eyes), the search space is restricted to a range on the order of cw, so as to avoid the possibility of false targets. Any candidates that satisfy all the matching constraints are accepted as possible correspondences, and stored away. If the total range of disparity over the entire image is within this cw range, then we are done. If not, however, then the same matching process is repeated at some desired spacing in depth, and the algorithm is swept across the entire range of disparity. While for each given alignment of the images, only one match is possible, it may be the case that matches for the same feature points will be found at very different alignment positions. If this is the case, then this false targets problem can be disambiguated by choosing the alternative that best agrees with the correspondence information obtained at coarser levels. Clearly, such a search algorithm requires a sweeping of fixation across the entire range of depths, and while it will result in high resolution depth information everywhere in the image, it does so at the expense of speed.

3. A Modified Marr-Poggio Stereo Matcher

We have incorporated all of these considerations into a new algorithm, which we describe below. While the modifications were made in part because of recent psychophysical evidence concerning the human stereo system, we will discuss its possible merits as a stereo system for such applications as automatic aerial cartography and robotics in the next section.

3.1. The Modified Algorithm

We will first outline the basic algorithm, and then provide more detailed descriptions of each of the steps. The basic steps of the matching algorithm can be summarized in the following manner. Note that steps 0-3 are identical to the original algorithm. The main concentration on modifying the algorithm has been at the matching stage. Also note that steps 4.1-4.3 are an instance of Marr's *principle of least commitment* [Marr, 1982].

3.1.1. Outline of the Algorithm

(0) **Loop over levels:** We initially choose the coarsest level of representation, i.e. the one corresponding to the largest image filter, and iterate by choosing successively finer levels of representation.

(1) **Convolution:** Given a level of representation, the left and right images are convolved with the $\nabla^2 G$ filters of the corresponding size.

(2) **Zero-crossings:** Given the convolved images, the nontrivial zero-crossings are located and marked with their contrast signs. These zero-crossings descriptions form the basic representations from which correspondences will be sought.

(3) **Loop over fixation position:** The relative alignments of the two images are choosen. The simplest method is to initially choose an alignment corresponding to some lower limit on the disparity of the images, and slowly increment this offset until some upper limit on the disparity is reached. This increment could be a pixel at a time, or in terms of some larger fraction of the width of the matching area for a given fixation position.

(4) **Matching:**

(4.1) *Feature point matching:* Given a pair of zero-crossing representations, from the current level, and given a fixation position defining the relative alignments of the two images, feature point matching is applied. For each feature point in one zero-crossing description, this involves searching an area of the other zero-crossing description for a zero-crossing of the same contrast sign. This area has a vertical extent about the same horizontal line in the other image that is limited to a small number of scan lines, and a horizontal extent, of width defined by the size of the underlying image filter, about the same position in the other image, offset by the current relative alignment.

(4.2) *Figural continuity:* Once all the feature points have been matched for the current level of representation and the current fixation alignment, figural continuity constraints are applied to prune the incorrect matches. This involves tracing the zero-crossing contours, searching for

contiguous matched segments of those contours whose length exceeds a threshold whose value can be determined *a priori* from the properties of the underlying $\nabla^2 G$ filters.

(4.3) *Disparity map update:* Any matched feature point contours which pass the figural continuity test are then added to disparity map, recording the relevant disparity for each feature point in the accepted contour segments.

(5) **Loop:** Once this computation of disparities within the defined range about the current image alignment has been completed, the fixation position is updated by looping to step (3).

(6) **Disambiguation:** When all the fixation positions have been processed, we are left with a disparity map representation that contains all matched zero-crossing segments, with their associated disparities. We now check this map for possible double matches. Any such ambiguities are resolved by checking the disparities within the same region of the representation at the previous level (if there is one) and accepting only those disparity values at the current level that are consistent with those values (i.e. lie within a predefined range of the coarser level disparities). If this disambiguation does not succeed, either because there is no coarser level, because there are no disparity values within the same image region at the coarser level, because none of the current level disparities lie within range of the coarser level ones, or because more than one of the current level disparities are consistent with coarser level disparities, then all the alternatives are discarded.

(7) **Loop:** Once the final disparity map for the current level has been completed, the process proceeds to the next finer level of representation, by looping to step (0).

(8) **Consistency:** When all the levels of disparity information have been computed, one final test is possible. Each disparity value at the finest level of representation can be tested for consistency by checking that, within the same region of the previous disparity representation, there is at least one disparity value that is consistent with the current value.

3.1.2. Detailed description of the algorithm

We now turn to a more detailed description of the different stages of the algorithm.

(1) **Convolutions:** As in the previous implementation, convolve the images L, R with $\nabla^2 G(w)$ filters, for different values of w. For notational convenience, we let

$$LC_w(x,y) = \nabla^2 G(w) * L$$
$$RC_w(x,y) = \nabla^2 G(w) * R$$

denote the left and right convolutions, that is, for different widths w, the convolved image forms a two-dimensional array indexed by x and y. Generally, we use only 3 or 4 values of w, for example, $w = 5, 9, 17, 33$ pixels.

(2) **Zero-Crossings:** As in the previous implementation, compute the zero-crossings of the convolved images. We let

$$LP_w(x,y) = \text{positive zero-crossings of } LC_w(x,y)$$
$$LN_w(x,y) = \text{negative zero-crossings of } LC_w(x,y)$$
$$LH_w(x,y) = \text{horizontal zero-crossings of } LC_w(x,y)$$
$$LZ_w(x,y) = \text{all zero-crossings of } LC_w(x,y)$$
$$RP_w(x,y) = \text{positive zero-crossing of } RC_w(x,y)$$
$$RN_w(x,y) = \text{negative zero-crossings of } RC_w(x,y)$$
$$RH_w(x,y) = \text{horizontal zero-crossings of } RC_w(x,y)$$
$$RZ_w(x,y) = \text{all zero-crossings of } RC_w(x,y).$$

Each of these is a bit map.

(3) **Fixation position:** Initially choose the alignment of the two images to correspond to some preset lower limit, and increment by a specified amount until the alignment exceeds some preset upper limit.

(4) **Matching:** The matching algorithm can be subdivided into three sections. First, the feature points are matched; then, figural continuity is applied to the resulting matches; and finally, any ambiguities between matches are resolved.

(4.1) **Feature point matching.** The feature point matching portion of the algorithm can be summarized as follows. Suppose we are dealing with zero-crossing descriptions corresponding to some particular filter of size w_0. Given a disparity d_0, we construct an $N \times N \times 2w_0$ local disparity array M:

$$M(x,y,r) = \left\{ LP_{w_0}(x,y) \bigwedge \left[\bigvee_{v=y-\epsilon}^{y+\epsilon} RP_{w_0}(x+d_0+r,v) \right] \right\}$$
$$\bigvee \left\{ LN_{w_0}(x,y) \bigwedge \left[\bigvee_{v=y-\epsilon}^{y+\epsilon} RN_{w_0}(x+d_0+r,v) \right] \right\}$$

where $0 \leq x \leq N$, $0 \leq y \leq N$, and $-w \leq r \leq w$. Thus, each slice of $M(x,y,r_0)$ given by a value r_0 of r is a set of matched feature points, within a vertical range of $\pm\epsilon$, for a local disparity value r about the current convergence value d_0. Note that positive zero-crossings are matched to positive ones, and negatives to negatives, over a vertical range of $\pm\epsilon$, and over a horizontal range of $\pm w$ about the current alignment.

