Massachusetts Institute of Technology

Artificial Intelligence Laboratory

A.I. Memo 780 May 1984

# PRISM: A Practical Real-Time Imaging Stereo Matcher

H.K.Nishihara

## Abstract

A binocular-stereo-matching algorithm for making rapid visual range measurements in noisy images is described. This technique is developed for application to problems in robotics where noise tolerance, reliability, and speed are predominant issues. A high speed pipelined convolver for preprocessing images and an *unstructured light* technique for improving signal quality are introduced to help enhance performance to meet the demands of this task domain. These optimizations, however, are not sufficient. A closer examination of the problems encountered suggests that broader interpretations of both the objective of binocular stereo and of the zero-crossing theory of Marr and Poggio are required. In this paper, we restrict ourselves to the problem of making a single primitive surface measurement. For example, to determine whether or not a specified volume of space is occupied, to measure the range to a surface at an indicated image location, or to determine the elevation gradient at that position. In this framework we make a subtle but important shift from the explicit use of zero-crossing contours (in band-pass filtered images) as the elements matched between left and right images, to the use of the signs between zero-crossings. With this change, we obtain a simpler algorithm with a reduced sensitivity to noise and a more predictable behavior. The PRISM system incorporates this algorithm with the unstructured light technique and a high speed digital convolver. It has been used successfully by others as a sensor in a path planning system and a bin picking system.

**Acknowledgments:** 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 contract N00014-80-C-0505, in part by the System Development Foundation and in part by National Science Foundation Grant 79-23110MCS. The research described in this report was originally presented at the Third International Conference on Robot Vision and Sensory Controls, SPIE Proceedings **449**. This report will appear in a forthcoming issue of Optical Engineering.

# 1. Introduction

This paper presents an approach to solving the binocular-stereo-matching problem which places special emphasis on the practical issues of noise tolerance, reliability and speed. It is strongly influenced by Marr and Poggio's[1] zero-crossing theory, but differs from recent implementations in the way zero-crossing information is used to drive the matching and in the product the matcher is designed to produce.

## 1.1. A robust, high-speed stereo system

Binocular stereo is a technique for measuring range, by triangulation, to selected locations in a scene imaged by two cameras. Figure 1 illustrates the imaging geometry. The vergence angle indicated in the diagram from the cameras to a selected target can be determined from the relative positions of that target in the left and right camera images in conjunction with the angle between the camera axes. The primary computational problem of binocular stereo is to identify corresponding locations in the two images. Once the position of the same physical surface point is known in both images, the vergence angle indicated in the diagram can be determined and from that, the distance to the surface from the cameras.

### 1.1.1. Design goals

Four design objectives have guided this study. The first is *noise tolerance.* We want to understand how matching can be accomplished in the presence of moderate to large noise levels which occur anytime surface contrast is low compared with sensor noise and other inter-image distortions. The second is to achieve *competent performance* for at least one of the three stereo measurements—volume occupancy, range measurement, and detection of elevation discontinuities.[2] The third is to operate at a *practical speed* for robotics. Our emphasis here will be to streamline the computation to increase speed and use processing resources efficiently. This forces a careful consideration of the relative cost of producing a measurement in different ways vis-a-vis their contribution to the final product of the algorithm. Finally we require *simplicity.*[3–5] An algorithm is difficult to analyze extensively if it involves many serial decisions or many special cases. Of course a more complex algorithm might be necessary to obtain a desired level of performance in a specialized domain,

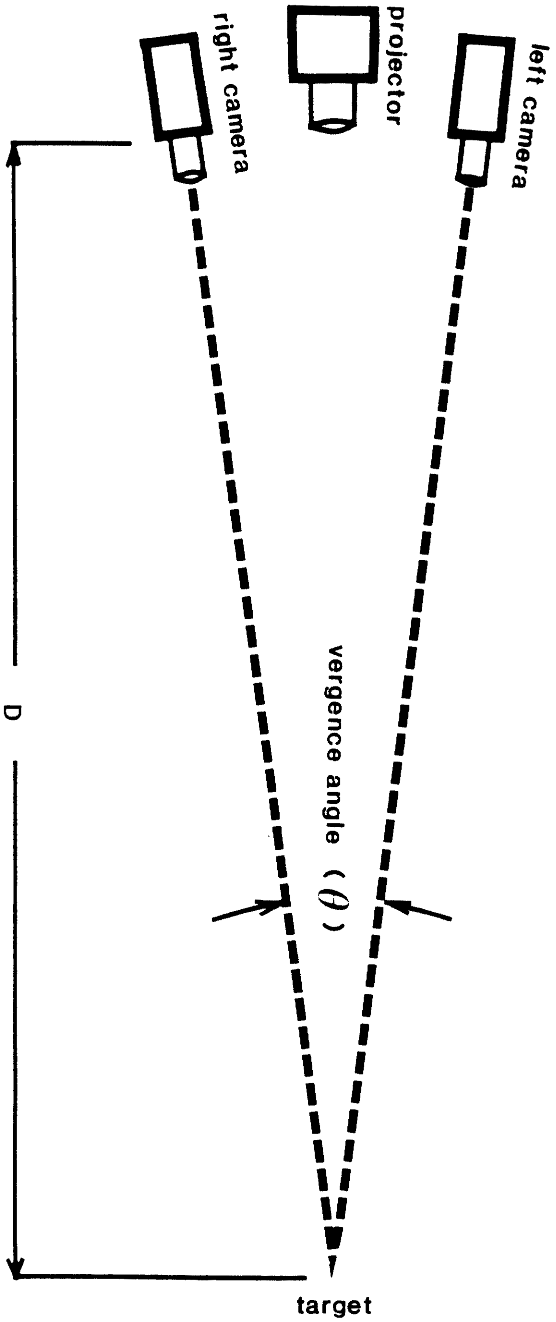

Figure 1. The imaging geometry used in the PRISM stereo matcher. Two vidicon cameras are mounted 40cm apart and 150cm from the target scene. The object of binocular stereo is to identify corresponding physical surface locations in the two camera images so that the vergence angle *θ* can be determined. With this and the camera separation, the distance *D* to that surface location can be calculated. A slide projector is situated between the cameras and projects a random dot texture onto the target surfaces to provide fixed surface markings on objects lacking sufficient natural texture.

but the strategy of this paper is to concentrate on simpler modular techniques which have the advantage of greater generality. These can then be incorporated flexibly into the design of more complex systems for special applications over a more diverse range.

### 1.1.2. Overview of the PRISM system

The PRISM system was developed in light of these design considerations. It relies both on a determination of the nature of the distortions that occur in noisy images and on changing the product of the matcher to more closely agree with the requirements of specific tasks that might be presented to a robotics vision system. The algorithm we use has its roots in the zero-crossing theory of Marr and Poggio, which will be described in the next section, but does not explicitly match zero-crossing contours. Algorithms which match zero-crossing contours tend to be very sensitive to local distortions due to system noise and are prevented from operating well on signals with moderate noise levels even though a substantial amount of information may still be present. Instead the PRISM system is based on the matching of a sign representation which is a dual of the zero-crossing representation. The initial design task of the implementation has been to rapidly detect obstacles in a robotics work space and determine their rough extents and heights.

The basic flow and organization of the computation is shown in figure 2. The scene (in front of a Unimation PUMA manipulator) is illuminated with an *unstructured* texture pattern by a slide projector as indicated in figure 1 to provide suitable matching targets on the otherwise clean surfaces common in industrial settings. A pair of inexpensive vidicon cameras, mounted above the workspace, digitize two $576 \times 454$ pixel images. The digitized video signals are then fed to a high speed digital convolver that preprocesses the images to produce filtered images at three scales of resolution for both left and right originals. A matching algorithm is then applied to the coarse pair to obtain a coarse $8 \times 6$ array of disparity measurements—disparity is defined here to be a vector in the image plane giving the translation required to bring the two images into registration at a designated location in one of the images. Then the same matching algorithm is applied to the medium

