Motion Stereo Using Ego-Motion Complex Logarithmic Mapping[1]

Ramesh Jain, Sandra L. Bartlett
and
Nancy O'Brien

Electrical Engineering and Computer Science
The University of Michigan
Ann Arbor, MI 48109

February 1986

CENTER FOR RESEARCH ON INTEGRATED MANUFACTURING

Robot Systems Division

COLLEGE OF ENGINEERING

THE UNIVERSITY OF MICHIGAN

ANN ARBOR, MICHIGAN 48109-1109

This work was partially supported by NSF Grant no. MCS8219739 and AFOSR contract no. F49620-82-C-0089.

Abstract

Stereo information can be obtained using a moving camera. If a dynamic scene is acquired using a translating camera and the camera motion parameters are known, then the analysis of the scene may be facilitated by Ego-Motion Complex Logarithmic Mapping (ECLM). It is shown in this paper that by using the Complex Logarithmic Mapping (CLM) with respect to the *focus of expansion*, the depth of stationary components can be determined easily in the transformed image sequence. The proposed approach for depth recovery avoids the difficult problems of establishing correspondence and computation of optical flow, by using the ego-motion information. An added advantage of the CLM will be the invariances it offers. We report our experiments with synthetic data to show the sensitivity of the depth recovery, and show results of real scenes to demonstrate the efficacy of the proposed motion stereo in applications such as autonomous navigation.

TABLE OF CONTENTS

1. INTRODUCTION

Depth determination is a continuing problem in computer vision. Research in this area is motivated by the perceived need for autonomous vehicles, vision for the blind, realistic flight trainers, models of the human visual system, etc. There is a plethora of depth determination techniques. Many different stereo systems for depth determination have been developed just in the last few years. The Marr-Poggio [Mar82] theory of human stereo vision forms a basis for several of them. Grimson [Gri81] implemented a version of it by using the Laplacian of the Gaussian to find matchable elements - zero-crossings of the same sign and similar orientation - in each of two images. Four different filter sizes were used and matches with the coarser filters were used to limit the search area for matches in images from the finer filters. Disparity of the matched points in the finest filter were used to determine depth. Smitley and Bajcsy [SmB84] use a method similar to Grimson. However, they use a DOG operator on an image smoothed with a non-linear filter which they designed, and a two stage matching process. Burr [Bur84], Ikeuchi [Ike84], and Grimson [Gri83] address the problem of recovering surface details by modifying the Marr-Poggio approach. Homma and Fu [HoF84] match regions in two stereo images instead of lines or points, where a region corresponds to an object surface having a continuous depth and change of depth.

Most of the above schemes use the human visual system as a model for camera location. Many other camera geometries have been reported in the literature. Wu, Wang, and Bajcsy [WWB83] and Nevatia [Nev76] obtain multiple images of the object for 3-D data acquisition by placing the object on a turntable and rotating it. Luh and Klaasen [LuK79] describe a system for robot collision avoidance in a work area. Three cameras are mounted orthogonally so they can each view the work area. Matthias and Thorpe [MaT84] used the algorithm for the Stanford Cart as a basis for developing a robot navigation system. Because of the time required to process the information in the nine images obtained in the original implementation, they used only two images. To compensate for the loss of information, they use information from the previous pair of pictures and added some constraints to the correspondence algorithm . The net result was significantly faster runtime, with slightly degraded performance.

One of the main problems in stereo vision is the time it takes to process the two or more images. Safranek and Kak [SaK83] propose a hardware system to implement the Marr-Poggio method, with provisions for using the peaks and valleys of the filtered images as well as the zero-crossings. A second serious problem encountered in stereo vision algorithms is matching the points, lines, regions, etc. in the two images. Several methods to solve this problem in a reasonable amount of time can be found in the survey article by Baranad and Fischler [BaF82]. Stockman and Esteva [StE84] describe a "cluster space stereo" which simplifies the correspondence problem. They assume an industrial type environment with few objects and known geometrical models of those objects. Berthod and Long

[BeL84] solve the correspondence problem by mapping it to a graph matching problem which they then solve using a parallel optimization method.

Gu, Yang, and Huang [GYH84] suggest a matching method for two images of a moving object. Their matching is based on circuits in the graph representation of objects. The method can be directly applied to stereo matching, however. Nishihara [Nis84] has developed a real-time, noise tolerant stereo matcher. To achieve noise tolerance, he uses the sign of the convolution, rather than the zero-crossings, as these degrade more gracefully in the presence of noise.

Stereo information can also be obtained using a single moving camera. Itoh, Miyauchi, and Ozawa [IMO84] present a method of obtaining depth information using a single camera. The camera is constrained to move along its visual axis. Corners of a cube are used as the points to be matched. Depth information is obtained from camera parameters and the geometry of the situation. Zacharias, Caglayan, and Sinacori [ZaC85,ZCS83] use a relaxed version of this camera/scene relationship to model a pilot's ability to navigate "by the seat of his pants". The problem they address is estimating self motion, which is essentially the dual of object depth determination. They assume a stationary, textured visual field. N points in the visual field are considered. For each one an "impact time" is calculated, that is, the amount of time until the pilot would crash into the point if he were aimed at it. This quantity is derived from a unit length line-of-sight vector and a rate of change of this vector over time.

Jain and O'Brien [JaO84, OBJ84] initiated some studies using the same camera configuration that Itoh, et al, describe. However, they approached this

problem very differently; they map images into a complex log space where the movement of the objects in two dimensions due to the camera motion becomes a translation along one axis in the new space. Given this phenomenon, the correspondence problem is greatly reduced, since only a small strip of the new space needs to be searched. This constraint is similar to the epi-polar constraint in stereo. In addition, the transform is scale and rotation invariant. It also has an analogue in the human visual system - the mapping of the retinal space into the striate cortex is very closely approximated by the CLM.

Recently, many researchers have been studying systems which acquire images using a moving camera. Optical flow has been studied with the aim of recovering information about the environment and the motion of the observer. In this paper, we discuss some characteristics of optical flow that are useful in analyzing dynamic scenes, and then discuss briefly the Ego-Motion Polar (EMP) transformation that is useful in separating stationary and nonstationary components of a scene acquired using a translating camera [Jai83]. Next we present some aspects of the CLM and then introduce the Ego-Motion Complex Logarithmic Mapping (ECLM). We show that the ECLM maintains projection invariance for arbitrary translational motion of the observer and can be used in recovering the depth of stationary objects. We also present aspects of the mapping that are useful in finding the precision of the information that can be recovered using this mapping. Finally we discuss our experience in recovering depth in laboratory scenes. Our aim in this paper is to show that motion stereo, in the case of a translating observer, can be efficiently implemented in the ECLM space.

2. EGO-MOTION POLAR TRANSFORM

It has been shown that optical flow carries information about the structure of the environment and the motion of the observer [Clo80,Gib79,Lee80,Pra80]. Optical flow is the instantaneous field of velocity at image points due to the motion of scene points relative to the viewer. When an observer moves through a scene, all points in the scene are in motion relative to him. Points which are close to the observer appear to move relatively faster than points which are further away. The flow vectors due to the stationary components of a scene intersect at a point, as shown in Figure 1. This point is called the *Focus of Expansion* (FOE). It has been shown that the FOE plays a vital role in the recovery of information from the optical flow field [Lee80, Pra80].

The last few years have seen increasing efforts to use optical flow in the analysis of dynamic scenes. Many approaches have been proposed for the computation of optical flow from images of a dynamic scene. Most of these proposed approaches consider two frames of a scene to compute the flow field. Based on the research reported in the literature, it appears that the computation of acceptable quality optical flow for real world scenes is a very difficult problem [BrH83]. Moreover, the methods for the recovery of the information are sensitive to the noise in the optical flow. Thus, in most realistic applications, the information obtained from the computed flow fields may not be reliable.

In many applications, such as robot vision, either the camera moves under computer control or the camera motion parameters can be obtained using some means. If the camera displacement between two frames is dX, dY, dZ, then the

FOE for these frames is

$$F_x = \frac{dX}{dZ} \tag{1}$$

$$F_y = \frac{dY}{dZ} \tag{2}$$

The information about the camera motion parameters, and hence about the FOE, may be used in the recovery of information. Jain tried to exploit characteristics of optical flow without actually computing the optical flow field [Jai83, Jai84]. He used a transformation on images acquired using a moving camera to segment a dynamic scene into its stationary and nonstationary components. This transformation, called the Ego-motion Polar transform (EMP), is centered around the FOE and converts the original image $I(x,y)$ into an image $I^*(r,\theta)$ using

$$I^*(r,\theta) = I(x,y) \tag{3}$$

where

$$r = \sqrt{(x-F_x)^2+(y-F_y)^2} \tag{4}$$

and

$$\theta = arctan\frac{(y-F_y)}{(x-F_x)} \tag{5}$$

This transformation is shown in Figure 2.

It is shown in [Jai84], that for a moving observer, all stationary points in a scene will show only horizontal displacement in the EMP transformed image. This fact can be used to determine whether an object is moving or not. As shown in Figure 3, the apparent motion of stationary points is converted from

assorted directions, depending on their locations in the image plane, to unidirectional motion in the EMP space. The stationarity of an object is judged by the absence of a θ component for the region corresponding to the object in the EMP space. An algorithm was developed to implement this scheme. The results for real world scenes are reported in [Jai84].