(4.2) **Figural continuity.**

In order to distinguish correct from random feature point matches, we apply a figural continuity constraint, by restricting the accepted matches to those extended contour segments whose length is sufficiently large. First, we need a means of defining a path along a zero-crossing contour. If $LZ_{w_0}(x,y) = 1$, that is if there is a zero-crossing at this point, then we define $f_{L,w_0} = (u,v)$ to be the next point along the zero-crossing contour. In other words, if the vector $\mathbf{r} = (x,y)$ is an index into the zero-crossing array, and if $LZ_{w_0}(x_0,y_0) = LZ_{w_0}(\mathbf{r}_0) = 1$ then the ordered sequence

$$\mathbf{r}_0, \quad f_{L,w_0}(\mathbf{r}_0), \quad f_{L,w_0}(f_{L,w_0}(\mathbf{r}_0)), \ldots$$

traces out a zero-crossing contour.

Then, given a threshold ϵ on the expected error rate ($0 \leq \epsilon < 1$), we need a threshold on the length of the matched contour segments. By the previous discussion, this is given by

$$\ell_j = \min \{k \mid P_{k,k-j} < \epsilon\}$$

where $P_{k,k-j}$ is given by equation (1). Thus, we let ℓ_0, ℓ_1, ℓ_2 denote the contour lengths required by contours of 0, 1 and 2 gaps respectively. Then the procedure for figural continuity can be specified as follows.

Figural Continuity Procedure

Compress all the matches into one representation:

$$MT(x,y) = \bigvee_{r=-w}^{w} M(x,y,r) \qquad \forall x, y.$$

Initialize the output array:

$$SM(x,y) = 0 \qquad \forall x, y.$$

For each point $\mathbf{r}_0 = (x_0, y_0)$ such that $MT(\mathbf{r}_0) = 1$, apply the following procedure. Set:

$g = 0$; gap counter
$\ell = 1$; length counter
$S = \{\mathbf{r}_0\}$; contour tested
$\mathbf{p} = \mathbf{r}_0$; contour pointer.

(0) **If** $f_{L,w_0}(\mathbf{p}) = \mathbf{r}_0$
then we have completed tracing the contour, and it is not long enough, so exit without saving the contour;
else,
  if $LH_{w_0}(f_{L,w_0}(\mathbf{p})) = 1$
  then the next point is a horizontal zero-crossing, so go to (1);
  else,
    if $MT(f_{L,w_0}(\mathbf{p})) = 0$
    then there is a gap so increment the gap counter: $g = g + 1$
    and go to (1);
    else increment the length counter: $\ell = \ell + 1$
    and continue.

(1) **If** $g > 2$
then the gap is too large, so exit without storing the contour;
else,
  if $g = 2$,
  then,
    if $\ell \geq \ell_2$
    then save the contour: $\forall \mathbf{p} \in S$, set $SM(\mathbf{p}) = SM(\mathbf{p}) \bigvee MT(\mathbf{p})$
    else go to (2).
  else,
    if $g = 1$,
    then,
      if $\ell \geq \ell_1$

then save the contour: $\forall \mathbf{p} \in S$, set $SM(\mathbf{p}) = SM(\mathbf{p}) \bigvee MT(\mathbf{p})$
else go to (2).

else,

if $g = 0$,
then,

if $\ell \geq \ell_0$
then save the contour: $\forall \mathbf{p} \in S$, set $SM(\mathbf{p}) = SM(\mathbf{p}) \bigvee MT(\mathbf{p})$
else go to (2).

(2) Increment the contour collection, setting $S = S \bigcup \{f_{L,w_0}(\mathbf{p})\}$
and increment the contour pointer, setting $\mathbf{p} = f_{L,w_0}(\mathbf{p})$.
Go to (0).

■

(4.3) **Disparity updating.**

When this procedure is finished, $SM(\mathbf{p})$ contains all the matches for this alignment that pass the figural continuity constraint. Now, we need to update the global disparity array $D_{w_0}(x, y, d)$. This is accomplished by looping over all values of $\mathbf{p}$ and applying the following procedure.

Disparity Update Procedure

If

$$SM(\mathbf{p}) = 1,$$

then set

$$D_{w_0}\left(\mathbf{p}, \frac{\sum_{o=-w}^{w} SM(\mathbf{p}, o)(d_0 + o)}{\sum_{o=-w}^{w} SM(\mathbf{p}, o)}\right) = 1.$$

■

That is, we mark a 1 at the point in the three-dimensional disparity array corresponding to the average disparity of the local matches. Thus for each d, the set

$$\{D_{w_0}(\mathbf{p}, d) \mid \forall \mathbf{p}\}$$

is a disparity slice of the matched images.

To create the total disparity array D, we can simply let d_0 range between preset limits d_ℓ to d_h, and iterate over the previous steps. Note that this is an extremely simple control strategy, which could clearly be augmented, for example along the lines suggested in the original Marr-Poggio theory. In cases where a detailed, fine resolution, disparity map is desired, this simple control mechanism should suffice. In situations in which speed is a critical factor, an attention focussing mechanism that uses coarse disparity information to guide finer resolution matching is probably essential.

The above algorithm has been specified for a single operator size w_0 and can be applied at each of the four sizes specified earlier. The original Marr-Poggio theory proposed that a coarse to fine matching strategy be used to guide the matching at finer resolution representations, in part

because the ambiguity of such matches increases with the increasing density of the zero-crossings. While we have split off the control strategy aspects of this proposal by sweeping the images through the entire range of possible disparities for each operator, the use of multiple resolution operators as a means of disambiguation still remains a possibility.

(5) **Loop:** Simply loop to step (3) to increment over all possible image alignments.

(6) **Disambiguation.** In particular, while only a single match will be assigned a zero-crossing point, for each alignment of the images, d_0, it is possible that more than one contour will be matched to the point, as the disparity sweeps through the range $d_\ell \leq d_0 \leq d_h$. We can use the disparity information obtained at coarser channels to help disambiguate this case. For each channel size w_0, we perform the following operations.

First, we project the disparity array, setting, $\forall \mathbf{p}$:

$$PD_{w_0}(\mathbf{p}) = \begin{cases} d, & \text{if } D_{w_0}(\mathbf{p}, a) = \delta_{ad} \\ \text{null}, & \text{if } D_{w_0}(\mathbf{p}, a) = 0, \forall a \\ ?, & \text{if otherwise.} \end{cases}$$

Thus, if there is exactly one match, $PD_{w_0}(\mathbf{p})$ equals the disparity value of that match; if there is no match, it is set to null; and if there is more than one match, $PD_{w_0}(\mathbf{p})$ is marked with the special character "?". If w_0 is currently set to the largest possible filter size, then nothing can be done. If it is set to a smaller filter size, however, then let w_ℓ denote the next largest filter size and proceed in the following manner.

Disambiguation Procedure

For each point $\mathbf{p}$ such that $PD_{w_0}(\mathbf{p}) = ?$, let

$$A = \{a \mid D_{w_0}(\mathbf{p}, a) = 1\}$$

denote the set of possible matches for this point.

If there is a point $\mathbf{p}'$ in a neighbourhood $\mathcal{N}_{w_\ell}(\mathbf{p})$ about this point, such that

$$PD_{w_\ell}(\mathbf{p}') \neq \text{null}$$

and

$$PD_{w_\ell}(\mathbf{p}') \neq ?$$

and such that

$$|PD_{w_\ell}(\mathbf{p}') - a_i| \leq \frac{w}{2}$$

for some $a_i \in A$,

then a_i is a legitimate disparity value.

If there is exactly one legitimate element a_i of A,

then set

$$PD_{w_0}(\mathbf{p}) = a_i$$

else set

$$PD_{w_0}(\mathbf{p}) = \text{null}.$$

In this manner, we create the disparity map PD_{w_0} for the current filter size w_0.

(7) **Loop:** We can iterate this procedure over decreasing values of w_0. When this is finished, we have a series of disparity maps PD_w of increasing resolution as w decreases.