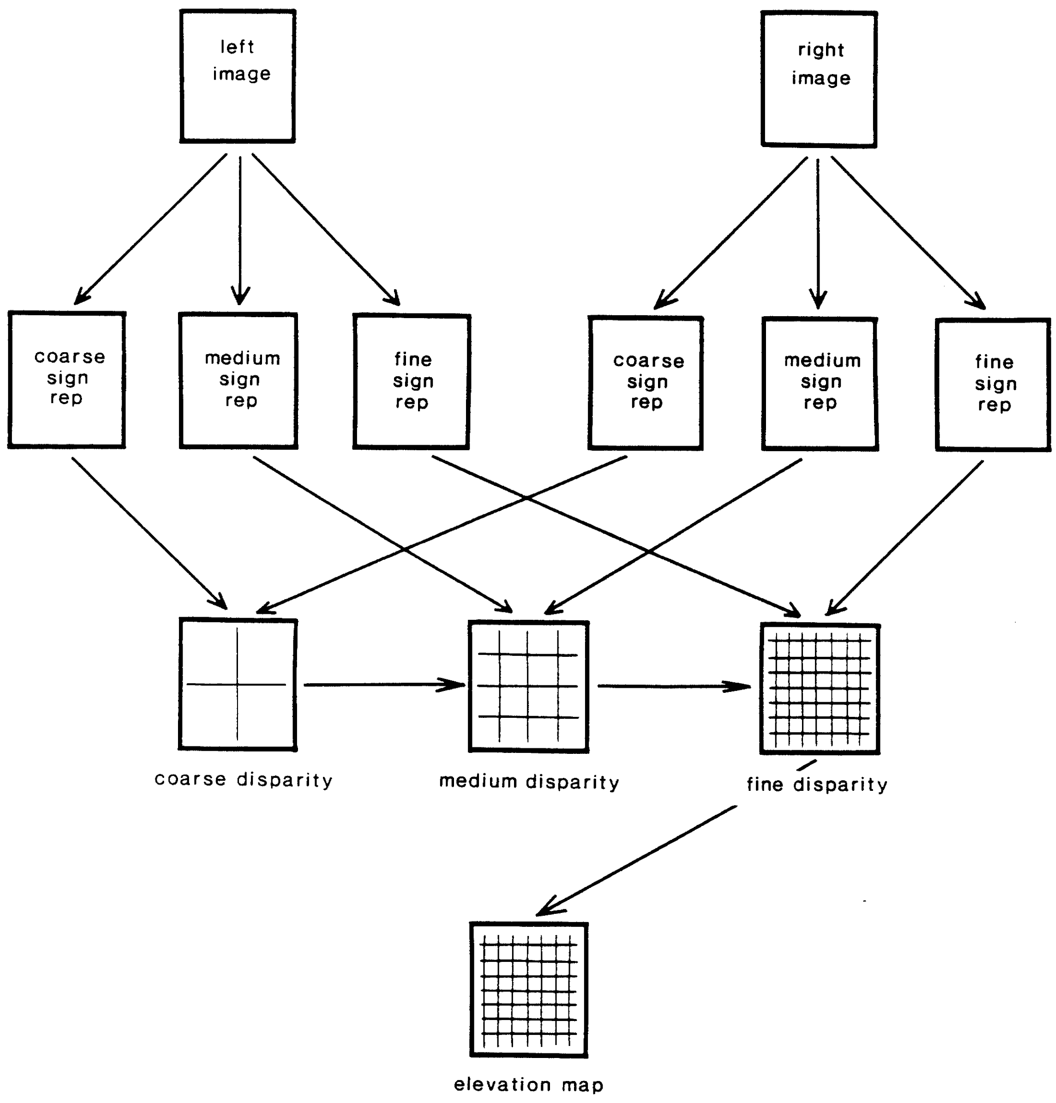

Figure 2. Information flow in the PRISM system. Left and right images are digitized and each is filtered to extract image information at three scales of resolution by convolution with two-dimensional $\nabla^2 G$ operators (with diameters of 32, 20, and 10 pixels). Each of these convolved images is hard clipped to produce a binary sign representation which is stored. The left and right coarse, medium, and fine scale sign representations are matched pair-wise to produce coarse, medium and fine scale disparity maps over the visual field of the cameras. The matching at the three scales is loosely coupled with coarser results guiding the search at the next finer scale. Finally, the fine scale disparity map is translated by means of a lookup table into an elevation map.

resolution filtered images to produce an 17 × 13 disparity array. The algorithm in this case starts its search for each measurement using the disparity measured at the corresponding location at the coarser scale. This loose coupling speeds up the search significantly. Finally, the third pair of filtered images is matched in the same way to produce a 36 × 26 array of disparity measurements. The system then transforms the disparities in this last array into an array of absolute elevations (millimeters above the work space surface). The entire process from raw images to a 36 × 26 array of elevation measurements with a resolution of about one part in 100 takes approximately 30 seconds.

## 1.2. Background

Intensity based area-correlation techniques have been investigated extensively for commercial applications in stereo-photogrammetry (see for instance [6]). Two of the best and most recent research efforts with correlation based approaches are due to Moravec[7] and Gennery,[8] who developed stereo systems for vehicular autonomous navigation. Tsai[9] has recently proposed two new methods which use as many as eight perspective views with known positions and orientations of the cameras. The correlation functions obtained by his method are significantly more peaked than conventional two-frame area correlation and allow a much smaller window size for the correlation measurements.

Symbolic matching techniques abstract away from the direct comparison of raw intensity values, permitting more explicit control over the image information used for matching.[10] Arnold and Binford[11,12] showed how constraints derived from the geometry of physical surfaces can be introduced to guide such matchers. Baker and Binford[13–15] developed a sophisticated matching system using many such constraints. Ohta and Kanade[16] have investigated dynamic programming methods for searching large spaces of possible correspondences between symbolic primitives from left and right images.

A third effort has its roots in the study of stereo matching in the human visual system.[17,18] Julesz showed conclusively with his random dot stereograms that stereo matching was an early process in the visual system that could function independently of monocular recognition. The psychophysical constraints that resulted from his work with the random dot stereogram stimulated thinking about how such computations could be accomplished. The prevailing intuition at the time had been that stereo involved some kind of parallel correlation or pattern matching process which operated on the fine detail of images.

### 1.2.1. Channels and zero-crossings

More recently Marr and Poggio began a concerted effort to develop a computational theory of stereo matching consistent with and guided by the apparently modular biological solution. Their first computational model[19] was designed to solve Julesz's random dot stereograms and was noteworthy for its explicit formulation of the computational assumptions—*continuity* and *uniqueness* of the imaged surfaces—required for solving the otherwise underdetermined matching problem.

Later, concentrating on the differences between their cooperative model and a larger body of psychophysical and physiological data they formulated an alternate model.[1] A key component was the observation that the matching problem was simplified if range and resolution were not both required simultaneously. Taken to either extreme, a high resolution short range matcher or a coarse resolution long range matcher, could be based on the use of simple matching primitives defined at a scale appropriate to the range-resolution trade-off selected. The resolution proposed for this matching module was very low, allowing just three disparity values: crossed, near zero, and uncrossed—a target at a crossed disparity would require a further crossing of the eyes (a larger vergence angle) to cause it to appear at the same position in left and right images. The nearest compatible candidate match was to be used and so the effective disparity range of the matcher depended directly on the spacing between neighboring matching primitives in the image.

A battery of such matching modules operating in parallel—each with a primitive size different from the others by about an octave—could produce high resolution measurements over a large range of disparities. Eye movements were proposed as a simple way to allow information obtained from matching coarse-scale primitives to bring modules operating on a finer scale into their respective ranges. The matching primitives required by their approach had to be scale specific and sensitive to reliable surface markings at that scale. Intensity values, from corresponding locations in the two stereo images, can differ due to the effects of camera position, camera response characteristics, and noise. These effects make the direct comparison of intensity a relatively weak indicator of the presence of a true correspondence. Image intensities coincident with larger intensity gradients in the images will be better localized spatially than at other locations in the images. Thus the locations of maxima in the intensity gradient are good places to compare intensities. In fact, the positions of these gradient maxima can be used directly as the primitive image features used for matching. On these grounds, Marr and Poggio based the definition of their matching primitive on local maxima in the intensity gradient—or equivalently zero-crossings in the second derivative. These directional derivatives being measured along the direction parallel to the line connecting the two cameras or more generally along what are called *epipolar* lines when lens and perspective distortions are taken into account. This zero-crossing primitive provides locally optimized reference points distributed uniformly over the image.

The second derivative alone, however, is a high pass operator so its zeros would be correlated with the fine scale structure of the images. To obtain the required scale specificity, a low pass filter can be used to first attenuate the undesired high frequency detail. A convolution operator that comes close to the combined requirements of attenuating high spatial frequencies while preserving the geometric structure at coarser scales is the two-dimensional Gaussian.

The matching primitive was therefore defined as zero-crossings in the second derivative of the image after Gaussian smoothing, with the primitive scale determined by the space constant of the Gaussian. Differential operators commute with convolution, so the above can also be formulated as zero-crossings in the image after convolution with the function:

$$\frac{\partial^2}{\partial x^2} G_\sigma(x, y) \qquad (1)$$

where $G_\sigma(x, y)$ is a two-dimensional Gaussian with space constant $\sigma$.