3. COMPLEX LOGARITHMIC MAPPING

Schwartz showed that the *retino-striate mapping* can be approximated using a Complex Logarithmic Mapping (CLM). Retino-striate mapping, a common feature of vertebrate sensory information processing, is a spatial mapping of the peripheral sensory receptive surfaces onto corresponding parts of the central nervous system [Sch77]. In our own human vision system, as well as those of lower animals, it has been found that the excitement of the striate cortex can be approximated by a Complex Logarithmic Mapping (CLM) of the eye's retinal image. In other words, what we see as the real world and what is focused on the retinas of our eyes, is reconfigured onto the striate cortex by a process similar to complex logarithmic mapping [Sch77,Sch80,Sch81,Sch82] before it is examined or interpreted in our brain. Schwartz further argued that this mapping is responsible for the scale, rotation and projection invariances in the human visual system. As is well known, these invariances play a vital role in human visual perception. Cavanaugh [Cav78,Cav81], however, showed that Schwartz's claims about the CLM resulting in the invariances are correct only under certain conditions. The rotation and scale invariances are obtained if the object is in the center of the

image and the rotation and scale changes are with respect to the origin. The projection invariance is obtained only if the direction of the observer's gaze and motion are the same.

Let us look at the mathematical definition of CLM. Complex log mapping may be written mathematically as

$$w = \log z \tag{6}$$

where w and z are complex variables:

$$z = x + iy = r(\cos\theta + i\sin\theta) = re^{i\theta} \tag{7}$$

and

$$w = u(z) + iv(z) \tag{8}$$

In this way, a function or image in z-space with coordinates x and y is mapped to w-space with coordinates u and v. The mapping is obtained from the simplified equations:

$$u(r,\theta) = \log r \tag{9}$$

$$v(r,\theta) = \theta \tag{10}$$

There are many attractive features of this mapping [ChW79, BGT79, SaT80]. From the psychological viewpoint, it is the only analytic function which maps a circular region, such as an image on the retina, into a rectangular region. This is a desirable feature for the study and modelling of the human visual system. The mappings of two regular patterns are shown in Figure 4 to result in similarly regular patterns. It is seen in Figure 4a that concentric circles in an

image or the z-plane become vertical lines in the mapped w-plane. This becomes obvious when one examines the CLM definition above. A single circle maps to a single vertical line since the constant radius, r, at all angles, θ, of the circle gives a constant u coordinate for all v coordinates in the mapped space. Similarly in Figure 4b, an image of radial lines which have constant angle but variable radii, result in a map of horizontal lines.

Through these mappings, we can demonstrate some of the invariances of CLM that may be helpful in image understanding. The first such invariance is that of rotation. In Figure 4a, we saw that for a circle, all possible angular orientations of a point at the given radius will map to the same vertical line. Thus, if an object is rotated between successive images, this will result in only a vertical displacement of the mapped image. This same result can be seen in Figure 4b. As a radial line rotates about the origin, its entire horizontal line mapping moves only vertically.

Another characteristic of CLM is size invariance. This also can be seen in Figure 4. As a point moves out from the origin along a radial line in Figure 4b, its mapping stays on the same horizontal line moving only from left to right. The mappings of the concentric circles of Figure 4a remain vertical lines and only move horizontally as the circles change in size.

A third important invariance is that of projection. When an observer translates in space, the images of objects appear to remain unchanged. Thus, though the images of stationary objects do change on the retina, the object perceived on the striate cortex does not change. This is due to the fact that in the

CLM space, translation of the observer only causes the object image to be displaced in the horizontal direction; the size and shape of the object image remain unchanged.

These invariances may be useful for object recognition. Reitboeck and Altmann [ReA84] note that size and rotation variations become translations in the complex log space when the object is in the center of the image. They propose applying a translation invariant transform to the resultant images to get templates suitable for matching with templates of known objects for recognition. The Fourier transform is dismissed as a candidate since there is no evidence that it is used anywhere in the visual system. The authors propose a C-transform, which is more consistent with operations that neurons can do. The Laplacian of the Gaussian operator is applied to the images before they are mapped into the complex log space. All the operations the authors do on the images can be performed with special hardware.

Massone, Sandini, and Tagliasco [MST85] study some of the characteristics of the CLM. They present a sampling algorithm based on the human visual system. A template matching approach is used for object recognition, where the center of gravity of a binary image of the object is used for the origin around which the object is mapped. Templates of the known object are created in this manner, as are maps of unknown objects.

Messner and Szu [MeS85] propose an architecture to perform an algorithm that simulates the CLM. A nonuniform sampling grid similar to the sampling algorithm of Massone, et al, is hardwired into a uniform grid. They show that

this mechanism duplicates the properties of the CLM.

Reeves, Prokop, Andrews, and Kuhl [RPA84] compare the performance of several shape recognition methods based on moments and Fourier descriptors. Each of these methods requires a large library (hundreds) of views to compare with the transformed image. This, along with the fact that these methods have no analogue in the human visual system, makes them less attractive then a complex log based approach.

Arsenault, Hsu, and Chalasinska-Macukow [AHC84] present a rotation and space invariant pattern recognition system based on matched filters. The various order circular harmonic components are used to differentiate between similar objects. However, there is no way to tell, a priori, what order component will be needed in a given situation. More work will have to be done before the performance of this method can be compared with the CLM.

Chaikin and Weiman[ChW79] have pointed out several advantages of the CLM in computer vision systems. In particular they show that the CLM space may allow iconic processing and can be implemented in hardware. Other advantages of the CLM for industrial applications are suggested in [ChW79, SWC81]. Thus, we see that efforts have been made by several researchers to use features of CLM for object recognition. Some efforts are in progress to have a hardware device that can transform an image from cartesian space to its CLM representation in real time [Ken83]. One very attractive feature of Complex Logarithmic Mapping is that it is conformal and, hence, unlike most other commonly used transformations in image processing, does not lose spatial connectivity of points.

An important property for information recovery from images is that a surface in the real world is mapped into a single region in an image. This surface coherence is used in recovering the structure of surfaces from the corresponding regions in images. The CLM preserves regions and hence also allows recovery of surface structure.

4. EGO-MOTION COMPLEX LOG MAPPING

To achieve the invariances, which are so important, the images must be obtained under certain constraints. The scale and rotation invariances are present only if the object is centered in the image, and the scale and rotation changes are with respect to the origin. In other cases, these invariances are not obtained. The projection invariance is only obtained by a camera translating along its optical axis. In this case the direction of the gaze and the direction of the motion are the same. This is a serious constraint. Indeed, in this case the FOE is (0,0) and hence the projection invariance really is the same as the scale invariance. If the observer motion is translational and is known, then the FOE is also known. The CLM is then taken so that all radii, r, and angles, θ, are in reference to this calculated FOE. This transformation is called Ego-Motion Complex Logarithmic Mapping (ECLM), since the mapping is performed with regard to the motion of the camera/observer. Let us consider this transformation for a point in the 3-D space.

When the observer moves in the direction of his gaze, the (X,Y,Z) coordinates of objects which are stationary relative to the observer, change only in the Z coordinate. With the perspective projection, the invariance resulting from the

ECLM gives only a horizontal displacement between images for corresponding points. This is very similar to the size invariance and can be compared to it and visualized with a little thought.

For a stationary point in the environment, with real world coordinates (X,Y,Z) relative to the observer at a time instant, the perspective projection, (x,y), of this point onto the image plane, is given by

$$x = \frac{X}{Z} \tag{11}$$

$$y = \frac{Y}{Z} \tag{12}$$

assuming that the projection plane is parallel to the XY plane at $Z=1$. For a translational motion along the direction of the gaze of the observer, the relationship between the distance, r, of the projection of the point from the FOE, and the distance, Z, of the point from the observer is

$$\frac{dr}{dZ} = \frac{d\sqrt{x^2+y^2}}{dZ} = -\frac{r}{Z} \tag{13}$$

By the chain rule

$$\frac{du}{dZ} = \frac{du}{dr} * \frac{dr}{dZ} \tag{14}$$

and from equation (9),

$$\frac{du}{dr} = \frac{1}{r} \tag{15}$$

Therefore, we have

$$\frac{du}{dZ} = -\frac{1}{Z} \tag{16}$$

Similarly, to find $\frac{dv}{dz}$,

$$\frac{d\theta}{dZ} = \frac{d(\tan^{-1}\frac{y}{x})}{dZ} = 0 \tag{17}$$

and

$$\frac{dv}{dZ} = \frac{dv}{d\theta} * \frac{d\theta}{dZ} = 0 \tag{18}$$

In equation (16) we see that the depth, Z, of a point can be determined from the horizontal displacement, du, in the ECLM for that point, and from the velocity, dZ, of the observer. Furthermore, the axial movement of the observer will result in *only* a horizontal change in the mapping of the image points since $dv/dZ = 0$. There will be no vertical movement of the mapped points and thus correspondence of points between the two stereo pictures will become easier. Note that this is similar to the epi-polar constraint used in the lateral stereo. Now, assuming that there is sufficient control of the camera to be able to determine the amount of its movement, both variables necessary to determine image depths are readily available. Thus, it is possible to recover depth, in principle, if the camera motion is along its optical axis.