(8) **Consistency.** The disambiguation process described above can be considered as a type of consistency check. That is, if there are two contours that, to within the limits of the figural continuity constraint, match a given contour, we can use coarser level information to eliminate the incorrect match. This relies on the assumption that the correct contour will be accepted by figural continuity. There may also be circumstances in which the correct contour is not accepted, for example because it is occluded in one of the images, but in which an incorrect contour passes the figural continuity constraint, and is accepted as a correct match. While this occurs very rarely (empirical observations suggest that less than 0.005 of the matched zero-crossing contours have this problem), it is possible to apply a consistency check to the computed disparity maps to remove this possibility.

Consistency Procedure

Given two adjacent filter sizes $w_s < w_\ell, \forall \mathbf{p}$,
if $PD_{w_s}(\mathbf{p}) \neq$ null
then,
  if $N_{w_\ell}(\mathbf{p})$ is empty, leave $PD_{w_s}(\mathbf{p})$ as it stands,
  else,
    if there is a point $\mathbf{p}' \in N_{w_\ell}(\mathbf{p})$ such that $|PD_{w_s}(\mathbf{p}) - PD_{w_\ell}(\mathbf{p}')| < \frac{w}{2}$
    then leave $PD_{w_s}(\mathbf{p})$ as it stands,
    else, set $PD_{w_s}(\mathbf{p}) =$ null as it is not consistent with the coarser resolution disparity map.

4. Examples

We will examine two different types of stereo imagery in this section, a laboratory scene with many of the characteristics of industrial robotics situations, and aerial photographs of natural and artificial terrain. The intent is both to provide a means of examining the performance of the stereo algorithm outlined in the previous section, and to consider the potential applicability of such algorithms to automated stereo acquisition of depth information, both in robotics and cartography.

4.1. Laboratory Scenes

We consider first an example of a laboratory scene, shown in Figure 2. The scene is composed of a set of wooden blocks, of different shapes and lying at different distances from the cameras.

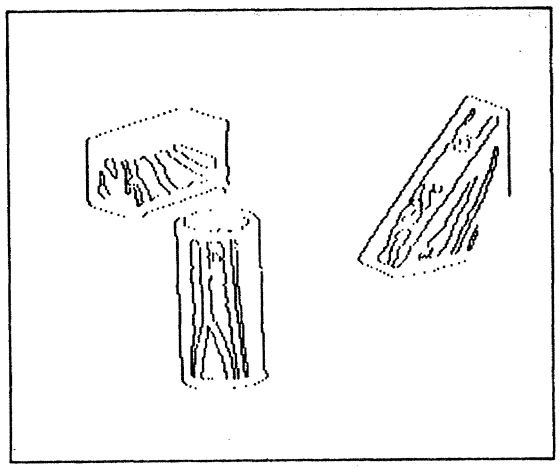

Figure 6. The set of matched zero-crossings for the blocks image.

The images were taken with an Hitachi CCD camera, and are 288 by 224 pixels each. The images contain grey-levels from 0 to 255, although the contrast range is more on the order of 10 to 110. The cameras were positioned roughly 1500 mm from the foremost point in the image, namely the front of the cylinder, with a separation of roughly 290 mm. By roughly, we mean that the distances were measured to an accuracy of a few millimeters.

The left and right images were convolved with four different sized $\nabla^2 G$ filters, with central widths given by $w = 17, 13, 9$ and 5 pixels each. These convolutions are illustrated in Figure 3.

The zero-crossings obtained from each of these convolutions are shown in Figure 4. Note by comparison to the convolutions that most of the zero-crossings in the support plane have very shallow gradients, corresponding to low contrast changes in the images. The positions of such zero-crossings tend to be sensitive to noise, an issue to which we will return shortly. As has been demonstrated in earlier implementations of the Marr-Poggio model, the density of the zero-crossings is directly proportional to the size of the $\nabla^2 G$ filter. Note also that the zero-crossings of the largest operator tend to capture coarse features of the objects, such as their occluding boundaries, while the zero-crossings of the smaller operators tend to capture in addition finer details, such as the wood grain on the objects.

The set of zero-crossings from the finest level operator to which a matching zero-crossing is assigned by the algorithm is displayed in Figure 6. Note that the figural continuity constraint has removed virtually all of the matches corresponding to the shallow zero-crossings of the background plane. As we noted earlier, these shallow zero-crossings tend to be sensitive to noise in the system, and as a consequence there can be a noticeable variation in the position of such zero-crossings, due to this noise component. One of the advantages of the algorithm presented here is that the variation in zero-crossing position due to noise will generally violate the figural continuity constraint, and hence such matches, with inherently noise disparity information attached to them, will be pruned from the final disparity data. We should note, however, that there may be other edge detection techniques that are more effective at removing such noise-sensitive features prior

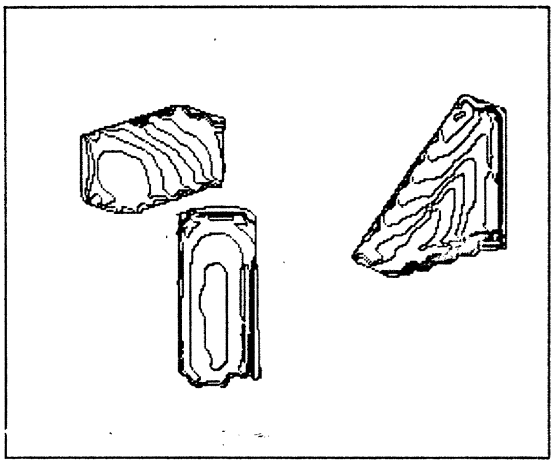

Figure 7. Contour map of the blocks image.

to the matching stage [for example, Canny, 1983].

The vertical disparity in this set of images covers a range of ±3 lines. To obtain the results displayed here, the algorithm was run at three different vertical alignments, and the results of each pass of the algorithm were merged into a single disparity array.

Finally, in order to display the results of the stereo algorithm, we apply the following process. We first interpolate the disparity information provided by the finest level channel, using a model of visual surface reconstruction based on the image irradiance equation [Grimson, 1982, 1983a, 1983b]. To do this, we use a portion of an efficient multi-grid implementation of an alternative but similar surface interpolation model, developed by Terzopoulos [1983, 1984]. Given the output of this process, which is a dense reconstruction of the disparity over the image, we plot isometric disparity contours, as shown in Figure 7.

The isometric disparity contours clearly demonstrate the local variations in depth of the objects, as computed by the stereo algorithm. It can be seen that the isometric disparity contours are not perfectly parallel, as might be expected from the shape of the blocks. This indicates that while overall the computed shape of the objects is correct, there may be a certain amount of local variation in the disparity values, leading to a distortion of the isometric contours. This is further illustrated in Figure 8, which shows a perspective view of the reconstructed surfaces of the blocks.

To further evalute the performance of the algorithm, especially the extent of this local variation, we performed the following additional tests. First, the disparity information was converted to actual distance values, based on the separation of the cameras, the angles of convergence of the cameras and the size of each individual pixel. These parameters were measured for the geometry used to record the original stereo images, and thus, the distances from the camera to points in the image were computed. The following table records the computed and measured distances, in millimeters, for a selected set of points in the image.