Marr and Hildreth[20] later showed that zero-crossings in the Laplacian of a Gaussian convolved (low-pass filtered) image avoided some technical problems associated with the use of directional linear filters. Moving the Laplacian inside the convolution, as above, yields a circularly symmetric linear operator:

$$\begin{aligned}\nabla^2 G &= \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)G(x,y)\\ &= \left(1 - \frac{4r^2}{w^2}\right)e^{-\frac{4r^2}{w^2}} \qquad (2)\end{aligned}$$

where $G$ is a two dimensional Gaussian function, $w$ is the diameter of the positive central region of the operator and is proportional to the Gaussian's space constant, and $r^2 = x^2 + y^2$.

The circularly symmetric $\nabla^2 G$ operator models closely the impulse response of some retinal ganglion cells.[21−23] Its shape is also closely approximated by a difference of two Gaussians with different space constants but normalized volumes—this property permitted a substantial simplification in the design of a $\nabla^2 G$ convolver.[24]

### 1.2.2. Problems encountered

A program written by Grimson[25] to test Marr and Poggio's algorithm on random dot stereograms showed general agreement with the results obtained by Julesz for static patterns. Soon thereafter N. Larson and I designed a high speed digital convolver[24] in hardware for $\nabla^2 G$ convolutions as part of a real-time implementation of the zero-crossing theory. During simulation tests for possible hardware designs for the stereo matcher, I prepared a scene incorporating a complex but known shape—an instant coffee jar—and spattered it with black paint after painting it mat white to provide a good texture for stereo matching. This stereo pair was intended to provide a better feel for natural disparity variations and image characteristics than could be obtained with synthetic random dot patterns. Though the bottle image was carefully produced using 35mm negatives scanned on a high resolution low noise scanning microdensitometer, several important differences with performance on noise free random dot patterns became apparent. Early versions of Grimson's program could match most of the image, but random errors occurred and there was a marked sensitivity to the small vertical misalignments—on the order of a pixel across the image pair—introduced by the imaging and digitization process.

Research at MIT to improve performance to a level adequate for practical application took three directions from this point. Grimson continued development of his program, investigating ways to reduce the error rate in the matching and making use of repetitive runs of his algorithm at each vertical disparity to deal with the vertical alignment problem. In his present algorithm, he switches from area statistics for pruning unacceptable candidate matches to a figural continuity technique following the ideas of Baker and Binford[13-15] and Mayhew and Frisby.[26] The latter approach requires an extended correspondence along the length of a zero-crossing contour for a candidate match to be accepted. Much of the emphasis of this approach has been directed toward the question of filling in the gaps between actual disparity measurements using surface interpolation techniques.

At about the same time, Kass[27] took a different tack, attempting to explicitly handle the vertical disparity problem by using a more elaborate primitive that could be reliably located vertically as well as horizontally. He obtained encouraging results on the bottle image using zero-crossing primitives augmented by contour curvature attributes as well as the orientation information used by Grimson's early program. He later was able to extend his idea to a more general class of pixel based primitives from a study of stochastic models of images. For example, first and second partial derivatives of the image in orthogonal directions and at scales of resolution separated by an octave or more are relatively independent measures. For three scales of resolution a 12 vector is obtained for each pixel in the image. Kass showed that pixel to pixel matching can yield low false positive and negative error rates over a relatively large two-dimensional search space even in the presence of moderate noise.

Both the Grimson and Kass techniques in their present forms appear to perform reasonably well on natural images. They both, however, are computationally demanding algorithms requiring on the order of $mn$ searches for matching each primitive where $m$ and $n$ are the vertical and horizontal disparity ranges respectively in pixels. The processing time for both types of algorithms in optimized lisp and microcode on the MIT Lisp machine for $512^2$ images with 100 pixel disparity ranges is on the order of an hour.

Marr and Poggio's idea of trading off resolution for range seems to be largely abandoned in both techniques. In addition, though not a serious issue for robotics, both approaches appear to have lost much of their biological plausibility in attempting to overcome technical problems.

## 2. The sign representation

The explicit matching of zero-crossing points in binocular stereo may not be the best way to use the information they carry. In the presence of noise the zero-crossing positions are better modeled by probability density functions than as contours at fixed locations. The actual zero-crossing geometry may fluctuate widely within the regions where zero-crossings are likely. Thus to be noise tolerant, a matching algorithm must not be sensitive to such variations. In this section an alternate representation, the convolution sign, is proposed as a more explicit representation of the stable position information present in the $\nabla^2 G$ convolution.

### 2.1. Stability

Noise will cause zero-crossing points to move by an amount proportional to the noise amplitude and inversely proportional to the convolution gradient at the zero-crossing.[28,29] If the spacing between zero-crossings is relatively large compared with this amount of movement, the region of constant sign between zeros will be stable over a large range of signal to noise ratios. The precision to which the positions of these regions can be measured will vary uniformly with the signal quality.

The stability of the convolution sign degrades much more uniformly than does the coherence of zero-crossing contours. This is because the sign map representation imposes a different perspective, that of regions of likely constant sign in the convolved image. Information is still carried at locations where the sign changes, but the focus of the analysis is shifted to the regions rather than to their boundaries.

The effect of noise on sign stability for a particular $\nabla^2 G$ convolution can be calculated easily if the noise and signal are zero-mean Gaussian random processes with relative amplitudes after convolution of $\sigma_s$ and $\sigma_n$ respectively. The probability of a sign change due to noise at any position in such an image is the probability that the noise value is larger than the signal and of opposite sign. The following expression gives the proportion of points in the convolution that are likely to have their sign changed by the addition of noise:

$$\begin{aligned} P &= \frac{2}{\sqrt{2\pi}\sigma_s}\int_0^\infty e^{-\frac{c^2}{2\sigma_s^2}}\frac{1}{\sqrt{2\pi}\sigma_n}\int_c^\infty e^{-\frac{x^2}{2\sigma_n^2}}dxdc \\ &= \frac{1}{\pi}\int_0^\infty\int_{\frac{v\sigma_s}{\sigma_n}}^\infty e^{-\frac{1}{2}\left(u^2+v^2\right)}dudv \\ &= \frac{tan^{-1}\left(\frac{\sigma_n}{\sigma_s}\right)}{\pi} \end{aligned} \tag{3}$$

where $x$ and $c$ are the noise and signal values respectively. Note that even when the noise and signal levels are the same ($\sigma_s = \sigma_n$), P is only $\frac{1}{4}$ ($P = \frac{1}{2}$ when $\sigma_n \gg \sigma_s$).

## 2.2. Autocorrelation of the convolution sign

The $\nabla^2 G$ sign representation is characterized by the autocorrelation function—this also gives us some information about the zero-crossing representation which is difficult to analyze directly. We show here the autocorrelation of the sign of the $\nabla^2 G$ filtered image assuming both white noise and a pink noise model for the image.

Let the image $I(x,y)$ be a Gaussian random process with uniform spectrum and let:

$$C(x,y) = \nabla^2 G * I(x,y) \tag{4}$$

where $*$ denotes a two-dimensional convolution. The autocorrelation of $C(x,y)$ when the image $I(x,y)$ is taken to be Gaussian white noise has the form:

$$R_c(\tau) = k\left(1 - \frac{4\tau^2}{w^2} + \frac{2\tau^4}{w^4}\right)e^{-\frac{2\tau^2}{w^2}} \tag{5}$$

where $k$ is a constant. The autocorrelation $R_s(\tau)$ of the sign of (4), $S(x,y) = sgn(C(x,y))$ obeys an arcsin law when $C$ is a Gaussian random process:[30,31]

$$R_s(\tau) = \frac{2}{\pi} sin^{-1}\left(\frac{R_c(\tau)}{R_c(0)}\right) \qquad (6)$$

A more accurate model of real images[32,33] has two-dimensional power spectrum proportional to:

$$\frac{1}{\left(f_0^2 + f^2\right)^{\frac{3}{2}}} \qquad (7)$$

where $f_0$ is a constant. We call this a *pink* noise model because it is weighted toward the longer spatial frequencies. The autocorrelation function of (7) is proportional to:

$$R_p(\tau) = e^{-\alpha|\tau|} \qquad (8)$$

The autocorrelation of the $\nabla^2 G$ convolution of this noise model can be expressed in terms of (5) as:

$$R_{c'}(\tau) = R_c(\tau) * e^{-\alpha|\tau|} \qquad (9)$$

The principal effect of this convolution is that it broadens the central peak of the function. Figure 3 plots Eq. (6), the autocorrelation function if $S(x,y)$, using $R_{c'}(\tau)$, for the pink noise image model, and compares it with an empirical measurement.

### 2.3. Significance of the sign representation

We can see from the above properties of the autocorrelation function that the $\nabla^2 G$ convolution sign representation is distinct from the raw intensity image and the full $\nabla^2 G$ convolution in three important aspects:

First with regard to *resolution*, $R_s(\tau)$ is sharply peaked and so is capable of a high resolution disparity measurement. This is also the case for $R_p(\tau)$, direct correlation on the intensity image, but it is not true for correlation on the full convolution, $C(x,y)$, since $R_c(\tau)$ has a Gaussian shape near the origin. The sharpness of $R_s$ is due to the nonlinear $sgn$ function which ties all information in the signal to the zero-crossing locations.