What is more interesting is that the depth can be recovered using the above technique even if the camera motion is not along its optical axis. To see that the depth can be recovered for an arbitrary translatory motion of the camera, let us assume that the polar transform is taken with respect to the point (a,b) in the

image plane. Then

$$r = \sqrt{(x-a)^2 + (y-b)^2}$$
$$u = \log r = \log \left(\sqrt{(x-a)^2 + (y-b)^2} \right) \tag{19}$$

Now

$$\frac{du}{dZ} = \frac{d}{dZ} \log r = \frac{1}{r} \frac{dr}{dZ} \tag{20}$$

Let us substitute for x and y from equations (11) nad (12), and evaluate dr/dZ.

$$\begin{aligned} \frac{dr}{dZ} &= \frac{d\sqrt{(X/Z-a)^2+(Y/Z-b)^2}}{dZ} \\ &= \frac{1}{2\sqrt{(X/Z-a)^2+(Y/Z-b)^2}} \left[2(X/Z-a)\frac{ZdX/dZ-X}{Z^2} + 2(Y/Z-b)\frac{ZdY/dZ-Y}{Z^2} \right] \\ &= \frac{1}{\sqrt{(X/Z-a)^2)+(Y/Z-b)^2}} \cdot \frac{1}{Z} \cdot \left[(X/Z-a)(dX/dZ-X/Z)+(Y/Z-b)(dY/dZ-Y/Z) \right] \end{aligned} \tag{21}$$

Hence

$$\frac{du}{dZ} = \frac{1}{(X/Z-a)^2+(Y/Z-b)^2} \cdot \frac{1}{Z} \cdot \left[(X/Z-a)(dX/dZ-X/Z)+(Y/Z-b)(dY/dZ-Y/Z) \right]$$

Now suppose that we let (a,b) be the FOE, i.e.

$$a = \frac{dX}{dZ} \quad \text{and} \quad b = \frac{dY}{dZ}$$

Then, substituting for dX/dZ and dY/dZ

$$\begin{aligned} \frac{du}{dZ} &= \frac{1}{(X/Z-a)^2+(Y/Z-b)^2} \cdot \frac{1}{Z} \cdot \left[-(X/Z-a)^2 - (Y/Z-b)^2 \right] \\ &= -\frac{1}{Z} \end{aligned}$$

Now let us examine dv/dZ, when v is calculated with respect to any FOE (a,b).

$$v = \theta = \tan^{-1} \frac{(y-b)}{(x-a)}$$

$$\frac{dv}{dZ} = 1+\frac{1}{\frac{(y-b)^2}{(x-a)^2}} \quad \frac{d}{dZ} \quad \frac{(y-b)}{x-a}$$

Considering only the second factor of this equation, and substituting for x and y

$$\frac{d}{dZ} \quad \frac{(y-b)}{(x-a)} = \frac{d}{dZ} \quad \frac{(\frac{Y}{Z}-b)}{(\frac{X}{Z}-a)}$$

$$= \frac{(\frac{X}{Z}-a)(z\frac{dY}{dZ}-Y)/Z^2 - (\frac{Y}{Z}-b)(Z\frac{dX}{dZ}-X)/Z^2}{(\frac{X}{Z}-a)^2}$$

$$= \frac{(\frac{X}{Z}-a)(\frac{dY}{dZ}-a)(\frac{dY}{dZ}-\frac{Y}{Z}) - (b-\frac{Y}{Z})(\frac{X}{Z}-\frac{dX}{dZ})}{Z(\frac{X}{Z}-a)^2}$$

Remembering that $dX/dZ = a$ and $dY/dZ = b$

$$\frac{d}{dZ} \quad \frac{(y-b)}{(x-a)} = \frac{(\frac{X}{Z}-a)(b-\frac{Y}{Z})-(\frac{X}{Z}-a)(b-\frac{Y}{Z})}{Z(\frac{X}{Z}-a)^2}$$

$$= 0$$

Therefore,

$$\frac{dv}{dZ} = 0$$

Note that when the polar coordinates are obtained with respect to the FOE, then the displacement in the u direction depends only on the Z coordinate of the point. For other values of (a,b) the above property will not be true.

Another interesting feature of this stereo approach is that, if required, we can obtain many frames for solving ambiguities that cannot be resolved based only on two frames. Moravec [Mor81] developed a technique for interpolating over nine frames which he used with common stereo. This technique may be even more applicable to motion stereo, because the series of frames can be naturally extended each time the observer moves. The frame sequence can be constantly updated by merely pushing back the current series by one time instant and adding a new frame to the front of the sequence.

5. THE MAPPING

Mathematically, each point in the image space corresponds to exactly one point in the space transformed through CLM. However, in computer vision systems where only a finite amount of memory space and computation time is available, an image can only be stored as a finite number of pixels and only a finite number of intensities are representable. This quantization of the image leads to ambiguity in the mapping, since an image pixel can map to a range of pixels in the transformed space. For example, for a pixel in the first quadrant with coordinates (x,y) at the lower left corner, the u coordinate in the transformed space will range from $\log\sqrt{x^2+y^2}$ to $\log\sqrt{(x+1)^2+(y+1)^2}$ and the v-coordinate will range from $\tan^{-1}\frac{y}{x+1}$ to $\tan^{-1}\frac{y+1}{x}$. These ranges can be quite wide for points close to the origin, or practically negligible for points far from the origin.

We considered several different interpolations of the image pixels to produce the CLM. One very simple method we examined involved merely computing the

range of each image pixel in the mapped space and setting each map pixel in this range to the corresponding image intensity. This procedure resulted in a very broken, choppy mapping. We also tried working inversely from the mapped space. The image point corresponding to each pixel in the CLM space was determined, and then various interpolations of the intensities of the image pixels around this point were tested. This method of inverse mapping resulted in a much smoother CLM.

The combination of image pixels we found that resulted in the most continuous and pleasing mapping was surprisingly simple. It involved merely adding the intensities of the portions of the image covered by a three pixel square centered at the point found from the inverse mapping. An indication of how this worked is shown in Figure 5. The results of such a mapping are shown in Figure 6. This method may be refined by assigning weights to the various areas of the square. In the future, if the CLM indeed becomes useful, it can easily be implemented in hardware.

6. AVAILABLE RESOLUTION AND THE BLINDSPOT

6.1. Available Resolution

The degree of resolution available for the determination of depth in motion stereo depends directly on the amount of resolution used for the images and the mappings. The resolution determines how large the pieces of information are that must be squeezed into each pixel. As the resolution increases, the information can be kept more exactly. With infinite resolution we could determine

depths exactly. Unfortunately infinite resolution requires infinite memory and computation, which we cannot provide.

To distribute the entire image onto a mapping. The factor which we use to distribute the u coordinate over the mapping will determine what depths can be recovered, since it will tell what increments of du are possible. The u scale factor was found in the following way: The resolution for both the images and the mappings was the same and shall be called ρ. We produced images so that the FOE was at the center of the image, therefore the x and y coordinates in an image could range from $-\frac{\rho}{2}$ to $+\frac{\rho}{2}$. The maximum possible value for u, then, is

$$u_{max} = \log\sqrt{(\frac{\rho}{2})^2 + (\frac{\rho}{2})^2} = \log(\frac{\rho}{2}\sqrt{2})$$

Thus, the u-axis of the CLM will range from 0 to u_{max} over ρ pixels, and every u in the CLM determined from the image pixel locations is multiplied by $\frac{\rho}{u_{max}}$ so that they are distributed over the entire CLM. We do not consider that u may have negative values since all mapping is done in reverse from the map to the image and there are only four pixels directly adjacent to the FOE which can have distance less than one and result in a negative u. These pixels will be interpolated into the map by their neighbors.

The smallest du that can be detected, for the camera displacement of one pixel, is $\frac{u_{max}}{\rho}$ and, therefore, the greatest distance that is recoverable is $\frac{\rho}{u_{max}}$ No u displacement between mappings for corresponding points in the CLM

indicates that the depth of that point is too much greater than this to be determined. Other dus can be found in integer multiples of $\frac{u_{max}}{\rho}$, i.e.,

$$du = n \times \frac{u_{max}}{\rho} \qquad \text{for } n = 1,2,3,...$$

and the depths which can be determined will be

$$Z = \frac{1}{n} \times \frac{\rho}{u_{max}} \qquad \text{for } n = 1,2,3,...$$

See Table 1 for depths recoverable with different resolutions.

It is interesting to note that as n increases, the depth Z decreases and there is increasing accuracy available. In other words, the depth of points that are closer to the observer can be more precisely determined than points that are further away. This is similar to what we observe in our own vision, that we can perceive depth best for objects which are close to us.

The numbers in Table 1 tell us some of the limitations of axial motion stereo. At a very low resolution of 128×128, the greatest distance that can be recovered is about 28 units (where 1 unit is the distance dZ traveled between images). All points that are much further away than this will have no u displacement. Also, for small u displacements of very few pixels in the map, there are large intervals between the depths that can be recovered. Specifically, the second furthest depth that can be determined is exactly half of the furthest depth! This is quite a large distance, especially relatively.

6.2. The Blindspot

The fact that closer points can be more precisely gauged makes good practical sense. The importance of five feet difference in depth between 1 and 6 feet is much greater than the same difference between 25 and 30 feet. For objects that are close to us, usually a decision must be made about how to treat or avoid that object before a similar decision is made for more distant objects. It would be advantageous, however, to have more precision available.

As was shown above, the available resolution depends on u_{max}. In fact, the resolution depends on the range of u values for the mapping. In the above discussion, the lower bound, u_{min}, was assumed to be 0. The resolution at the periphery can be improved by sacrificing some vision along the line of sight. We can introduce a blind spot at the FOE in the image, and thus increase u_{min}. Depending on the size of the blindspot, the range $[u_{min}, u_{max}]$ will change and hence the resolution will also change. The resolution at the periphery will improve with increasing size of the blindspot. This can be seen from Figure 7.

7. EXPERIMENTS

7.1. Synthetic Images

We studied the precision of the recovered depth as a function of the location of the object in the visual field, and also as a function of the camera movement by generating synthetic images and simulating the movement of a single "block" from a centered location toward the right. A series of 3 images taken from dif-

ferent distances is simulated for each position of the block as it moves. Distances of the object from the original "camera position" are calculated using the first and second images, the second and third images, and the first and third images. The percent error in the distance calculation for each set of images is plotted for each position of the block. The plots in Figure 8 show results of our study. The percent error is a function of the distance of the camera from the object, the distance the camera moves, and the size of the object in the image. Data for two sizes of objects, three camera distances, and two camera movement values are included. For all this data, the location of the block after the camera "moves" is calculated using perspective projection and then "digitized". For comparison, a set of data is included for which the location of the block in the second image is not "digitized", that is, the double precision values are mapped rather than their closest integer equivalent. Note that the error is constant in this case. This shows that the fluctuation of the error is the result of digitization error.