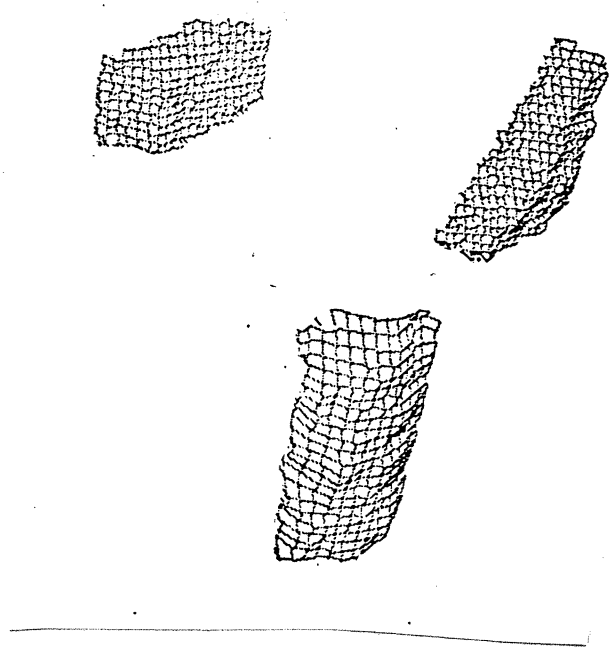

Figure 8. Perspective view of the reconstructed blocks surfaces.

Table 1 – Computation of Distance			
Points	Computed	Measured	Difference
Cylinder front	1506	1517	11
Wedge front	1647	1665	18
Block front	1743	1758	15
Cylinder to block	237	241	4
Cylinder to wedge	141	148	7
Cylinder radius – left	16	17	1
Cylinder radius – right	18	17	1
Wedge – depth extent	33	35	2
Block – depth extent	47	50	3

The first three entries record absolute depth measurements, and it can be seen that the computed distances to the fronts of the three objects are off by approximately 15 mm, out of a sensing distance of 1500 mm, or roughly 1%. Note that this transformation to absolute distance is sensitive not only to errors in the computation of stereo correspondence, but also to errors in the measurement of the camera geometry. Given the coarseness with which the camera parameters were computed, it is likely that this is the major source of error in the computation of absolute distance.

The remaining entries of the table record relative computed distances, both for separations of the objects, and for the depth extent of the objects. The fourth and fifth entries record the computed and measured relative separations of the objects. The final four entries record the radius of the cylinder, as measured to the left and right of the front of the cylinder, and the change

in depth across the block and wedge, for this particular viewing angle. On average, the error in relative depth tends to be on the order of 5–7 mm, out of a total depth range of 300 mm. To put this in the context of the stereo algorithm, we note that for this camera geometry, an error in stereo matching of one pixel would give rise to a depth error of 5-10 mm, depending on the actual location in the image. Thus, the errors in relative depth are essentially on the order of a pixel in disparity.

4.2. Aerial Photographs

The second type of images to which we have applied the stereo algorithm are aerial photographs, both of natural terrain and man-made structures. The performance of the modified stereo algorithm on all the images is summarized in the following table.

Table II – Stereo Summary

	Blocks	UBC	Ft.Sill	Phoenix	Boeing
Size	288 × 224	320 × 320	512 × 512	512 × 512	320 × 320
Disparity Range	56	13	51	41	13
Zero-crossings	11013	16801	32907	31403	10642
Matched Z-C's	1780	12310	16073	23890	6608
Matching Errors	0	9	286	78	167
After Consistency	0	0	0	0	33

The row labelled *size* indicates the dimensions of the images. The row labelled *disparity range* lists the disparity range of each image pair, in pixels. In the row labelled *zero-crossings*, we indicate the total number of zero-crossing pixels, including horizontal ones. In the row labelled *matched z-c's*, the number of such zero-crossings that are assigned a match is indicated. In the row labelled *matching errors*, the number of zero-crossings pixels that are assigned an incorrect match are listed. Note that we distinguish here between matching errors and localization errors. Matching errors are those that arise when incorrect zero-crossings contours are matched, independent of the accuracy of the contours themselves. Such errors tend to be relatively large in disparity. Localization errors are those that arise due to error in position of the zero-crossing contour itself. Such errors usually tend to be relatively small. The row labelled *after consistency* lists the number of such matching errors that remain after the consistency constraint is applied between different resolution disparity maps.

The images themselves are illustrated in Figures 9-20. For each one, we show the stereo images, the disparity map obtained by matching the zero-crossings are the finest level of representation, and a contour map based on this disparity map. The disparity maps are displayed using intensity to encode height, so that the brighter disparity points are closer. To obtain a contour map representation of the results, we have applied a surface reconstruction algorithm [Grimson 1982, 1983a, Terzopoulos, 1983, 1984] to the stereo data.

The first pair of images, from the Phoenix area, are illustrated in Figure 9, and were supplied courtesy of the Defense Mapping Agency. A second stereo pair of natural terrain, from the Fort Sill, Oklahoma area, are illustrated in Figure 12, and were supplied courtesy of the U.S. Army Engineering Topographic Laboratory. The next stereo pair, from the University of British

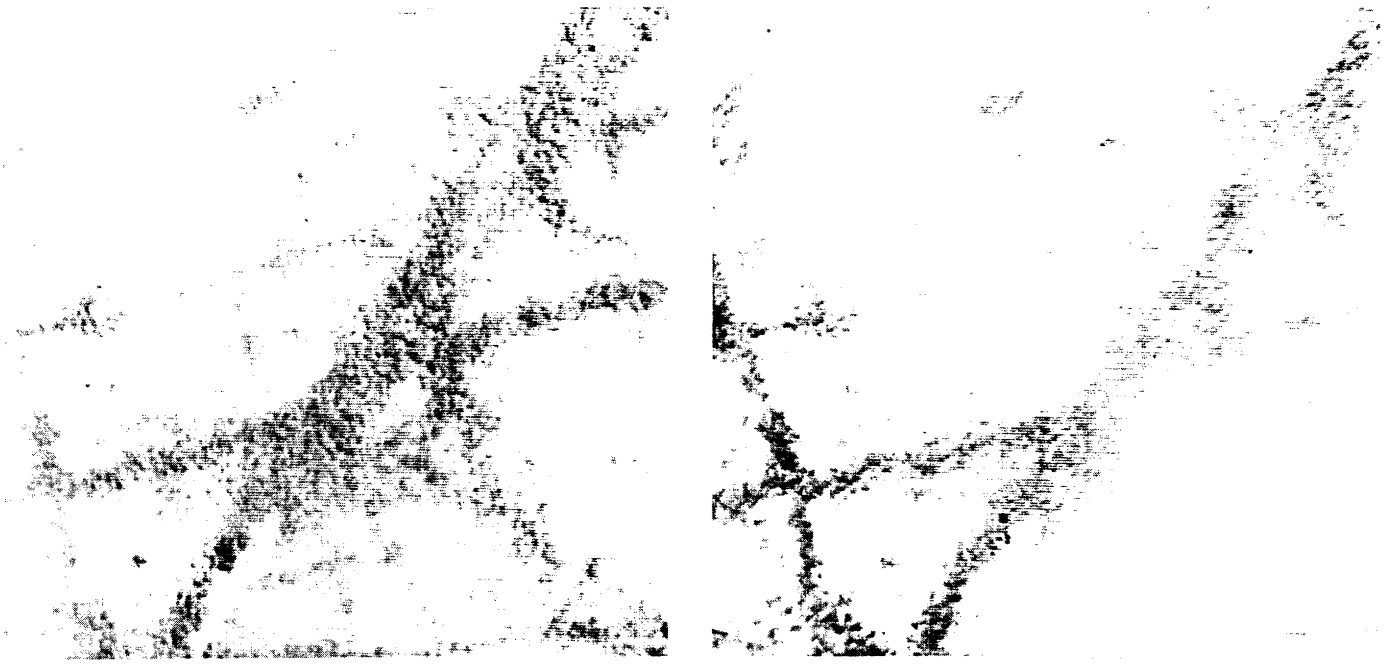

Figure 9. Natural terrain stereo pair (Ft. Sill).