Second, the detection *range* for finding the direction (in disparity) towards the correlation peak from single measurements is determined by the width of the

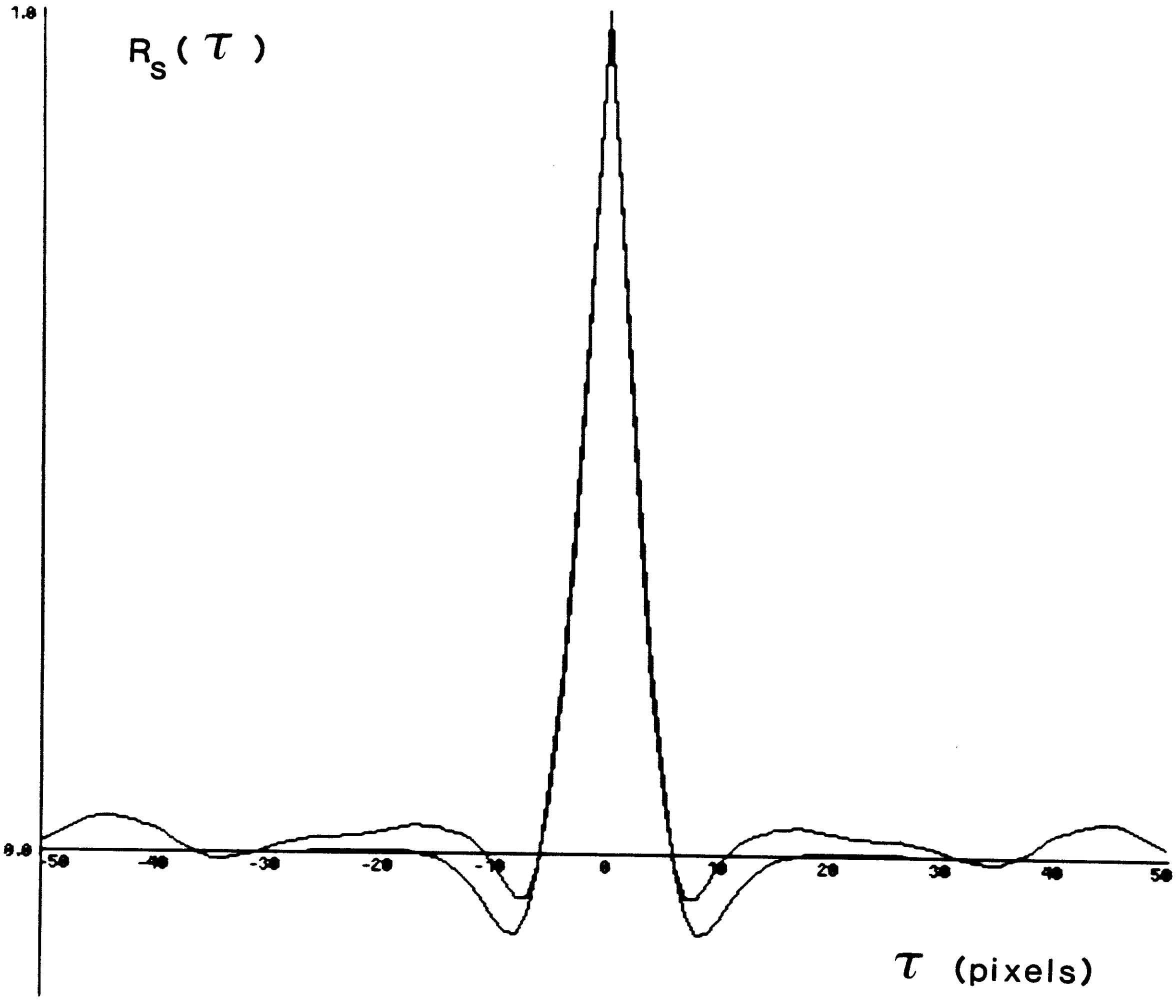

Figure 3. $R_s(\tau)$ using the pink noise image model of Eq. (8) (with $\alpha = 0.2$ and with $w = 8$ pixels). The value for $\alpha$ was measured from the right image of figure 4(b). This curve is overlayed with an empirical measurement of the autocorrelation of the corresponding sign array from figure 5(b).

autocorrelation peak. This width is controllable—a function of the convolution operator size ($w$)—for the full convolution and the sign representations, but not for the raw intensity image.

Finally, regarding *confidence*, $R_s(0) = 1$ independent of local image properties about the point of measurement such as mean intensity or contrast. This is not the case for the autocorrelation functions of either the raw image or the raw $\nabla^2 G$ convolution. This property of $R_s$ makes it possible to assess the significance of a correlation measurement $R_s(\tau_0)$ at an unknown $\tau_0$ directly.

## 3. A near/far module

The preceding results are used here to design an efficient module for determining the two-dimensional displacement, $\tau_0$, between patches out of the left and right images. This *near/far* module does not do point by point search. Instead, a single correlation measurement is made at a test disparity (provided as input) and a determination is made as to whether the correlation peak can be near by (within $w/2$). If there is a positive result, several additional correlation measurements are made at neighboring disparities to determine the shape of the correlation function over the test disparity. From this an estimate is made for the disparity at which the correlation peak occurs. The name of the module comes from the work of G. Poggio and Fisher[34] who described a class of neurons in primate visual cortex sensitive to either near or far disparities.

### 3.1. Measurement of physical surface parameters

The gradient of the two dimensional autocorrelation function, $R_s(\tau)$, points toward the proper alignment over a range of about $\frac{w}{2}$ as can be seen in figure 3. Thus a small set of correlation measurements is sufficient to determine whether or not a correlation peak is present within that range along with its direction and approximate distance if there is one. If, on the other hand, it is determined that a peak is not present within the detection range of the method, that fact can be indicated by the module and we will make it the responsibility of the user's program to decide what action to take next.

The principal surface parameter is distance from the cameras which manifests itself as a translational disparity between corresponding patches from the two images. We can also correlate against parameters other than translational disparity. For example, an elevation gradient on the physical surface viewed can be measured by correlations against compressive or shear distortions. These distortions are introduced between the left and right images by horizontal and vertical elevation gradients.

The autocorrelation surface $R_s(u,v)$ can be approximated by a cone of the form $\phi = 1 - a\sqrt{u^2+v^2}$ for points $(u,v)$ within $\frac{w}{2}$ of the origin. In an image coordinate frame where the correlation peak occurs at an unknown disparity $(u_p, v_p)$ and where there is camera noise and geometric distortion, the following gives a better approximation to the shape of the correlation function near $(u_p, v_p)$:

$$\phi = 1 - \sqrt{a(u-u_p)^2 + b(v-v_p)^2 + c} \qquad (10)$$

From Eq. (10) we obtain:

$$\psi^2 = (\phi - 1)^2 = a(u-u_p)^2 + b(v-v_p)^2 + c \qquad (11)$$

A one-dimensional slice through this surface along the u axis has the form:

$$\psi^2 = a(u-u_p)^2 + bv_p^2 + c \qquad (12)$$

If we measure this function at three points along this curve, we can solve for $u_p$. For example using the points $(-1,0), (0,0)$ and $(1,0)$ we get the equations:

$$\psi^2_{-1,0} = a(-1-u_p)^2 + bv_p^2 + c \qquad (13)$$

$$\psi^2_{0,0} = au_p^2 + bv_p^2 + c \qquad (14)$$

$$\psi^2_{1,0} = a(1-u_p)^2 + bv_p^2 + c \qquad (15)$$

This yields:

$$u_p = \frac{\psi^2_{-1,0} - \psi^2_{1,0}}{2\psi^2_{-1,0} - 4\psi^2_{0,0} + 2\psi^2_{1,0}} \qquad (16)$$

and similarly by measuring the correlation at two additional points we get:

$$v_p = \frac{\psi_{0,-1}^2 - \psi_{0,1}^2}{2\psi_{0,-1}^2 - 4\psi_{0,0}^2 + 2\psi_{0,1}^2} \tag{17}$$

### 3.2. Measurement of the autocorrelation

The standard deviation $\sigma$ of an autocorrelation estimate is inversely proportional to $\sqrt{A}$ where $A$ is the area over which the measurement is made. If measurements are made uniformly in a square patch of the image, $\sigma$ will have a $\frac{1}{d}$ dependence where $d$ is the diameter of the patch.