The effect of the initial distance of the camera on the percent error is studied directly by having the block remain at the same location relative to the center of the image. This was done with the block at different locations with respect to the center. Two sets of data with these constraints were collected. In one, the size of the block in the image remained constant; in the other, the size of the object in the space remained constant. The first case would be equivalent to moving the camera back and using a larger object. This was done to study the effect of the size of the object in the image on distance determination. In the other set, the size of the object in the image is adjusted, based on perspective

projection. In each series, data was collected until there was no change in the location of the block in the image after the camera "moved" or until 22 sets of data had been obtained. The results of the experiments are shown in Figure 9.

There does not seem to be a simple relationship between the error and any of the variables studied. For example, in some cases, a larger camera movement resulted in less error while in others it resulted in more. Digitization errors appear to dominate and hence working with higher resolution images will be a big help. It may be the case that the error is minimized when the map that is used "matches" the object and its location, that is, when the "mask" of this algorithm and the block fulfill the constraints of Massone, et al. It is an interesting idea that needs more research.

7.2. Depths Using Real Images

To study the efficacy of the proposed approach for real scenes, we performed several experiments in our laboratory. We mounted a camera on a PUMA robot. This set up allows us to move the camera in a desired direction by a desired amount.

The objects used were wooden blocks with dimensions less than 6 inches. We were forced to place objects within a small depth of each other due to the limited depth of field of the camera. By placing objects far away, we could do some more experiments, but our laboratory set-up did not allow us to perform these experiments. Moreover, we were more interested in studying the efficacy of the proposed approach, which could be done by considering a limited set of

scenes. In the first image the lower right corner of the farthest block was placed in the center of the image. A paper triangle was mounted on a block and positioned so an acute angle lined up with the block corner. The position of the triangle was marked and experiments were done to guarantee the reproducability of its position. The image was obtained with the triangle removed, then the camera moved 6 inches forward. The triangle was replaced and the camera location adjusted so the acute angle and the corner of the block lined up. This process was repeated until 5 images had been obtained. The 512 x 512 images were shrunk to 128 x 128. Three frames of the sequence obtained are shown in Figure 10.

Corners were found and the coordinates in complex log space of these points were calculated. The corner detector used simple masks to find corners, so the corner images produced were noisy. No effort was made to remedy this since it served to make the problem more challenging. The corners obtained using our algorithms are shown in Figure 11. Note that the location of the corners is noisy and many false corners are detected.

The corners were mapped tp the ECLM space. Figure 12 shows the corners of Figure 11 mapped in a composite frame in the ECLM space. The correspondence between corners was established in the ECLM space. For each pair of images, the coordinates in the ECLM space were calculated for all the "corners" and saved in a list. The algorithm knows the number of objects and the corners for each object. The number of corners detected in different images varies significantly, however. Theoretically, two matching points should have the same θ

value (v coordinate). Due to digitization error, blurring in the images from shrinking, and errors in corner detection, matching points do not have the same θ value. So a threshold for dv (the difference between the θ values) is used. Some heuristics were used to establish correspondence for points that are within the selected threshold.

The algorithm was run on every pair of images using three different thresholds for dv: 0.01, 0.02, and 0.05. A match is considered correct if the points belong to the same object. By using two simple rules, all errors are detected by the system. The rules are: a camera cannot see behind itself and if you pass an object, it won't be in the second image. This eliminates all matches with a distance less than the amount the camera moved. Figure 13 shows some examples of matches and the results of applying the rules to the matches.

The effect on the average depth determination with a tighter bound on allowable depths for individual pairs was also studied. An arbitrary value larger than the distance the camera moved was chosen as the lower bound. An arbitrary value of 100 was chosen as the upper bound.

The depth of an object was obtained by averaging the depth obtained for its corners. The data obtained using these real images reflects the noise inherent in the system, as demonstrated by the work with the synthetic images. Several trends are apparent, though they do not hold universally.

As the threshold increased for dv, the number of pairs that "matched" increases as does the error rate. However, since the system is so good at finding

the erroneous matches, the extra information from the larger number of points makes the distance determinations more accurate, overall.

The calculated depths tended to be lower than the real depths. We did not use the focal length of the camera in our depth computation, since for the camera we used, the focal length is not known. Without the focal length, the depth values should indicate the relative depth values, rather than the absolute depth. Depth determination from images where the camera was closer to the objects was more accurate. Larger camera movement gave better results when the camera was far from the objects, but not from closer positions. Some results are shown in Table 2. We ran the experiments for several other frame pairs. In some cases the results indicated wrong depth order for the objects.

Note that in the above table for each object two distances are given: the one on the left is generated using the two simple rules, the one on the right uses the arbitrary upper and lower bounds in addition to the two rules.

Using the upper and lower bound usually made no difference in the depth determination. However, this heuristic corrected several errors that the simple rules did not catch. In some cases the threshold for dv does not affect the depth determination. This is because the new points found with a higher threshold generated approximately the same depth value as the others. This gives an indication of the stability of the system.

8. CONCLUSION

Most research in stereo has considered lateral placement of cameras. The lateral placement of cameras is certainly a geometrically efficient placement for recovering the depth information. An arbitrary motion of the camera allows a rich set of camera geometries for recovering depth information. Structure-from-Motion [Ull79] has attracted significant interest recently. The research in structure from motion has shown that in realistic cases the approaches to structure from motion are too sensitive to be of much use. Two major problems in structure from motion are the correspondence problem and the sensitivity of the methods to solve equations to get the depth values. Another approach to determination of depth using motion is to use optical flow. The approaches based on optical flow have the problem of computing reliable optical flow that will allow computation of its derivatives for recovering the depth. Looming combines features of token based approaches for structure from motion and approaches using optical flow. In fact, looming has been known to be a depth recovery technique used by animals in many different situations. In this paper we have shown that looming can be efficiently implemented using the known camera motion. In the proposed approach, the most complex problem involved in structure-from-motion, that is the correspondence problem, is simplified appreciably. Similarly the problem of computation of optical flow is also avoided by exploiting the ego-motion information.

The study of the synthetic images showed that factors inherent in the theory and the digital nature of the data make it difficult to recover accurate depths

using this method. However, relative depths with bounds on the quantitative depth are certainly possible if information from several frames is used.

The results also have implications for determining correspondence of points and objects. By using information about which points are matched across frames, more accurate depth values can be obtained by eliminating bad matches that the "in the same object" criterion doesn't catch. This in turn can be used to refine knowledge about which points match and which points belong to the same object.

Our experiments with synthetic and real scenes indicate the efficacy of this approach for recovering depth of stationary objects in those situations where the observer is translating. The resolution of the depth recoverable using this approach is poorer compared to the regular lateral stereo for the same displacement of the camera. This limitatation will be more than offset when we consider the fact that in most cases the depth will be obtained as a by-product of other visual processes. A mobile robot will have to segment a scene as the first step in its understanding of the layout around it. As was shown in [Jai84], ECLM will be helpful in the segmentation of the scene. Since the depth recovery also requires the same mapping with little extra efforts, the depth of stationary objects may be recovered as a by-product of segmentation. Another major advantage of this method of depth recovery will be that in most applications, one may combine depth information obtained from several images, rather than from just two images.

REFERENCES

[AHC84] Arsenault, H.H. and Y.N. Hsu and K. Chalasinska-Macukow, "Rotation-Invariant Pattern Recognition", *Optical Engineering,* Vol. 23, pp. 705-709, Nov.-Dec. 1984.

[BaF82] Barnard, S.T. and M. A. Fischler, "Computational Stereo", *Computing Surveys,* Vol. 14, pp. 553-572, 1982.

[BeL84] Berthod, M. and P. Long, "Graph Matching by Parallel Optimization Methods: An Application to Stereo Vision", Seventh International Conference on Pattern Recognition, Vol. 2, pp. 841-843, Montreal, July 1984.

[BGT79] Braccini, C., G. Gamberdella, and V. Tagliasco, "A model of the early stages of human visual system," *Biological Cybernetics,* 44, 1982, pp.47-88

[BrH83] Bruss, A.R. and B.K.P. Horn, "Passive Navigation", *Computer Vision, Graphics, and Image Processing,* Vol.21, 1983.

[Bur84] Burr, D.J., "A Fast Filtering Operator for Robot Stereo Vision", Seventh International Conference on Pattern Recognition, Vol. 1, pp. 669-672, Montreal, July 1984.

[Cav78] Cavanaugh, P., "Size and position invariance in the visual system," *Perception,* vol. 7, pp.167-177, 1978.

[Cav81] Cavanaugh, P., "Size invariance: reply to Schwartz," *Perception,* col. 10, pp.469-474, 1981.

[ChW79] Chaikin, G. and C. Weiman, "Log spiral grids in computer pattern recognition", *Computer Graphics and Pattern Recognition,* vol.4, pp.197-226, 1979.

[Clo80] Clocksin, W.F., "Perception of surface slant and edge labels from optical flow: A computational approach," *Perception,* vol. 9, 1980, pp.253-269.

[Gib79] Gibson, J.J., *The ecological approach to visual perception,* Houghton Mifflen, Boston, 1979.

[Gri81] Grimson, W.E.L., *>From Images to Surfaces A Computational Study of the Human Early Visual System,* The MIT Press, Cambridge, Massachusetts, 1981.

[Gri83] Grimson, W.E.L., "Binocular Shading and Visual Surface Reconstruction", *Computer Vision, Graphics, and Image Processing,* Vol. 28, pp. 19-43, Oct. 1983.

[GYH84] Gu, W.K. and J.Y. Yang and T.S. Huang, "Matching Perspective Views of a 3-D Object Using Circuits", Seventh International Conference on Pattern Recognition, Vol. 1, pp. 441-443, Montreal, July 1984.

[HoF84] Homma, K. and K.S. Fu, "A Stereo Vision Method Based on Region Segmentation", IEEE Computer Society Workshop on Visual Languages, Hiroshima, pp. 14-19, Dec. 1984.

[Ike84] Ikeuchi, K., "Reconstructing a Depth Map From Intensity Maps," Seventh International Conference on Pattern Recognition, Vol. 2, pp. 736-738, Montreal, July 1984.