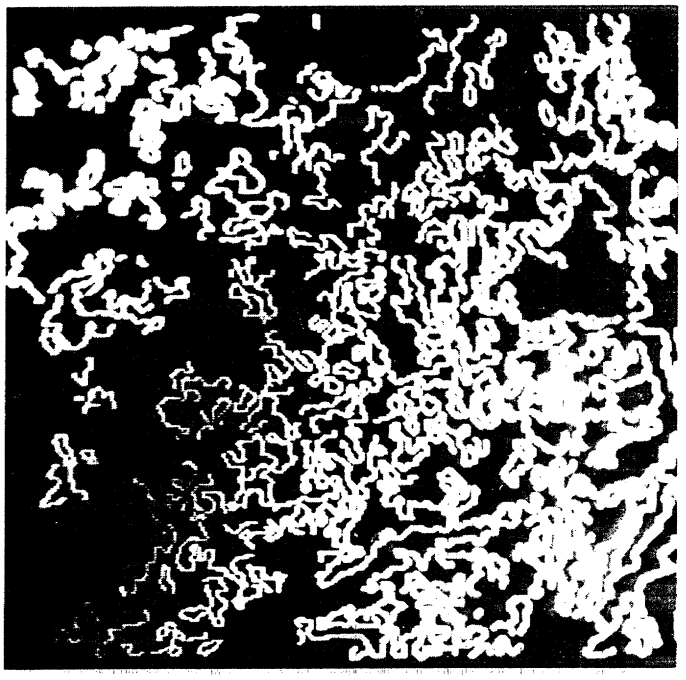

Figure 10. Disparity map (Ft. Sill).

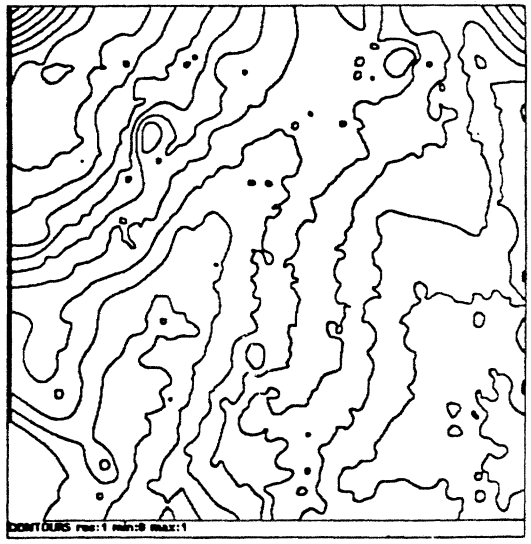

Figure 11. Contour map (Ft. Sill) based on matching before consistency check.

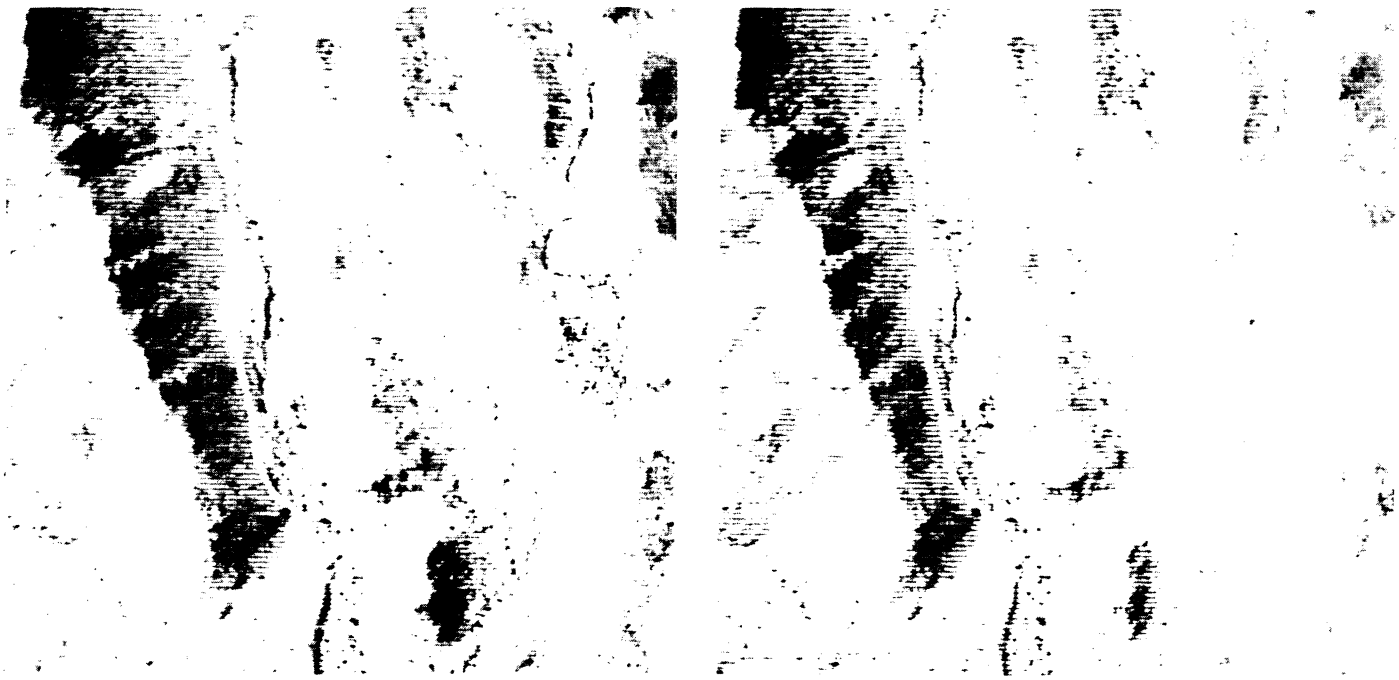

Figure 12. Natural terrain stereo pair (Pheonix).

Columbia, and supplied courtesy of UBC, are illustrated in Figure 16. The final stereo pair are of a highway interchange, and were supplied courtesy of Boeing Corporation.

A number of comments are in order concerning the performance of the algorithm, as indicated above. We note that in the case of the blocks scene, the percentage of matched zero-crossing to total

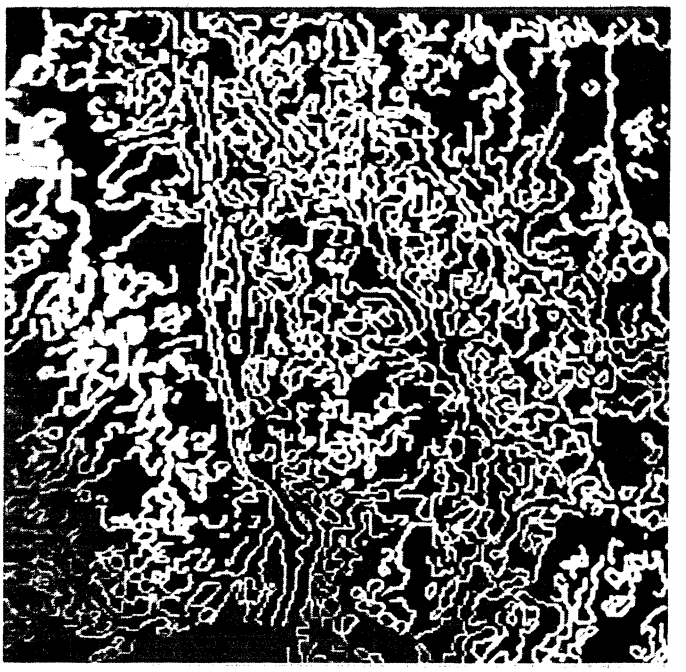

Figure 13. Disparity map (Pheonix).