There is also an important dependence on the size of the $\nabla^2 G$ convolution operator. The width of the central peak of the autocorrelation function is approximately $w$—the diameter of the $\nabla^2 G$ operator's positive center. Increasing the size of this operator increases the effective detection range of the measurement. However, it also reduces the independence of measurements at neighboring image points. Thus a larger patch diameter $d$ is required when $w$ is increased to maintain the same level of confidence in the correlation measurement.

If we write the estimate for the autocorrelation $R_s(\tau)$ measured on an $L \times M$ rectangle of $S(x,y)$ as:

$$V(\tau) = \frac{1}{LM} \int_0^M \int_0^L S(x,y) S(x+\tau,y) dx dy \tag{18}$$

and assume that $\tau$ is very large compared with the size of $L$ and $M$, so that the patches compared are uncorrelated, the expected variance of the measurement $V$ becomes:

$$\sigma^2 = \frac{4}{LM} \int_0^M \int_0^L (1 - \frac{v}{M})(1 - \frac{u}{L}) R_s(u,v)^2 du dv \tag{19}$$

Using the pink noise parameters from figure 3 and with Eq. (6) for $R_s$ in Eq. (19) we obtain the approximation:

$$\sigma \approx \frac{.5w}{d} \tag{20}$$

for the case $L = M = d$. Thus a patch diameter of $8w$ will give a standard deviation of about 0.06 for the autocorrelation measurement with Eq. (18).

Another important consideration when estimating the correlation between left and right image patches is the constancy of the disparity over the measurement patch. Both disparity discontinuities and gradients due to surface elevation variations reduce the height of the correlation peak by an amount proportional to the size of the patch, and thus limit the useful patch size.

### 3.3. An example: a surface velocity sensor

Sequential images from a single camera can be compared using the above technique to measure uniform surface translation during the interval between exposures. The above calculation was implemented using a Hitachi CCD area camera and a memory mapped frame grabber designed by N. Larson. The convolutions were carried out in hardware[24] and correlations and arithmetic calculations were done on an MIT Lisp machine. The 5 correlation measurements required for a single disparity measurement using Eqs. (16) and (17) can be accomplished in 32 milliseconds, including the time for convolution, when the measurements are made on a $32^2$ pixel support with $w = 4$ for the convolution operator. This arrangement handles a maximum inter-frame displacement of 2 pixels reliably on textured surfaces which corresponds to an acceleration of $\frac{\Delta x}{\Delta t^2} \approx 2000 pixels/sec^2$ and is capable of measuring small displacements to a resolution of about 0.1 pixel.

This example illustrates the behavior of the basic sensing module. We now apply it to binocular stereo simply by shifting to a two camera system and designing an appropriate control algorithm for using the near/far module.

## 4. The PRISM system

We will now use the near/far module to produce stereo measurements tailored to the requirements of specific tasks in robotics. The first application considered has been the problem of rapidly determining an elevation map over the visual field sufficient for obstacle avoidance tasks. The prime objectives for this are speed and reliability. Spatial and depth resolution requirements, on the other hand, are less stringent than would be the case for tasks such as shape description or part position measurement. In addition to developing control strategies for operating the near/far module, this section discusses techniques for insuring that adequate surface texture will be present on the imaged surfaces and for computing surface elevation from image disparity.

### 4.1. The design task

Our design goal is a system which produces a coarse surface elevation map over the camera field with a $36 \times 26$ tessellation. Its height range should be similar to the field diameter and height should be resolved to about 200 levels over that range. Thus we model the actual surface using square prisms with varying heights. We seek reliable measurements and so must pay attention to the height of the correlation at each peak found by the near/far module. The nature of the correlation measurement causes this method to be blind to surface details smaller than the patch size on which measurements are made. It will also fail to find a surface—though it will not make a false measurement—when surface orientations or heights are outside the matcher's design limits. To minimize errors in such situations, we further require the algorithm to abstain from reporting a prism elevation unless its measurement is solid. A user's program can either avoid suspect regions or employ alternate methods if they are available for rechecking those surface patches. This allows greater design simplicity without jeopardizing reliability.

### 4.2. Unstructured light

A potential handicap of binocular stereo as compared with structured light approaches for robotics vision is its dependence on surface markings for making range measurements. Industrial parts and surfaces are often without dense surface textures that can be registered. Light stripe or structured light techniques have the opposite problem. They suffer ambiguity problems when dealing with objects that have high contrast surface markings. Repeated surface markings can also cause serious problems for binocular stereo matching since false matches appear as good as the correct ones locally. Both situations—clean surfaces and regular patterns—are especially common in man-made environments.

These problems can be dealt with easily in most robotics tasks by illuminating the workspace with a suitable random texture pattern with a projector situated near the cameras as shown in figure 1. Unlike structured light techniques, our matching system begins with no a priori knowledge of the surface markings it is to use. It does not matter whether the markings are natural or artificially produced and so the projected pattern can mix with any texture markings already present on the imaged surfaces with no adverse consequences. A binary random dot pattern was generated with a 50 percent dot density on a computer and a 35mm slide was produced of the pattern from a high resolution display. A standard slide projector was then used to project this pattern onto the workspace. This produced a marked improvement in the signal to noise performance of the system even for already textured surfaces.

Figure 4 shows an example of this unstructured light technique in use. Figure 5 shows the convolution sign representations obtained from figure 4(b) with $w = 16$, 8, and 4 pixel $\nabla^2 G$ convolution operators. White and black indicate locations where the convolution was positive or negative respectively.

### 4.3. Control strategies

A simple way to use the near/far module to produce a disparity map for a pair of stereo images would be to apply it iteratively in a triple nested loop indexed over image position and disparity. The measurement patch size $d$ must be chosen to be sufficiently small to give adequate spatial resolution. The operator size ($w$) must be small enough relative to $d$ to give a good correlation estimate. These choices determine the number of steps required in the iteration. For a $32^2$ pixel measurement patch, a $10 \times 10$ operator ($w = 4$) will provide a low correlation variance and has a reliable detection range of $\pm 2$ pixels. If the camera geometry is arranged so that the elevation range requirement corresponds to a disparity range of 200 pixels, then as many as 50 correlation checks would be required at each image patch to determine the disparity there. If each near/far check takes 32 milliseconds, a $36 \times 26$ matrix of patches covering the image with an interpatch spacing of 16 pixels would take 1500 seconds to compute, in the worst case.

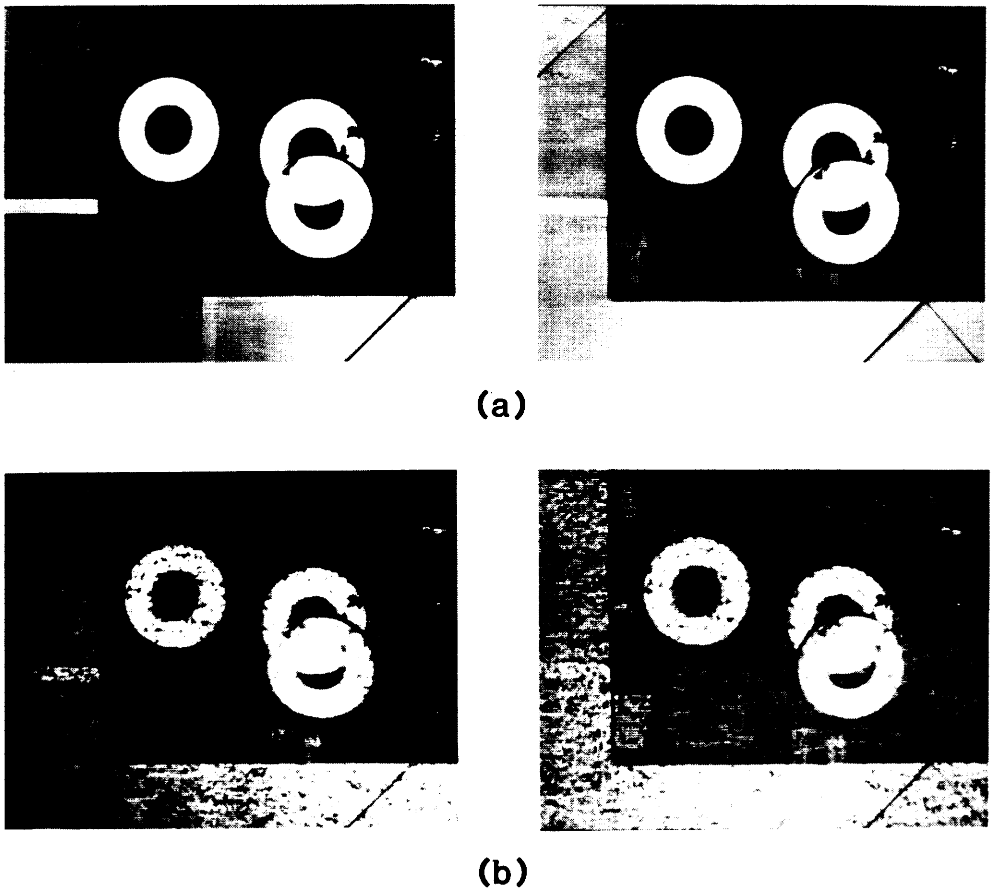

Figure 4. (a) Stereo pair of three plastic doughnuts approximately 10cm in diameter on a dark board 2.5cm thick. Aside from some tape marks and paint chips on the doughnuts, there is little texture present on the surfaces suitable for stereo matching. The cameras are 150cm above the work table and separated by 40cm. Low cost vidicons are used with 25mm lenses. A single $\frac{1}{30}$ second tv frame is grabbed into a 576 × 454 array for each image and redisplayed here. (b) The same stereo scene illuminated with a random dot texture pattern from a projector located near the TV cameras. This *unstructured light* technique provides a dense high contrast surface texture to drive stereo matching. This enables the matcher to operate with little dependence on the sample material or its surface markings.