[IMO84] Itoh, H. and A. Miyauchi and S. Ozawa, "Distance Measuring Method Using Only Simple Vision Constrained for Moving Robots", Seventh International Conference on Pattern Recognition, Vol. 1, pp. 192-195, Montreal, July 1984.

[Jai83] Jain, R., "Complex Logarithmic Mapping and the Focus of Expansion", SIGGRAPH/SIGART Workshop on MOTION: Representation and Perception, Toronto, April 1983.

[Jai84] Jain, R. "Segmentation of frame sequences obtained by a moving observer," *IEEE Trans. PAMI.* pp. 624-629, Sept. 1984.

[JaO85] Jain, R. and N. O'Brien, "Ego-Motion Complex Logarithmic Mapping" *SPIE,* Npv. 1984.

[Ken83] Kent, E., Personal Communication.

[Lee80] Lee, D.N., "The optic flow field: The foundation of vision," *Phil. Trans. Royal Society of London,* vol. B290, 1980, pp. 169-179.

[LuK79] Luh, J.Y.S. and J.A. Klaasen, "Real-time 3-D Vision by Off-shelf System with Multi-cameras for Robotic Collision Avoidance", First

International Conference on Computers and Applications, Beijing, pp.887-894, June 1979.

[Mar82] Marr, D., *Vision,* Freeman, 1982.

[MaT84] Mattias, L.H. and C. Thorpe, "Experience with Visual Robot Navigation", 84 Oceans Conference Record, Vol. 2, Washington, D.C., pp.594-597, Sept. 1984.

[MeS85] Messner, R.A. and H.H. Szu, "An Image Processing Architecture for Real Time Generation of Scale and Rotation Invariant Patterns", *Computer Vision, Graphics, and Image Processing,* Vol. 31, pp. 50-66, 1985.

[Mor81] Moravec, H.P., *Robot Rover Visual Navigation* UMI research Press, Ann Arbor, 1981.

[MST85] Massone, L. and G. Sandini and V. Tagliasco, ""Form-Invariant" Topological Mapping Strategy for 2D Shape Recognition", *Computer Vision, Graphics, and Image Processing,* Vol. 30, pp. 169-188, 1985.

[Nev76] Nevatia, R., "Depth Measurement by Motion Stereo" *Computer Graphics and Image Processing,* Vol.5, pp. 203-214, 1976.

[Nis84] Nishihara, H.K., "Practical Real-time Imaging Stereo Matcher", *Optical Engineering,* Vol. 23, pp. 536-545, Sept.-Oct. 1984.

[ObJ84] O'Brien, N. and R. Jain, "Axial Motion Stereo", Proc. of Workshop on Computer Vision, pp. 88-92, Anapolis, Maryland, April 1984.

[Pra80] Prazdny, K., "Egomotion and relative depth map from optical flow," *Biological Cybernetics,* vol. 36, 1980, pp. 87-102.

[ReA84] Reitboeck, H.J. and J. Altmann, "A Model for Size- and Rotation-Invariant Pattern Processing in the Visual System", *Biological Cybernetics,* Vol. 51, pp. 113-121, 1984.

[RPA84] Reeves, A.P. and R.J. Prokop and S.E. Andrews and F.P. Kuhl, "Three Dimensional Shape Analysis Using Moments and Fourier Descriptors", *Seventh International Conference on Pattern Recognition,* Vol. 1, pp. 447-450, Montreal, July 1984.

[SaK83] Safranek, R.J. and A.C. Kak, "Stereoscopic Depth Perception for Robot Vision: Algorithms and Architectures", Proc. IEEE International Conference on Computer Design: VLSI in Computers, Port Chester, N.Y., pp.76-79, Oct. 1983.

[SaT80] Sandini, G and V. Tagliasco, "An anthromomorphic retin-like structure for scene analysis," *Computer Graphics and Image Processing,* vol.14, pp.365-372, 1980.

[Sch77] Schwartz, E. L., "The development of specific visual connections in the monkey and goldfish: Outline of a geometric theory of receptotopic structure", *J. Theoretical Biology,* vol 69, pp.655-683.

[Sch80] Schwartz, E. L., "Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to coding," *Vision Research,* 20, 1980, pp.645-669.

[Sch81] Schwartz, E. L., "Cortical anatomy, size invariance, and spatial frequency analysis," *Perception,* vol.10, pp.455-468, 1981.

[Sch82] Schwartz, E.L., "Columnar architecture and computational anatomy in primate visual cortex: Segmentation and feature extraction via spatial frequency coded difference mapping," *Biological Cybernetics,* vol. 42, pp.157-168, 1982.

[SmB84] Smitley, D.L. and R. Bajcsy, "Stereo Processing of Aerial, Urban Images", Seventh International Conference on Pattern Recognition, Vol. 1, pp. 433-435, Montreal, July 1984.

[StE84] Stockman, G. and J.C. Esteva, "Use of Geometrical Constraints and Clustering to Determine 3D Pose", Seventh International Conference on Pattern Recognition, Vol. 2, pp. 742-744, Montreal, July 1984.

[SWC81] Schenker, P.S., K.M. Wong, and E.G. Cande "Fast adaptive algorithms for low-level scene analysis: Application of polar exponential grid (PEG) representation to high-speed, scale-and-rotation invariant target segmentation", *Proc. SPIE,* Vol. 281, Techniques and Applications of Image Understanding, pp.47-57, 1981.

[Ull79] Ullman, S., *The interpretation of visual motion,* MIT Press, Cambridge, Mass, 1979.

[WWB83] Wu, C.K. and D.Q. Wang and R.K. Bajcsy, "Acquiring 3-D Spatial Data of a Real Object", *Computer Vision, Graphics, and Image Processing*, Vol. 28, pp. 126-133, Oct. 1983.

[ZaC85] Zacharias, G.L. and A.K. Caglayan, "A Visual Cueing Model for Terrain-Following Applications", *Journal of Guidance, Control, and Dynamics*, Vol. 8, pp. 201-207, Mar.-Apr. 1985.

[ZCS83] Zacharias, G.L. and A.K. Caglayan and J.B. Sinacori, "A Model for Visual Flow-Field Cueing and Self-Motion Estimation", Proc. 1983 American Control Conference, pp. 1326-1330, San Francisco, June 1983.

Table 1

Recoverable Depths at Various Resolutions

ρ	1	2	3	4	5	6
64	16.8	8.4	5.6	4.2	3.4	2.8
128	28.4	14.2	9.5	7.1	5.7	4.7
256	49.2	24.6	16.4	12.3	9.8	8.2
512	86.9	43.5	29.0	21.7	17.4	14.5
1024	155.5	77.8	51.8	38.9	31.1	25.9

Table 2

Depths of Objects

Threshold	frame-pair	obj 1 (65)	obj 2 (83)	obj 3 (75)
0.01	1-3	50 50	74 74	57 57
0.01	2-4	58 58	79 79	77 77
0.01	1-4	53 53	72 72	56 56
0.01	1-5	52 52	71 71	51 51
0.02	1-3	50 50	67 67	54 54
0.02	2-4	59 59	78 78	71 71
0.02	1-4	49 49	72 72	56 56
0.02	1-5	50 50	71 71	51 51
0.05	1-3	50 50	67 67	61 61
0.05	2-4	208 51	72 72	64 64
0.05	1-4	82 49	72 72	56 56
0.05	1-5	60 50	71 71	52 52

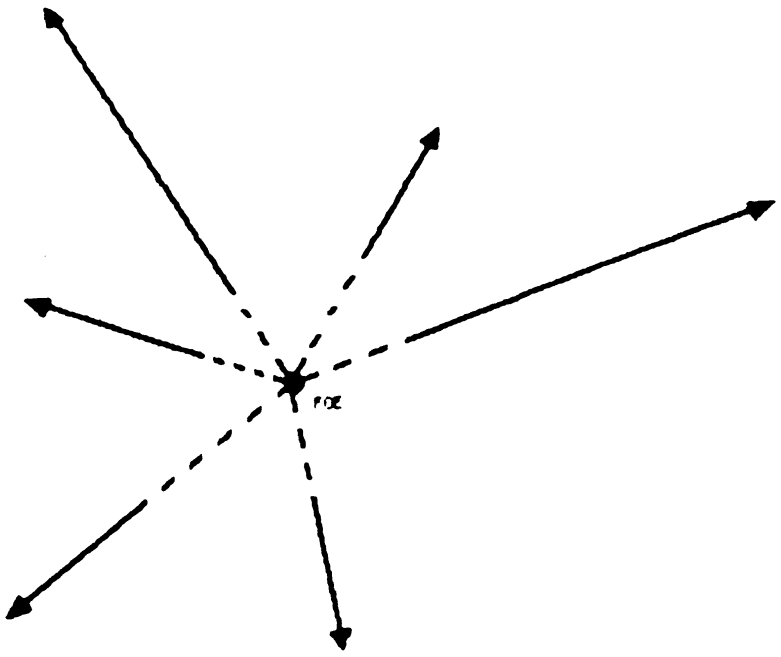

Figure 1 The optical flow field for a moving observer in a stationary environment. The vectors intersect at the *focus of expansion*.

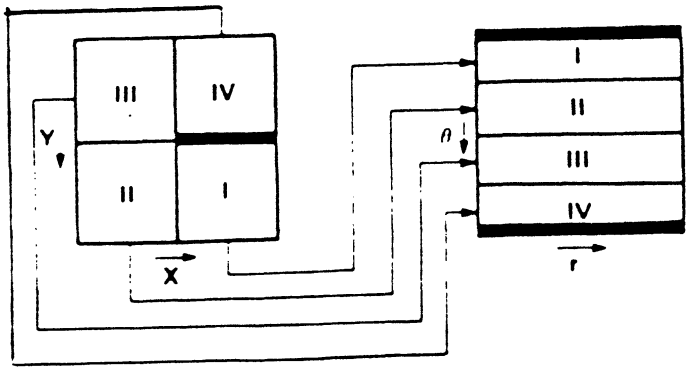

Figure 2 An image $I(x,y)$ is converted to another image $I(r,\theta)$ using the EMP transform.