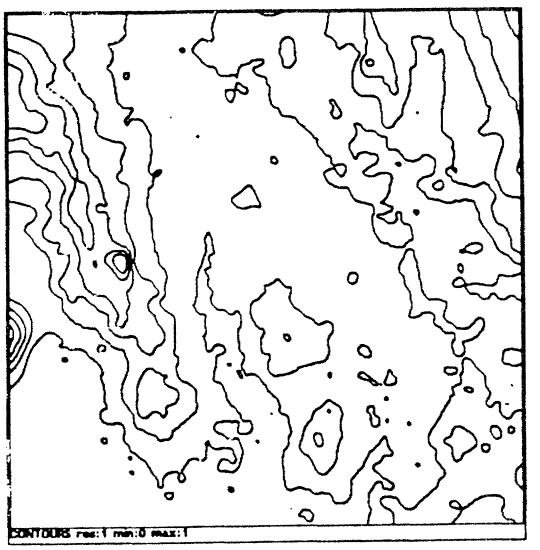

Figure 14. Contour map (Pheonix) based on matching before consistency check.

zero-crossing is small, on the order of .17 percent. Note, however, that many of the zero-crossings are shallow, unstable zero-crossing, corresponding to small fluctuations in the photometric process, as illustrated by Figure 3. If we consider only zero-crossing points on the blocks themselves, then the number of eligible zero-crossing points reduces to 2703, of which 1780 are assigned a match. Note further that this number of 1780 does not include any strictly horizontal zero-crossing points,

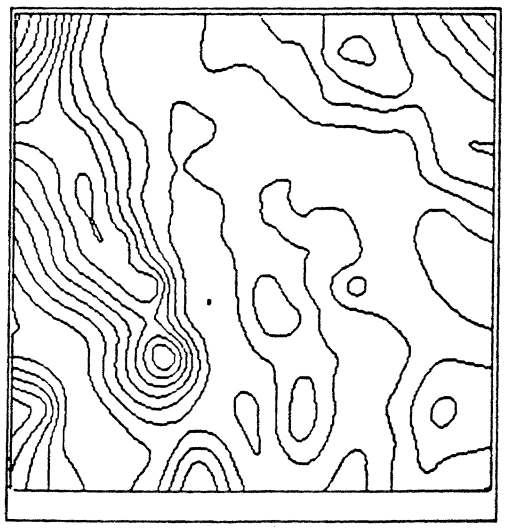

Figure 15. Contour map (Pheonix) based on matching after consistency check.

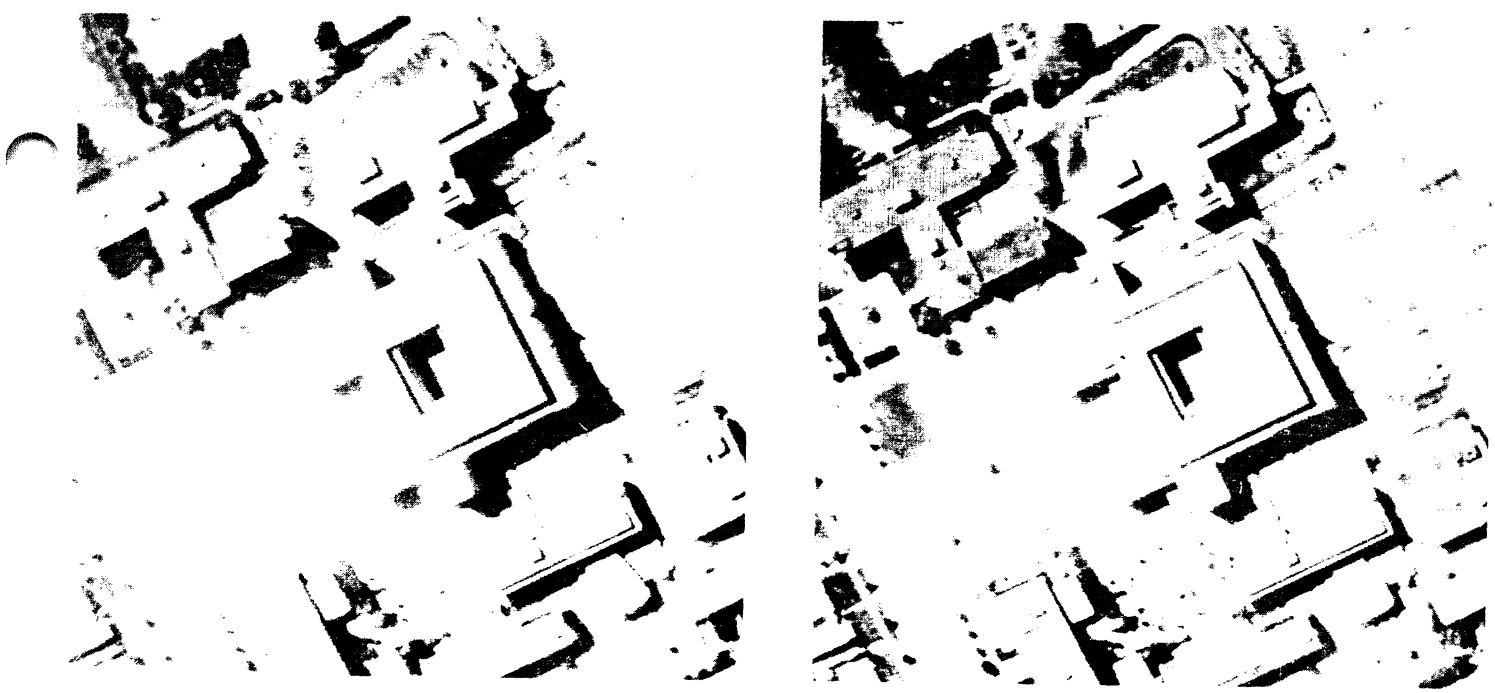

Figure 16. Natural terrain stereo pair (UBC).

nor does it include very small zero-crossing contours, which fall below the matching thresholds, and are hence unmatchable.

The Fort Sill image does provide some difficulty for the algorithm, particularly because the photometric properties of the images cause a certain amount of fluctuation in the positions of the

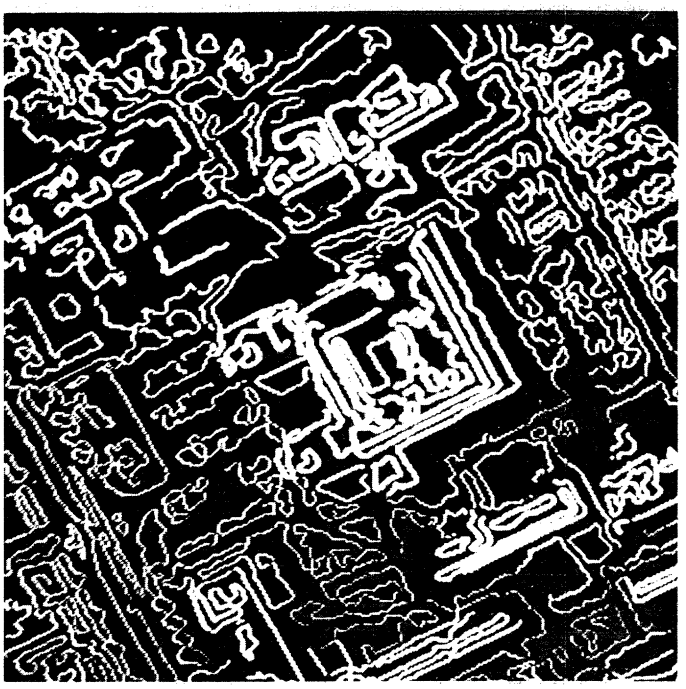

Figure 17. Disparity map (UBC).