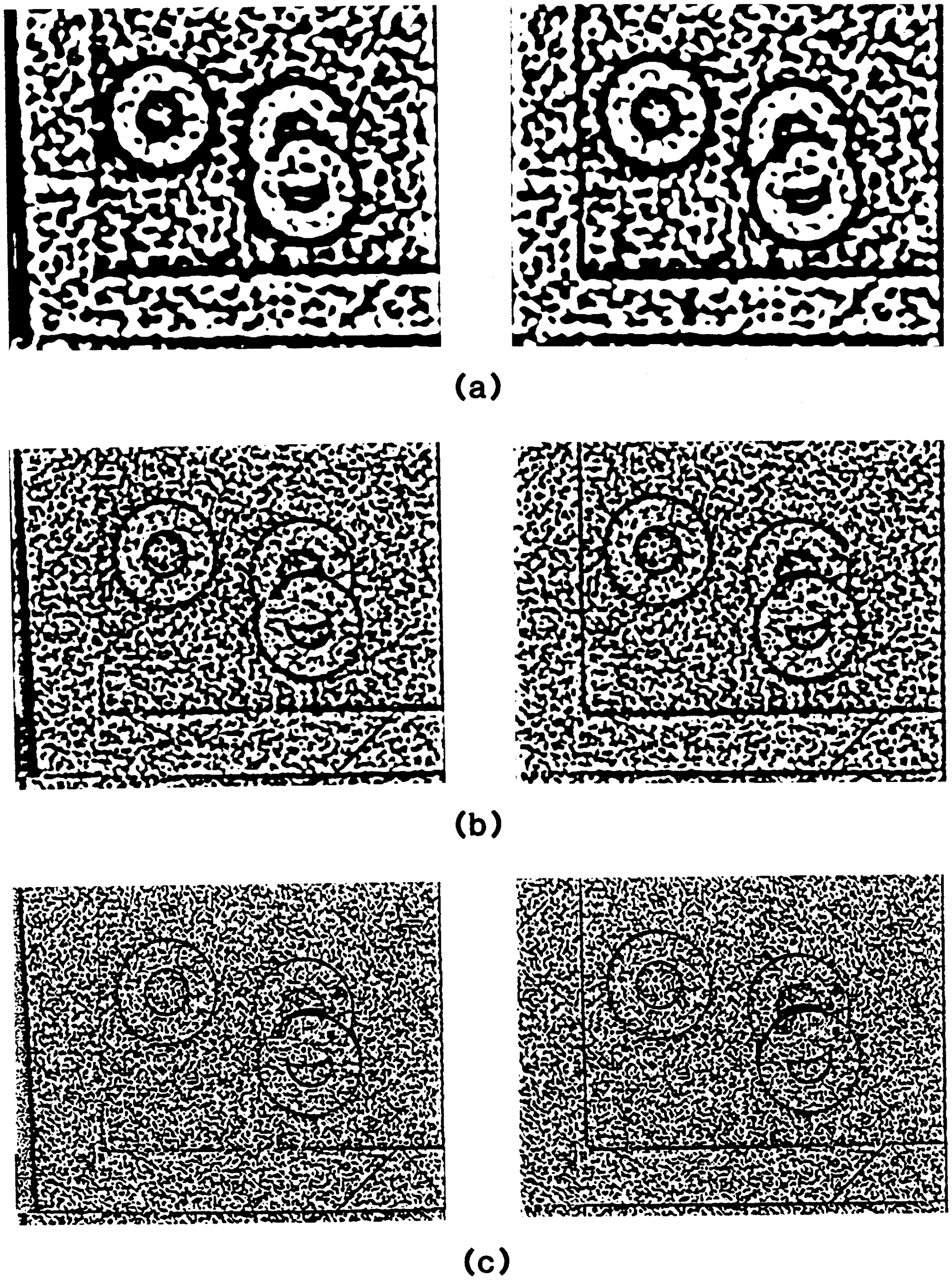

Figure 5. The images of figure 4(b) convolved with a difference of gaussian operator at three scales. The sign of the convolution is shown here using white and black to indicate positive and negative regions respectively. The first pair (a) is with a $32^2$ operator ($w = 16$ pixels); (b) is with a $20^2$ operator ($w = 8$ pixels); and (c) is produced with a $10^2$ operator ($w = 4$ pixels). The convolutions are carried out digitally in a pipelined convolver designed by N.Larson and the author. This set of 6 convolutions is accomplished in 1.5 seconds.

This search time can be reduced substantially by first doing a coarse resolution pass with a larger convolution operator and proportionally larger measurement patches as indicated in figure 2. With a $w = 16$ operator on $128^2$ patches, the detection range of the near/far module is $\pm 8$ pixels so at most only 13 checks are required at each patch location. The near/far module in its present implementation takes about the same amount of time for a check at this scale because the brunt of the computation time is in the convolution which is fixed at 1 microsecond per output point independent of the mask size. Thus in the worst case a $9 \times 7$ matrix covering the whole field can be computed in about 26 seconds. In practice, neighboring patches are often at similar elevations and taking advantage of this reduces the time required considerably—down to 4 seconds on the average.

Once elevation measurements are obtained at the coarse scale, a second pass is made with $w = 8$ pixel convolutions and $64^2$ patches, followed by a third pass with $w = 4$ pixels and $32^2$ patches as shown in figure 6. In most cases, a single near/far check is required at each patch location. The algorithm presently takes 30 to 40 seconds to produce a $36 \times 26$ matrix of disparity measurements using this three scale coarse-to-fine control.

## 4.4. Calibration

Disparities measured between the two cameras must be transformed to physical elevation values. This transformation follows the approximate relation:

$$\frac{\Delta\theta}{\theta} \approx -\frac{\Delta D}{D} \tag{21}$$

where $\theta$ is the vergence angle of the cameras and $D$ is the distance to the surface viewed (see figure 1). From this we can derive the relation:

$$\frac{\Delta S}{\Delta D} \approx \frac{l}{D} \tag{22}$$

where $\Delta S$ is the spatial pixel resolution in mm, $\Delta D$ is the depth resolution also in mm, and $l$ is the camera to camera separation. The desired absolute transformation to a physical height map however is complicated by the geometric distortion introduced by the cameras and the imaging geometry. In particular, a relative magnification