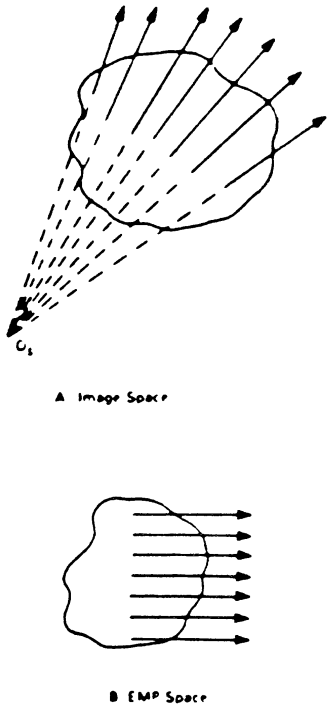

Figure 3 The assorted directions of the velocity vectors for a stationary object are transformed to one direction in the transformed image.

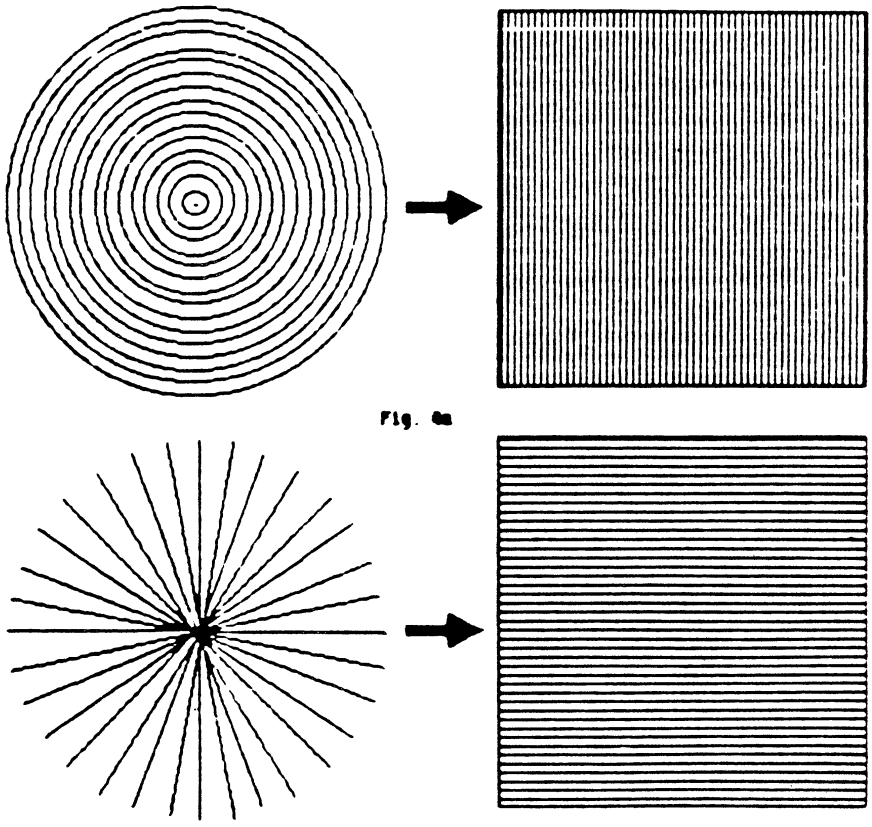

Figure 4 The CLM results in the transformation of certain regular patterns in the *z-plane* into another regular pattern in the *w-plane*.

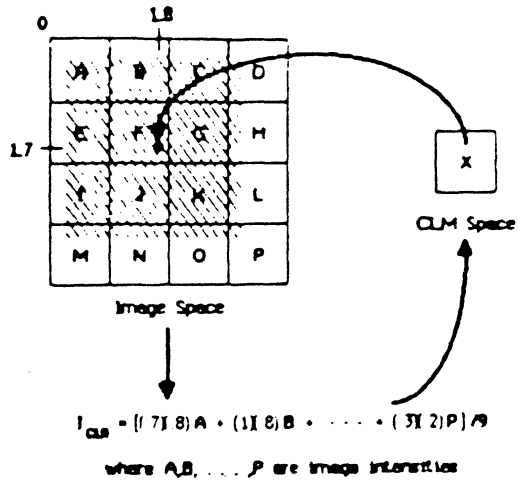

Figure 5 The interpolation scheme used for the ECLM mapping uses the area of pixels that contribute to the intensity at the point in the ECLM space.

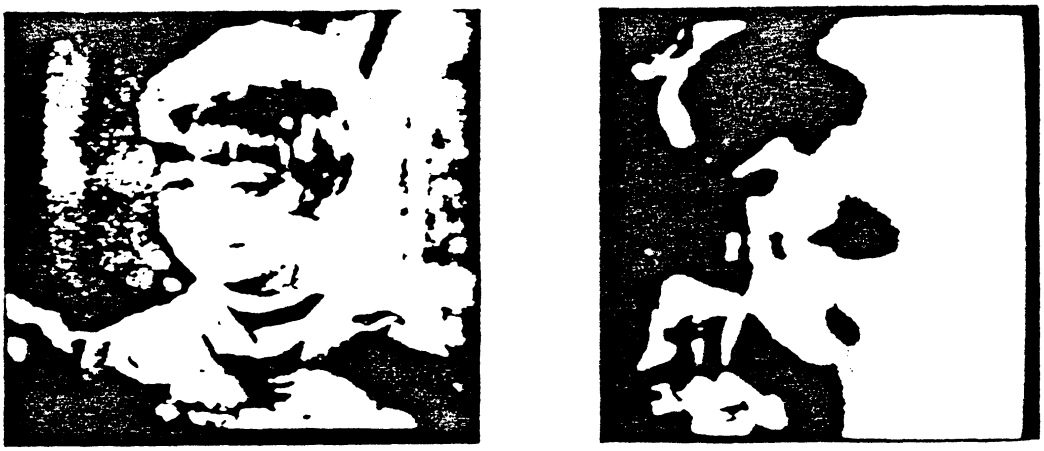

Figure 6 An image and its ECLM are shown in this figure.

Figure 7 The increase in the size of the blind spot results in an increase in the details of the peripheral areas.

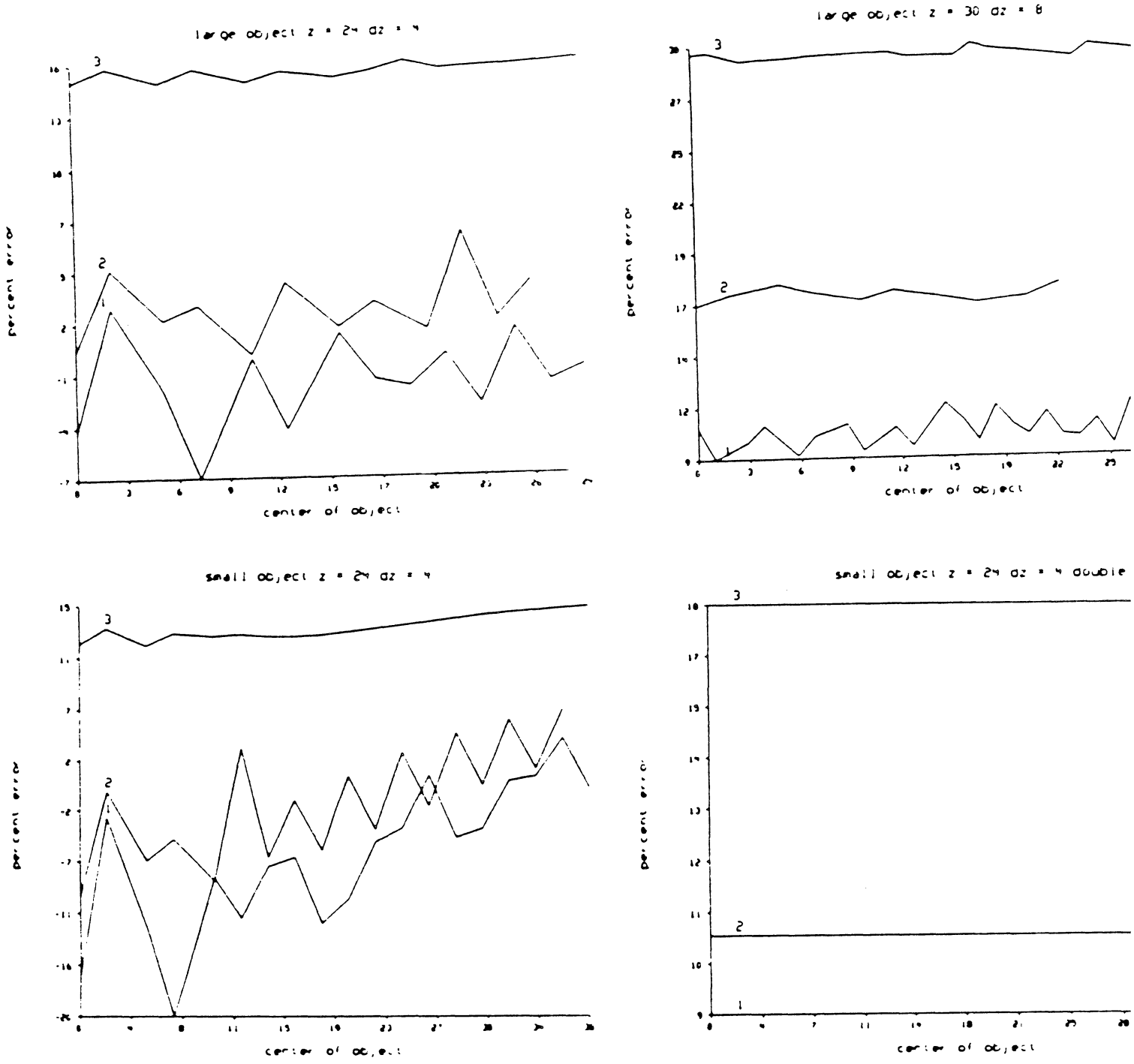

Figure 8 Plots showing the effect of eccentricity on the depth determination. Data for line 1 is from the first and second images, for line 2 from the second and third images, and for line 3 from the first and third images