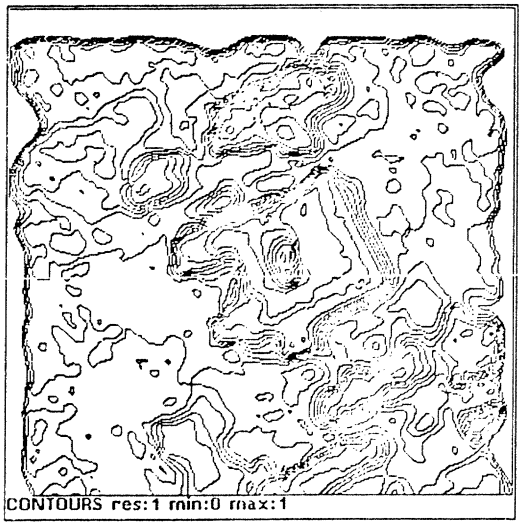

Figure 18. Contour map (UBC) based on matching before consistency check.

zero-crossing contours. As a consequence of the design of the matching procedure, which favors no match to possible incorrect matches, a large number of the potential zero-crossing points are not matched. Note, however, that the percentage of matched zero-crossings to total zero-crossings is somewhat misleading, since a large number of the total are not, in fact, matchable. In this case, at least ten percent of the zero-crossings in the left image are not present in the right since

Figure 19. Natural terrain stereo pair (Boeing).

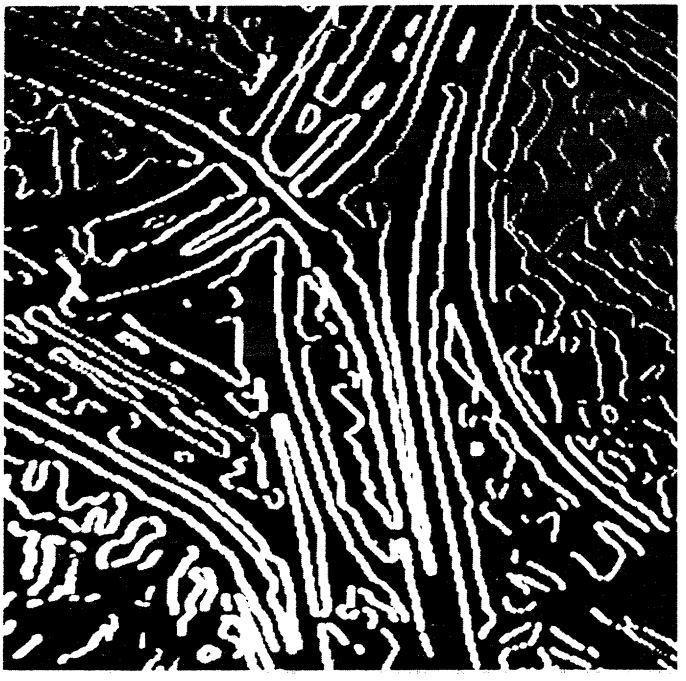

Figure 20. Disparity map (Boeing).

they lie beyond the edge of the image. We also note that the contour map displayed in Figure 11 is based on the results of the matching algorithm before the consistency check is applied. As a consequence, the effect of the single incorrectly matched contour in the upper left quadrant is clearly visible as a sudden dip in the contour map. This clearly demonstrates the need for a consistency check to remove obvious matching errors that survive the matching process itself.

In the Phoenix images, the contour map of Figure 14 is also generated from matching data without a consistency check. In figure 15, we apply the surface reconstruction algorithm to the data after applying the consistency check. We also have relaxed the tightness with which the reconstruction is forced to pass through the stereo data. It can be seen that the resulting contour map has removed the obvious matching defects and has a smoother set of contours. This smoother surface reconstruction is one means of removing possible localization errors in the matched data, as well as matching errors that survive the process.

While the Fort Sill image presents a great deal of difficulty to the algorithm due to large fluctuations in the positions and shapes of the zero-crossing contours, the Boeing image presents a different type of difficulty. Here, the large number of extended, parallel image contours presents a large set of potential ambiguities. In general, however, the algorithm is able to solve this problem, by relying on information from coarser channnels to disambiguate finer ones. Because the interpolation process is only applicable across smooth surfaces, and the Boeing image contains a large number of surface discontinuities, we have omitted the contour map for this image.

It is important to stress with all of the contour maps, and especially for the UBC images, that these illustrations are intended as a graphical means of displaying the performance of the stereo algorithm but not as a precise reconstruction of the underlying terrain. In particular, since one of the parameters of the surface reconstruction algorithm is the degree of smoothing applied to the reconstructed surface, the resulting contour maps may exhibit more smoothing than is warranted, due to the choice of this parameter. Nonetheless the qualitative performance of the stereo algorithm is still evident by the arrangement and spacing of the contours. In the case of the stereo pairs with buildings and other artifacts present, the application of the surface reconstruction algorithm directly to the results of the stereo algorithm is actually incorrect, since it attempts to fit a single surface over what are in fact several distinct surfaces. To be completely correct, the stereo depth data should be segmented into coherent regions, and then interpolated. Since this was not done, the resulting surface interpolation tends incorrectly to smooth over the discontinuities in depth. Nonetheless, the contour maps illustrated still demonstrate the basic performance of the stereo algorithm and the tightly clustered isometric contours help to indicate the separations of the different buildings from the ground.

5. Discussion

The modified Marr-Poggio-Grimson algorithm presented here was originally implemented in LISP on an MIT Lisp Machine, and then recoded in Lisp Machine microcode, for more efficient performance. The convolutions of the images were performed using a special purpose convolution

device [Nishihara and Larson, 1981]. While the time required to process an image is dependent on a large number of factors involving the complexity of the image, it is possible to give estimates on the performance of this implementation of the algorithm. Using a 320×320 image as a basis, we have observed the following timing characteristics. Each convolution of an image, including time required to interface the convolution device with the Lisp Machine, usually required on the order of 5 seconds. Each computation of a zero-crossings representation typically required on the order of 10 seconds. The amount of time required to match the zero-crossing representations was highly dependent on the number of fixation positions required (and thus on the total disparity range of the image). Matching at each such fixation position usually required on the order of $5 - 20$ seconds, depending on the structure of the zero-crossings contours. Finally, combining all the slices of the disparity map into a single consistent representation typically required on the order of $30 - 60$ seconds. Thus, for example, a single fine resolution channel processing of the UBC images normally took under 5 minutes in total, and the total time for running three different resolution channels was on the order of 10 minutes.

6. Acknowledgements

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 under Office of Naval Research contracts N00014-80-C-0505 and N00014-82-K-0334.

The aerial photographs of University of British Columbia were supplied courtesy of Bob Woodham, the photographs of Phoenix were supplied courtesy of John Unruh of the Defense Mapping Agency, the photographs of Fort Sill were supplied courtesy of George Lukes of the U.S. Army Engineering Topographic Labs, and the photographs of the highway interchance were supplied courtesy of the Boeing Corporation. The contour map of Figure 15 was kindly provided courtesy of Demetri Terzopoulos.

The author wishes to thank Tomaso Poggio, Ellen Hildreth, Berthold Horn, Tomas Lozano-Perez, John Mayhew, John Frisby, Mike Brady and Demetri Terzopoulos for many valuable comments and discussions, and Demetri Terzopoulos for kindly providing access to his extremely efficient surface reconstruction algorithm.

7. References

Arnold, R.D. and Binford, T.O. "Geometric constraints in stereo vision," *Proc. SPIE, San Diego* **238**, (1980), 281-292.