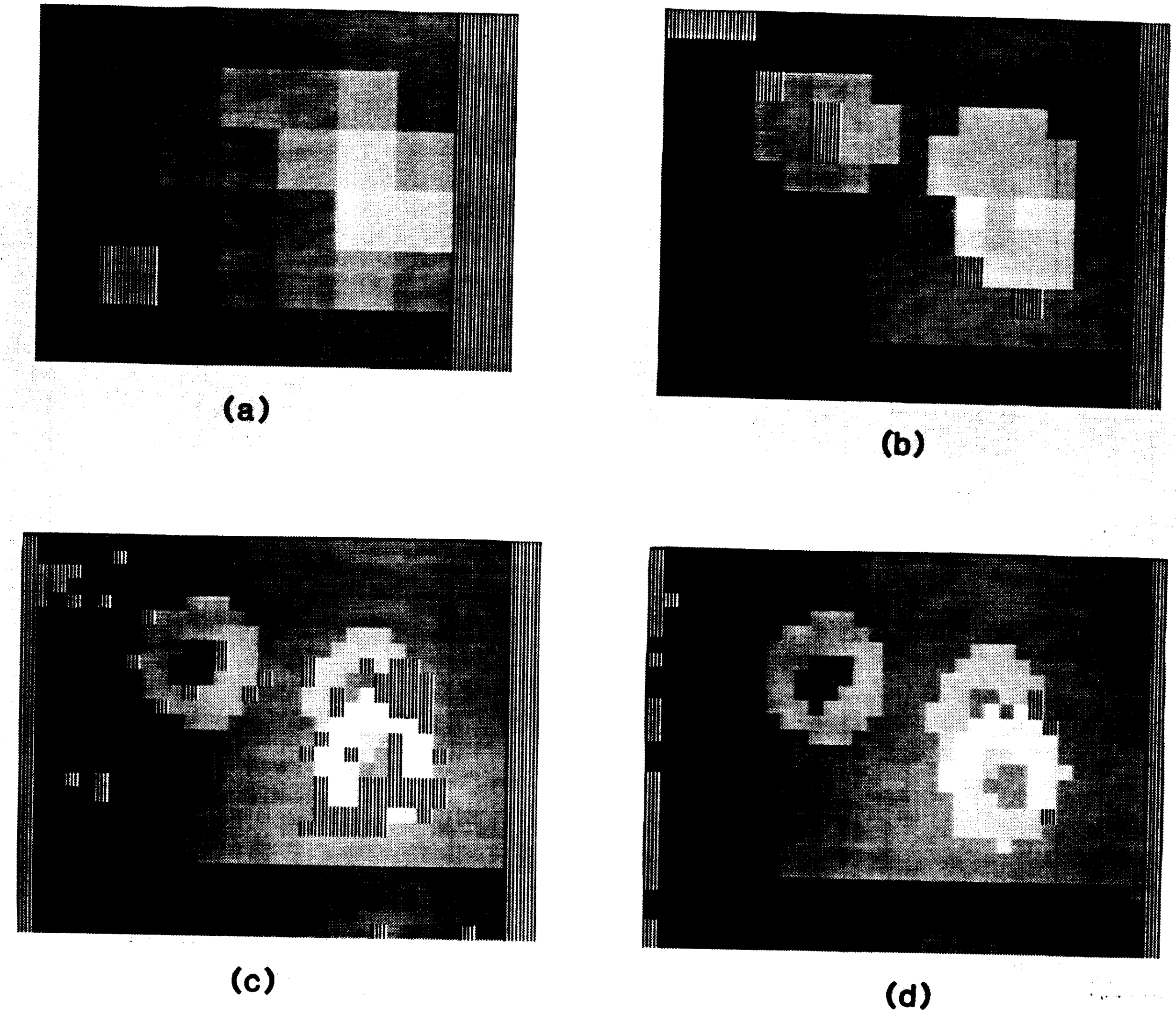

Figure 6. Disparity measurements produced by the PRISM matcher from the data in figure 5. Shading indicates the disparity magnitude—lighter means near. The vertical stripe texture is used to indicate patch positions at which no satisfactory correlation was obtained. The measurement patches overlap by 50 percent in both dimensions. (a) is obtained from figure 5(a) using $128^2$ patches; (b) from figure 5(b) using $64^2$ patches; and (c) from figure 5(c) using $32^2$ patches. In (d) the possibility that an elevation discontinuity passes through the unmatched patches in (c) is checked. In such cases, the side of the discontinuity having greater area in the patch is used.

between the left and right images is induced by perspective effects whenever there is a difference in distance from the target to the two cameras.

The PRISM system makes a restricted set of measurements, so a simple lookup table can be used for the disparity to height transformation. The present calibration procedure takes two disparity maps obtained from running the system on a flat surface at 0mm elevation and at 300 mm. The matcher produces both vertical and horizontal disparity measurements over both of these planes. A linear disparity to height mapping is computed from these calibration planes for each patch position of the $36 \times 26$ disparity array. A similar linear function is produced for estimating vertical disparity as a function of horizontal disparity at each patch position. The resulting system tolerates large camera and imaging distortions so long as these distortions are stable. At present, position measurements are accurate to 10mm in all three dimensions over the entire operating volume. Figure 7 illustrates the elevation values obtained by this method from the disparity measurements displayed in figure 6(d). Elevation measurements are repeatable to 2mm and absolute calibration to that precision should be possible with a more elaborate transformation table.

### 4.5. Test applications

The PRISM stereo matcher has been used as a vision input to two robotics manipulation systems with different application requirements.

#### 4.5.1. Brooks' path planning system

In this experiment, the PRISM system was used in a two channel configuration to provide an elevation map of a PUMA workspace to a collision avoidance system developed by Brooks.[35] The combined system was presented with the task of moving a part from a predefined workspace location to another. Large obstacles were to be placed at random in the workspace prior to the movement and the system was responsible for measuring work space elevation with PRISM and planning a trajectory that would safely accomplish a pick and place task free of collisions with any part of the manipulator or its payload. Brooks' program constructed a polyhedral model of the PRISM elevation array and mapped that into the manipulators configuration space where a search for a collision free path was made.

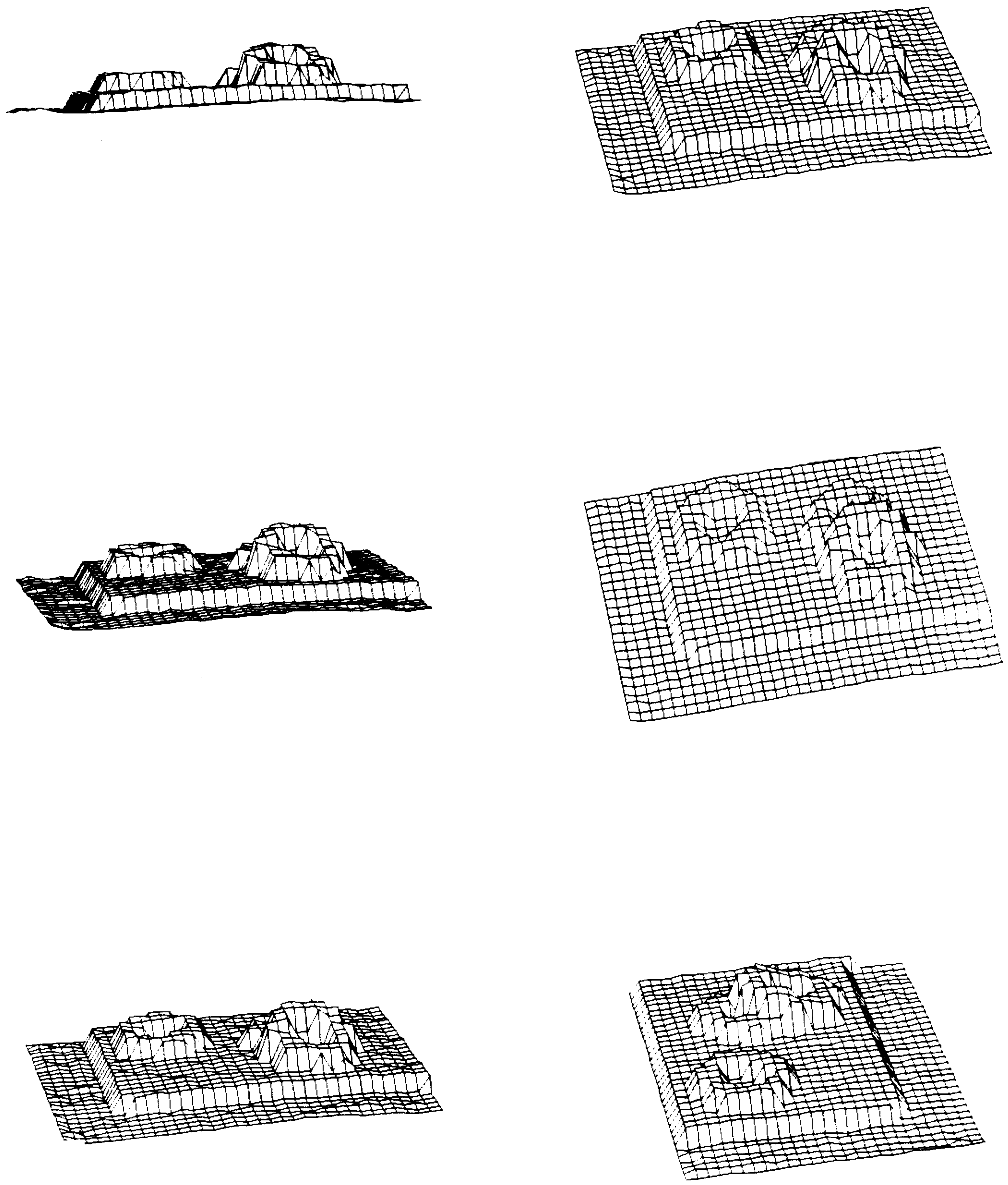

Figure 7. Perspective displays of data from figure 6(d) after conversion from disparity to physical height. Height variations along the edges of flat surfaces gives a rough measure of the repeatability of the measurements. The total time required for the process, from taking the pictures to having the elevation matrix shown here, was 30 seconds.