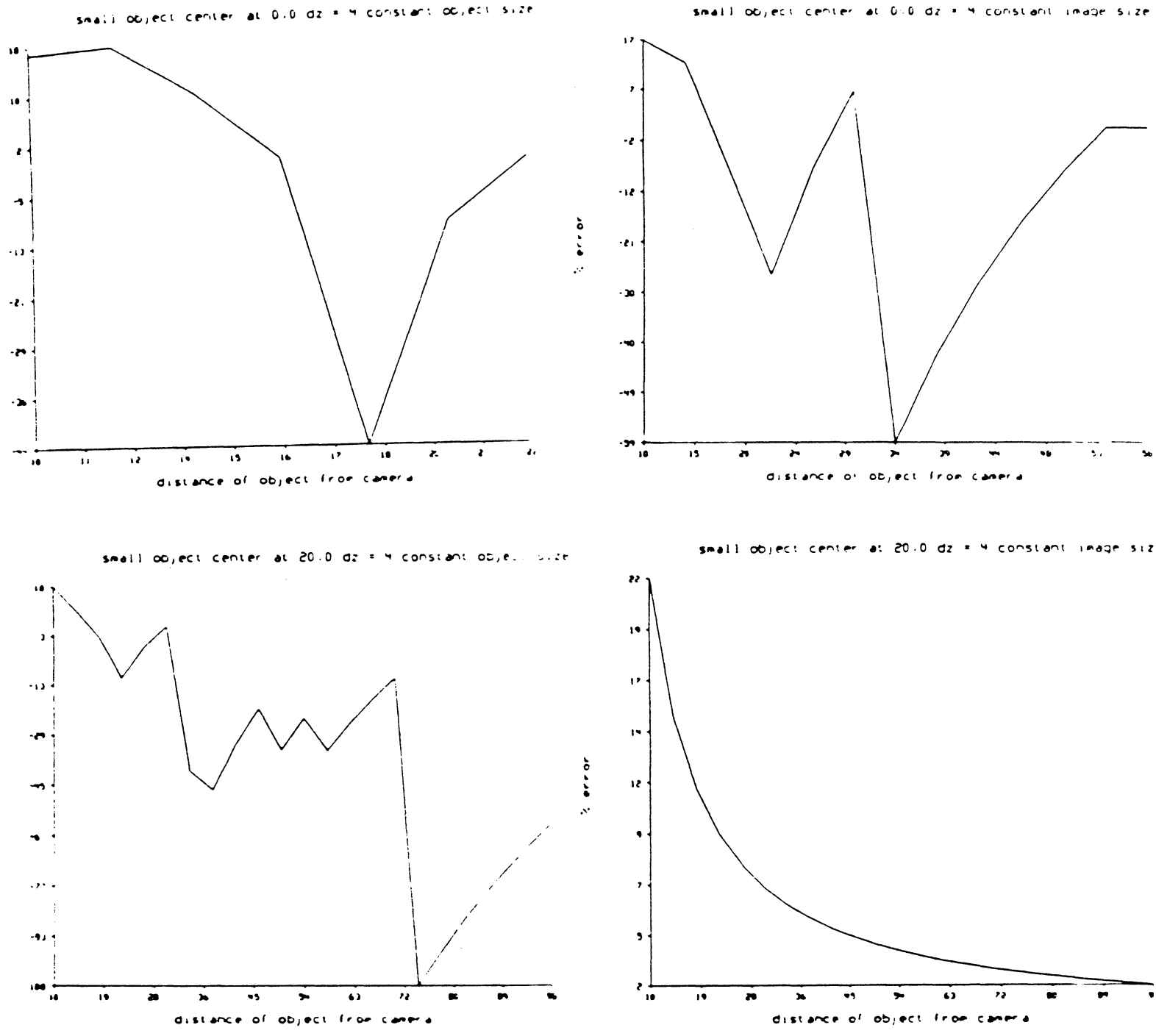

Figure 9 Plots showing the effect of the initial distance of the camera on depth determination. The camera moves the same amount at each distance.

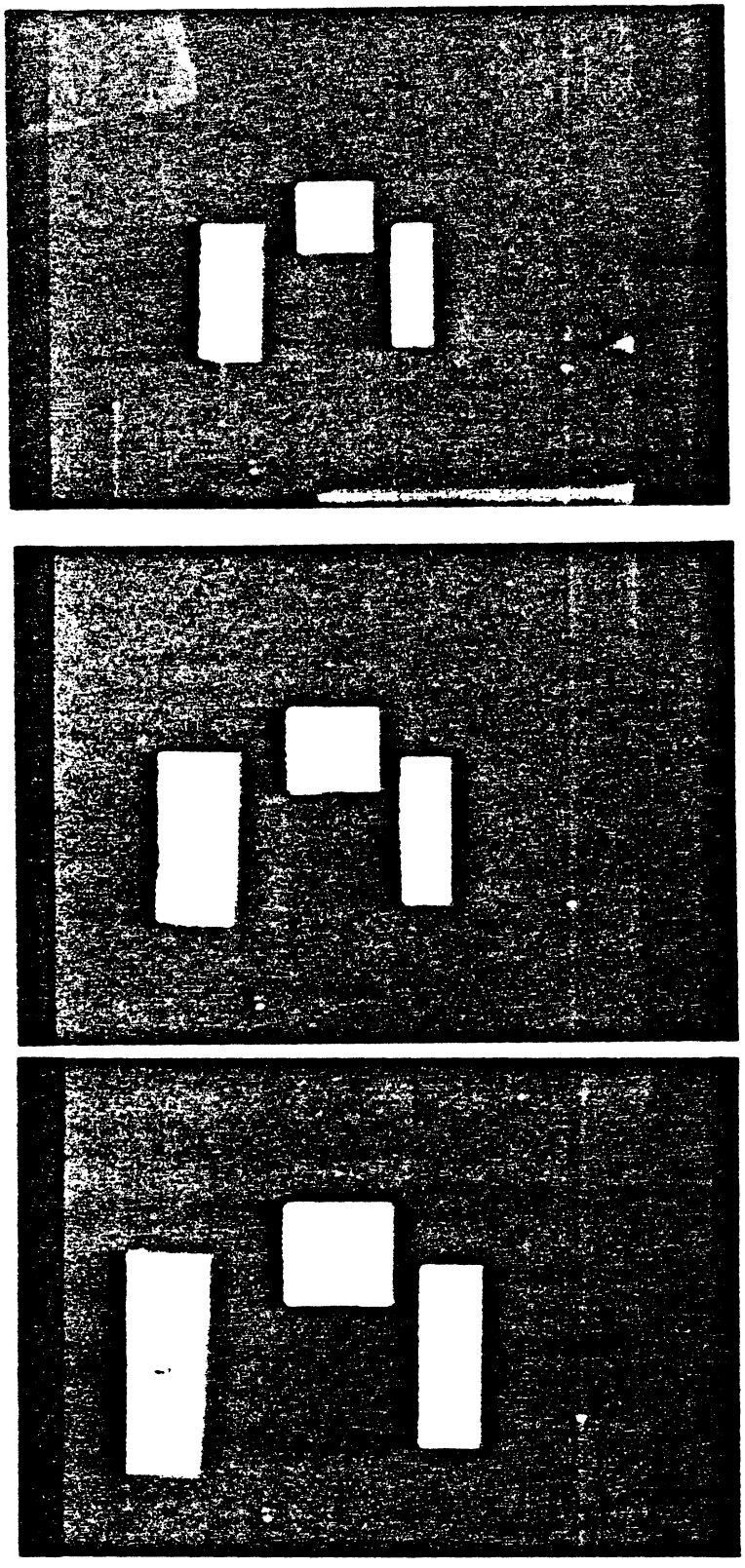

Figure 10 Three frames of the laboratory sequence used in our experiments.