Baker, H. H. "Depth from edge and intensity based stereo," Stanford University Technical Report STAN-CS-82-930, September, 1982.

Baker, H. H. and Binford, T. O. "Depth from Edge and Intensity Based Stereo," *Seventh International Joint Conference on Artificial Intelligence*, August 1981, 631-636.

Barnard, S. T. and Thompson, W. B. "Disparity analysis of images," *IEEE Pattern Analysis and Machine Intelligence* PAMI-2, 4, (1980), 333-340.

Berry, R. N. "Quantitative relations among vernier, real depth, and stereoscopic depth acuities," *J. Exp. Psychol.* **38** (1948), 708-721.

Canny, J. F. "Finding Edges and Lines in Images," Massachusetts Institute of Technology Artificial Intelligence Laboratory Technical Report TR-720, June 1983.

Crick, F.H.C., Marr, D. and Poggio, T. "An information-processing approach to understanding the visual cortex," in *The Cerebral Cortex*, Neurosciences Research Program, (1980) 505-533.

Duwaer, A.L. and van den Brink, G. "Diplopia thresholds and the initiation of vergence eye-movements," *Vision Research* **21** (1981a), 1727-1737.

Duwaer, A.L. and van den Brink, G. "What is the diplopia threshold?" *Perception and Psychophysics* **29** (1981b), 295-309.

Frisby, J.P> and Mayhew, J.E.W. "The role of spatial frequency tuned channels in vergence control," *Vision Res.* **20** (1980) 727-732.

Grimson, W.E.L. "A computer implementation of a theory of human stereo vision," *Phil. Trans. Roy. Soc. Lond.***B 292** (1981a), 217-253. (an earlier version appeared as MIT AI Lab Memo 565, 1980).

Grimson, W.E.L. *From Images to Surfaces: A computational study of the human early visual system* MIT Press, Cambridge, Ma., 1981b.

Grimson, W.E.L. "A computational theory of visual surface interpolation," *Phil. Trans. Roy. Soc. Lond.* **B 298** (1982), 395-427.

Grimson, W.E.L. "An implementation of a computational theory of visual surface interpolation," *Computer Vision, Graphics and Image Processing* **22** (1983a), 39-69.

Grimson, W.E.L. "Surface consistency constraints in vision," *Computer Vision, Graphics and Image Processing* (1983b), to appear.

Hildreth, E.C. Implementation of a theory of edge detection, S.M. Thesis, Department of Computer Science and Electrical Engineering, Massachusetts Institute of Technology, 1980. (see also MIT AI Lab Technical Report 597, 1980).

Howard, J. H. "A test for the judgement of distance," *Am. J. Ophthal.* **2** (1919), 656-675.

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

Julesz, B. *Foundations of Cyclopean Perception* University of Chicago Press, Chicago, 1971.

Kak, A. C. "Depth perception for robots," Purdue University Technical Report TR-EE 83-44, (1983), (also to appear as a chapter in *Handbook of Industrial Robotics*, S. Nof (ed), John-Wiley, New York).

Longuet-Higgins, H. C. "The roel of the vertical dimension in stereoscopic vision," *Perception* **11** (1982) 377-386.

Kass, M. "A computational framework for the visual correspondence problem," *Eighth International Joint Conference on Artificial Intelligence*, (1983), 1043-1045.

Kass, M. "Computing Stereo Correspondence," M.Sc. Thesis, Dept of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1984.

MacVicar-Whelan, P.J. and Binford, T.O. "Intensity discontinuity location to subpixel precision," *Seventh International Joint Conference on Artificial Intelligence*, (1981), 752-754.

Marr, D. " Representing visual information," *AAS 143rd Meeting, Symposium on some mathematical questions in biology, February 1977* Published in *Lectures in the Life Sciences* **10** (1978), 101-180.

Marr, D. *Vision* W.H. Freeman and Company, San Francisco, 1982.

Marr, D. and Hildreth, E. "Theory of edge detection," *Proc. Roy. Soc. Lond.* **B 207** (1980), 187-217.

Marr, D. and Poggio, T. "Some comments on a recent theory of stereopsis," *MIT AI Lab Memo 558*, (1980).

Marr, D. and Poggio, T. "A theory of human stereo vision," *Proc. Roy. Soc. Lond.***B 204** (1979), 301-328. (an earlier version appeared as MIT AI Lab Memo 451, 1977).

Marr, D., Poggio, T. and Hildreth, E. "The smallest channel in early human vision," *J. Opt. Soc. Am.* 70, 7 (1979), 868-870.

Mayhew, J.E.W. "The interpretation of stereo-disparity information: the computation of surface orientation and depth," *Perception* **11** (1982) 387-403.

Mayhew, J.E.W. and Frisby, J.P. "Psychophysical and computational studies towards a theory of human stereopsis," *Artificial Intelligence* **17** (1981), 349-385.

Mayhew, J.E.W. and Longuet-Higgins, H.C. "A computational model of binocular depth perception," *Nature Lond.* **297** (1982) 376-379.

Moravec, H.P. "Towards automatic visual obstacle avoidance," *Fifth International Joint Conference on Artificial Intelligence*, (1977), 584.

Moravec, H.P. "Obstacle avoidance and navigation in the real world by a seeing robot rover," *Stanford Artificial Intelligence Laboratory, AIM-340*, (1980).

Mowforth, P., Mayhew, J.E.W. and Frisby, J.P. "Vergence eye movements made in response to spatial-frequency-filtered random-dot stereograms," *Perception* **10** (1981) 299-304.

Nielsen. K. R. K. and Poggio T. "Vertical image registration in human stereopsis," *MIT Artificial Intelligence Laboratory Memo* 743, 1983.

Nishihara, H.K. and Larson, N. G. "Towards a real time implementation of the Marr and Poggio stereo matcher," *Proceedings of the DARPA Image Understanding Workshop*, April, 1981, Washington D.C., 114-120.

Nishihara, H.K. and Poggio, T. "Hidden cues in random line stereograms," *Nature* **300** (1982), 347-349.

Ohta, Y. and Kanade, T. "Stereo by intra- and inter-scanline search using dynamic programming," Carnegie-Mellon University Technical Report CMU-CS-83-162, 1983.

Prazdny, K. "The role of eye position information in algorithms for stereoscopic depth," *Proceedings AAAI*, 1982, 1-4.

Prazdny, K. "Computing convergence angle from random dot stereograms," *Proceedings Eighth IJCAI*, Karlsruhe, West Germany, 1983, 1050-1052.

Schumer, R.A. and Julesz, B. "Disparity limits for random-dot cinematograms for movement and form detection, and a learning effect," *Suppl. Invest. Ophthalmol. Visual Sci.* **22**, (1982) 272.

Terzopoulos, D. "Multi-level reconstruction of visual surfaces" in *Multiresolution image processing and analysis* A. Rosenfeld, (ed). Springer-Verlag 1983. (See also MIT Artificial Intelligence Laboratory Memo 671, 1982.)

Terzopoulos, D. "Multiresolution computation of visible-surface representations," Ph.D. Thesis, Dept of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, January, 1984.

Tyler, C. W. "Spatial limitations of human stereoscopic vision," *Proceedings, SPIE* **120** (1977).

Woodburne, L. S. "The effect of a constant visual angle upon the binocular discrimination of depth differences," *Am. J. Psych.* **46** (1934), 273-286.

Wilson, H.R. "Psychophysical evidence for spatial channels," in *Physical and Biological Processing of Images* O. J. Braddick and A. C. Sleigh, eds. Springer-Verlag, Beriln, 1983. (pp. 88-99).

Wilson, H.R. and Bergen, J.R. "A four mechanism model for threshold spatial vision," *Vision Research* **19** (1979), 19-32.