The vision system and the path planning system were developed separately and they were operated in parallel on separate Lisp machines with communication over a local area network. Interfacing PRISM with Brooks' path planner required little work and the first demonstration was successfully completed the first day the two systems were tested together.

### 4.5.2. Ikeuchi's bin of parts system

A second collaboration incorporated the PRISM system into Ikeuchi's bin picking system.[36,37] He uses photometric stereo[38–43] to measure local surface orientation at each pixel of the image. The technique uses three images all taken with the same camera but with the light source in a different position for each. The local surface orientation information is first used to segment the image into regions of continuous surface. A histogram of surface orientations—the *extended gaussian image* (EGI)—can then be made for each of these regions and used to recognize parts out of a catalogue of known shapes and determine their orientation in space. The photometric stereo technique, however, is not capable of measuring range and this information was provided by the PRISM system.

There were three phases in the bin picking study. The first was to pick up plastic doughnuts off of the top of the pile and stack them on a post. Information from the PRISM system was used in this case to determine the height of the selected part so that the hand could be made to approach that part along a trajectory different from the line of sight. The highest part of the doughnut was selected for the grasp point with no other precautions taken to avoid collisions with neighboring doughnuts. Figures 4-7 illustrate the processing from input, convolution, matching, and final elevation for a typical scene used in this phase. In the second phase, finger clearance was measured using the PRISM data to select the best grasp point around the doughnut circumference. This was done by projecting the finger *foot prints* along the direction of the approach trajectory until the first surface element in the PRISM elevation map was encountered. In the third phase, the clearance test was further elaborated to consider three different approach angles at each point around the doughnut circumference.

In both experiments, the PRISM system was used independently by Brooks and Ikeuchi as a tool in their respective systems. Except for a single change to the correlation threshold used to eliminate uncertain measurements, no significant modifications were made to the PRISM algorithm during the two month period when these demonstration systems were in operation. Over that period several hundred runs were made of the PRISM program. Failures to operate properly occurred infrequently and were due almost always to a failure of a mechanical relay to switch the frame grabber between left and right images at the proper time.

A further test was done running the PRISM program repeatedly 500 times (4 hours run time) without changing the scene—though room lighting and building motion with trains passing outside could not be controlled—and collecting a histogram of horizontal disparity measurements at each patch position in a scene like that shown in figure 4. The standard deviation about any clear surface position was less than a pixel in disparity about the central mean and no matches occurred to disparities more than a few pixels away from a disparity actually present in the measurement patch. Bimodal disparity distributions occurred in patches straddling surface height discontinuities because two correlation peaks occur at such locations.

## 5. Discussion

We have designed a simple and robust near/far stereo module following Marr and Poggio's idea of trading resolution for range. By restricting the analysis to a specific scale in the stereo images, disparity displacements comparable to that scale can be measured without search but detail much finer than that scale are lost. Operating the scale-specific module at several scales in a coarse-to-fine progression, allows performance to be tailored to the range and resolution requirements of specific applications. Two results of this work are important, first that a minimal mechanism[3] like the one proposed here may be capable of explaining much more complex aspects of stereo performance, and second that the sign of the $\nabla^2 G$ convolution can be matched more reliably in practice than is the case for the explicit matching of zero-crossings.

As illustrated in the previous section, surface topography information from a stereo vision system can be used under a broad range of operating conditions to successfully guide robotic manipulation systems in part position measurement and obstacle detection and avoidance tasks. Single range measurements can already be made in a fraction of a second and it appears that computation of an elevation map like that shown in figure 7 can be brought into the same time range with relatively simple special purpose hardware for the near/far module.

We are now working on the design of a surface proximity detector based on the near/far module. It will be attached to a manipulator hand to measure hand-to-part position relations with 10 to 30 near/far measurements per second. The device will be used to measure surface range, orientation, and flatness in a small field in front of the sensor. This measurement in conjunction with an edge position measurement may be capable of obtaining very high precision position and orientation. The problem of detecting depth discontinuities to support this task is presently under study using techniques analogous to those used by the near/far module.

## 6. References

1. Marr, D., Poggio, T., Proc. R. Soc. Lond. B 204, 301-328(1979).

2. Nishihara, H.K., Poggio, T., Proc. Int. Symp. Rob. Res., Bretton Woods, NH.(1983) to appear in MIT Press.

3. Marr, D., Nishihara, H.K., Proc. Roc. Soc. B 200, 269-294(1978).

4. Nishihara, H.K., Artificial Intelligence 17, 265-284(1981).

5. Nishihara, H.K., Physical and Biological Processing of Images, Eds.O.J.Braddick and A.C.Sleigh. Springer-Verlag, 335-348(1983).

6. Kelly R.E., McConnell P.R.H., Mildenberger S.J., Photogramm. Eng. Rem. Sens. 43, 1407-1417(1977).

7. Moravec, H. P., Ph.D. thesis, Stanford University, Stanford Artificial Intelligence Memo 340(1980).

8. Gennery, D. B., Ph.D. thesis, Stanford University, Artificial Intelligence Laboratory Memo 339(1980).

9. Tsai, R.Y., IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-5, 2, 159-173(1983).

10. Horn, B.K.P., Photogrammetric Engineering and Remote Sensing, 49, 535-536(1983).

11. Arnold, R. D., Proc. ARPA Image Understanding Workshop, L. Baumann, ed., Science Applications, Inc., 65-72(1978).

12. Arnold, R. D., Binford, T. O., SPIE, 238, 281-292(1980).

13. Baker, H. H., Proc. ARPA Image Understanding Workshop, L. Baumann, ed., Science Applications, Inc., 168-175(1980).

14. Baker, H. H., Ph.D. thesis, University of Illinois(1981).

15. Baker, H. H., Binford, T. O., Proc. 7th Intern. Joint Conf. on A. I., Vancouver, British Columbia, 631-636(1981).

16. Ohta, Y., Kanade, T., Carnegie-Mellon University Dept. of Computer Science Memo CMU-CS-83-162.

17. Julesz, B., Foundations of cyclopean perception, Chicago:University of Chicago Press(1971).

18. Poggio, G.F., Poggio, T., Ann. Rev. Neurosci. 7, 379-412(1984).

19. Marr, D., Poggio, T., Science, 194,283-287(1976).

20. Marr, D., Hildreth, E., Proc. R. Soc. Lond. B, 207, 187-217(1980).

21. Rodieck, R.W., Stone, J., J. Neurophysiol., 28, 833-849(1965).

22. Ratliff, F., Mach Bands: quantitative studies on neural networks in the retina, San Francisco: Holden-Day(1965).

23. Enroth-Cugell, C., Robson, J.G., J. Physiol. Lond., 187, 517-552(1966).

24. Nishihara, H.K., Larson, N. G., Proc. ARPA Image Understanding Workshop, L. Baumann, ed., Science Applications, Inc., 114-120(1981).

25. Grimson, W.E.L., From Images to Surfaces: A Computational Study of the Human Early Visual System., Cambridge, Mass.:MIT Press(1981).

26. Mayhew, J. E. W., Frisby, J. P., Artif. Intell. 17, 349-385(1981).

27. Kass, M., Proc. ARPA Image Understanding Workshop, L. Baumann, ed., Science Applications, Inc.,(1983).

28. Nishihara, H.K., SPIE, 360, 76-87(1982).

29. Nishihara, H.K., Poggio, T., Nature, 300, 347-349(1982).

30. Lawson, J.L., Uhlenbeck, G.E., 1950. Threshold signals. McGraw Hill.

31. Papoulis, A., Probability, random variables, and stochastic processes, McGraw Hill(1965).

32. Netravali, A., Limb, J., Proc. IEEE 68, 3(1980).

33. Kass, M., IEEE International Conf. on Systems, Man, and Cybernetics. Bombay and New Delhi, India(1983).

34. Poggio, G.F., Fisher, B., J. Neurophysiol. 40, 1392-1405(1977).

35. Brooks, R. A., Proc. Int. Symp. Rob. Res., Bretton Woods, NH.(1983) to appear in MIT Press.

36. Ikeuchi, K., Horn, B.K.P., Nagata, S., Callahan, T., Feingold, O., Proc. Int. Symp. Rob. Res., Bretton Woods, NH. to appear in MIT Press. Also available as Massachusetts Institute of Technology Artificial Intelligence Laboratory Memo 718(1983).

37. Ikeuchi, K., Nishihara, H.K., Horn, B.K.P., Sobalvarro, P., Nagata, S., Massachusetts Institute of Technology Artificial Intelligence Laboratory Memo 742(1984).

38. Horn, B.K.P., Woodham, R.J., Silver W.M., Massachusetts Institute of Technology Artificial Intelligence Laboratory Memo 490(1978).

39. Woodham, R. J., SPIE 155, 136-143(1978).