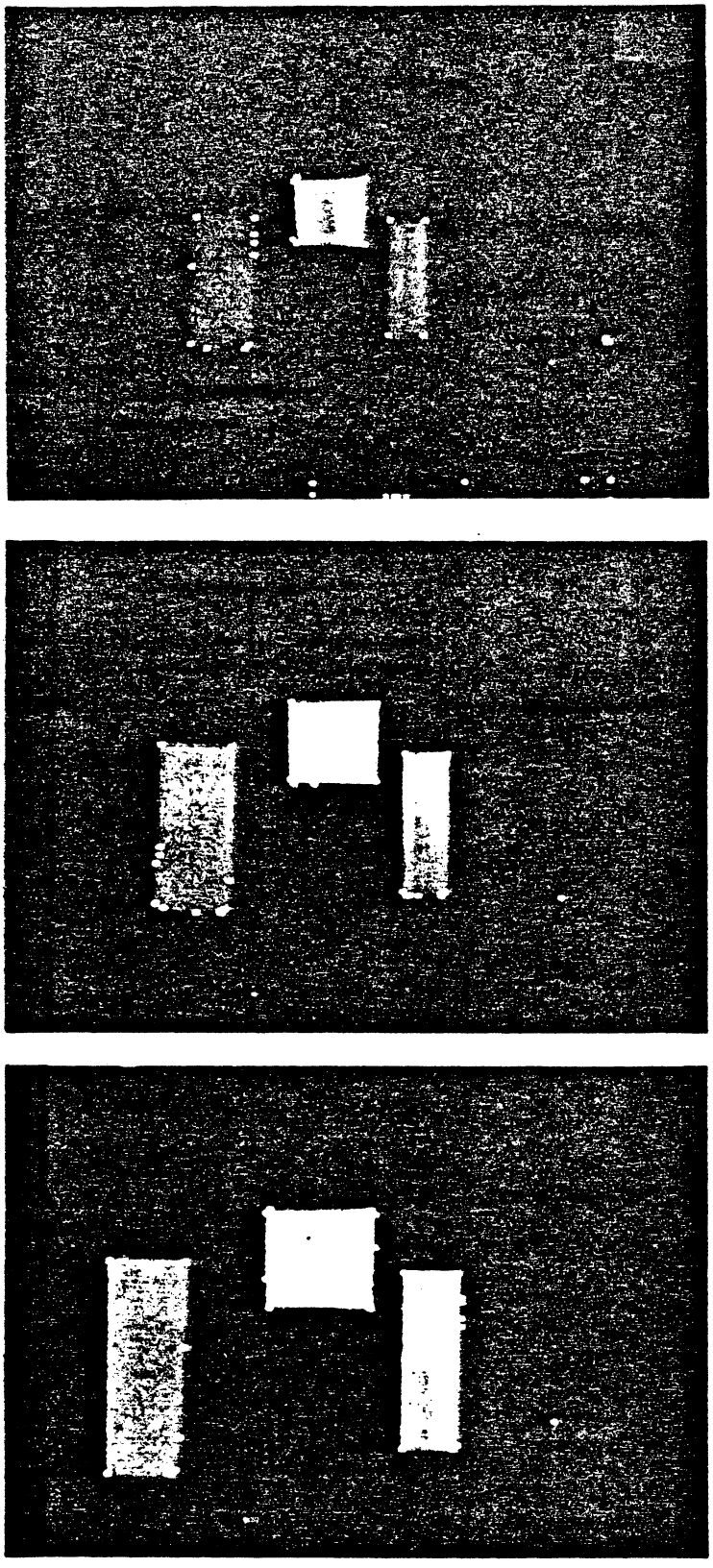

Figure 11 Corners detected in the frames. Note the poor quality of the corners.

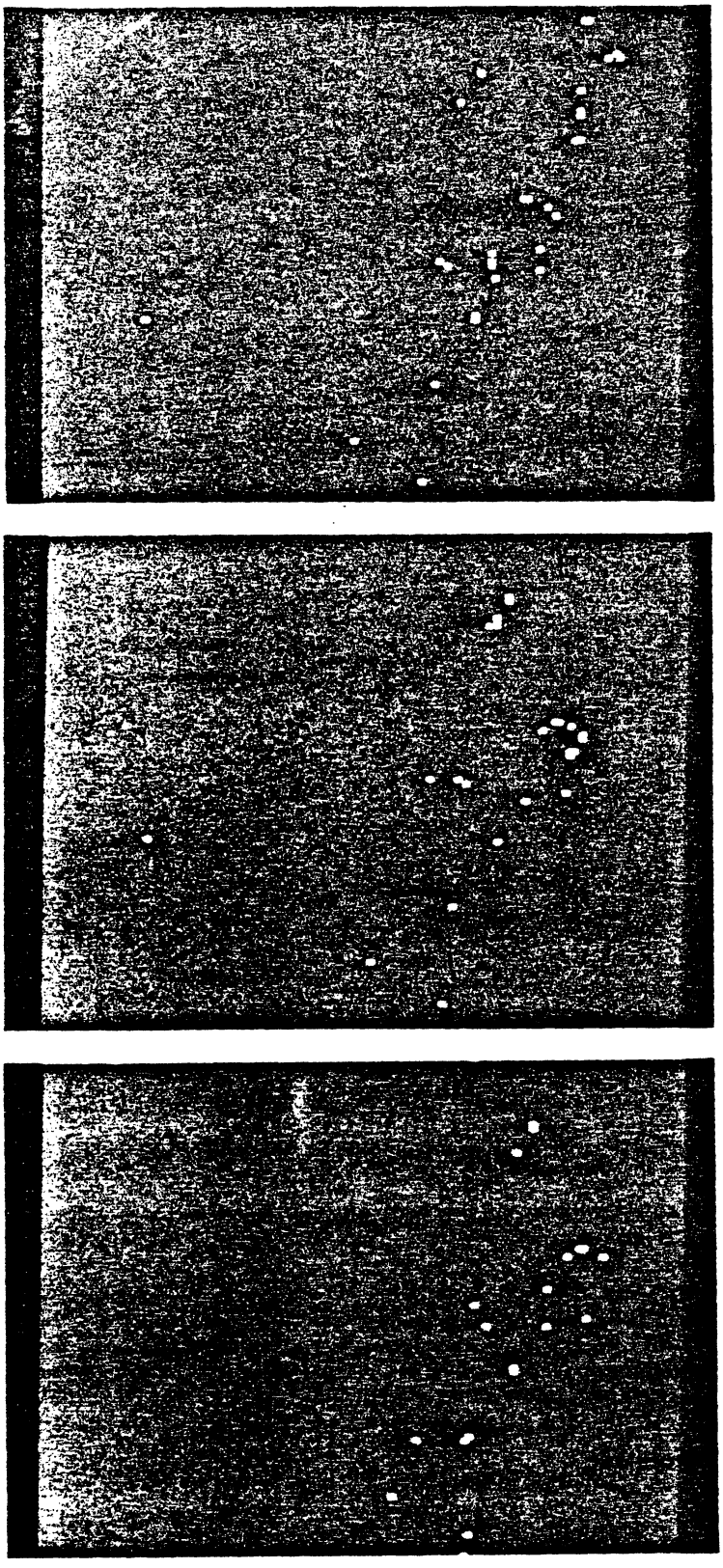

Figure 12 The corners of figure 11 in the ECLM space.

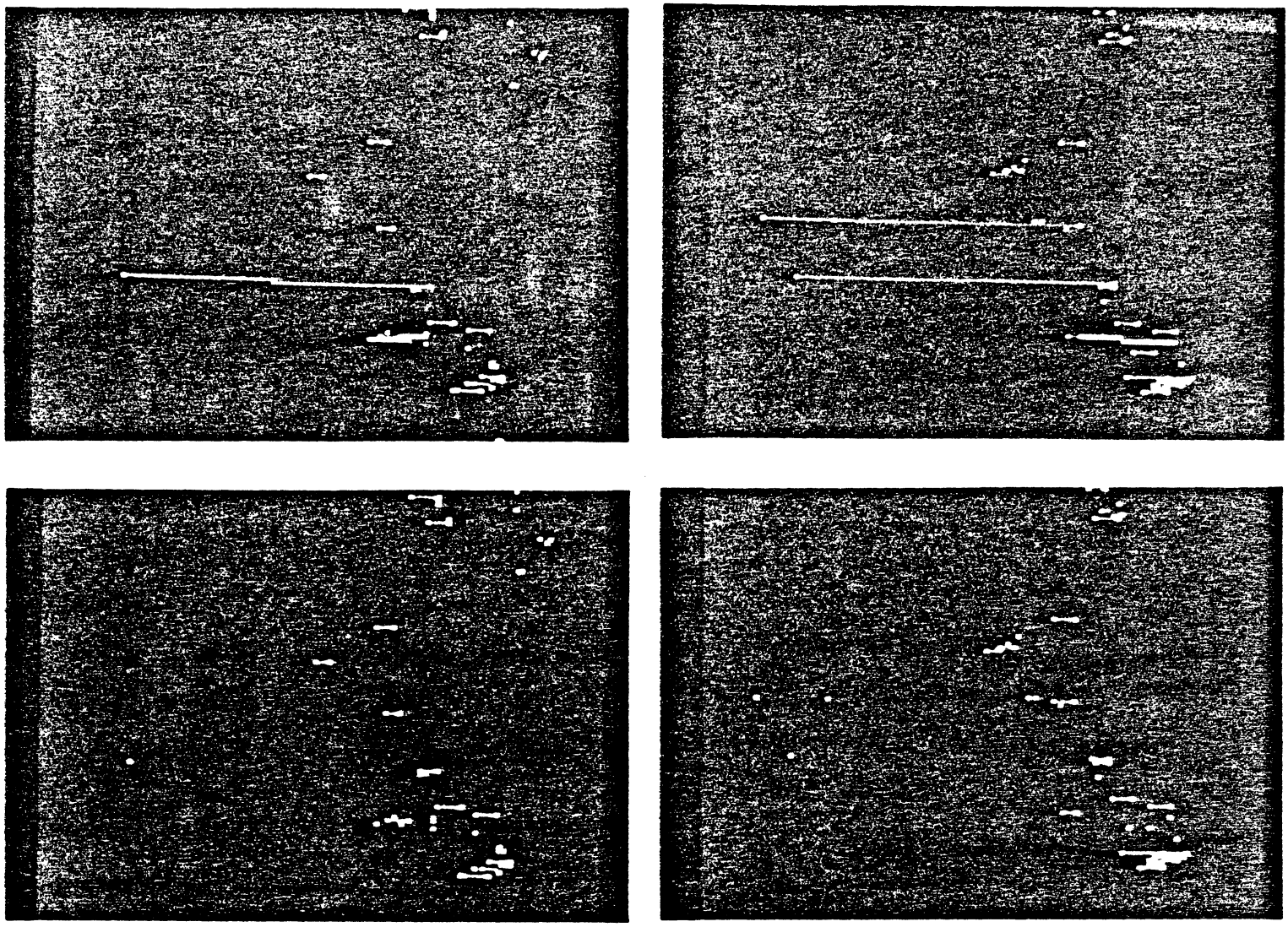

Figure 13 Match found by the algorithm. Images on the left are for frames 1 and 3. Images on the right are for frames 2 and 4. The top images show all the matches found. The bottom images show the corrected matches.