Katsushi Ikeuchi

Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Determining a Depth Map Using a Dual Photometric Stereo

Abstract

This paper describes a method for determining a depth map from a pair of surface-orientation maps obtained by a dual photometric stereo. A photometric stereo system *determines surface orientations by taking three images from the same position under different lighting conditions, based on the shading information. A photometric stereo system can determine surface orientations very rapidly, but cannot determine absolute depth values. This paper proposes a dual photometric stereo system to obtain absolute depth values.*

A dual photometric stereo *generates a pair of surface-orientation maps. Then, the surface-orientation maps can be segmented into isolated regions with respect to surface orientations, using a geodesic dome for grouping surface orientations. The resulting left and right regions are compared to pair corresponding regions. The following three kinds of constraints will be used to search for corresponding regions efficiently: a surface-orientation constraint, an area constraint, and an epipolar constraint. Region matching is done iteratively, starting from a coarse segmented result and proceeding to a fine segmented result, using a parent-children constraint. The horizontal difference in the position of the center of mass of a region pair determines the absolute depth value for the physical surface patch imaged onto that pair. This system takes only a few minutes on a Lisp machine to determine an absolute depth map in complicated scenes and could be used as an input device for a bin-picking system.*

1. Introduction

Vision is one of the most important subsystems of an intelligent robot. Without vision, a robot can repeat only one predetermined job sequence where objects are expected to be at predetermined places. Moreover, slight disturbances can cause unpredictable circumstances in the robot environment which might prevent the robot from continuing its job sequence. Such a robot system lacks flexibility and robustness.

The International Journal of Robotics Research,
Vol. 6, No. 1, Spring 1987.

Robot vision has been explored from various directions, including the binocular stereo method, the shape-from-shading method, the shape-from-texture method, and the shape-from-line-drawings method (Brady 1981). We have focused on the binocular stereo method and the photometric stereo method (part of the shape-from-shading method) because these methods can obtain the depth information robustly.

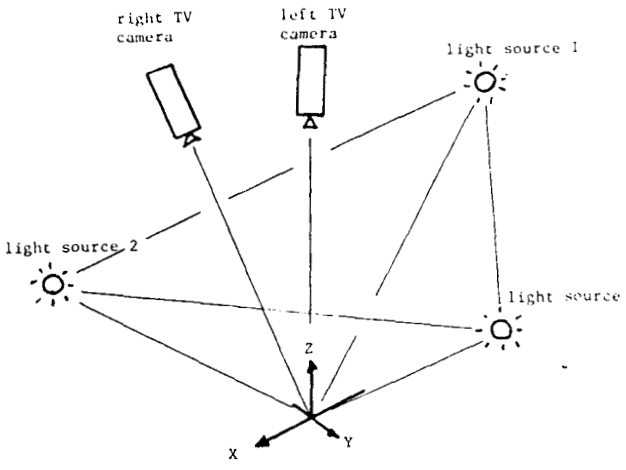
The binocular stereo method has been explored since the early days of robot vision research, because binocular stereo plays an important role in the human visual system. This area has been explored by Marr and Poggio (1979); Moravec (1979); Baker (1981); Grimson (1981); Barnard and Fishler (1982); Nishihara (1984); Thorpe (1984); and Ohta and Kanade (1985).

Binocular stereo methods are divided into two classes: one uses brightness correlation, and the other is based on feature matching. The *brightness correlation methods* divide the left and the right images into small windows and compare the brightness distribution in a window of the right image with distributions over candidate windows of the left image. The window pair with the highest correlation is declared to be a corresponding pair, and the horizontal difference between the windows gives the depth value. This correlation method is suitable for hardware implementation. It has, however, the following two defects:

1. If a window contains an occluding boundary, the apparent brightness distribution depends heavily on the viewer direction. Thus, it is difficult to determine a good match for such a window.
2. Depth values can be measured only at the center of the windows. In order to have robust matching, a larger window size is better. On the other hand, if the window size is large, sampling of the image is coarse.

The *feature matching method* searches for corresponding features in the left and the right images and re-

Fig. 1. Dual photometric stereo.



quires that feature points be extracted before matching. Depth information can be determined directly only at places where feature points exist. Many kinds of operators for extracting features have been proposed, including the so-called interest operator (Moravec 1979), the DOG filter (Marr and Hildreth 1980), and the Roberts operator (Roberts 1965). Since depth information can be determined only at feature points, it is desirable to have many such points. In this sense, the output of the DOG filter and the output of the Sobel operator are better than the output of the interest operator because output is produced by the DOG filter or by the Sobel in many more places in the images than by the interest operator. On the other hand, since the number of feature points is increased, it takes more time to compare all the candidate features. Thus, from the point of view of matching complexity, the output of the interest operator is better than the output of the DOG filter or the Sobel operator.

Although the binocular stereo method is robust and well explored, it takes a long time to compute depth with this method. Also the binocular stereo method determines depth only at the place where matching features are obtained; it requires an interpolation to get a continuous depth map. The photometric stereo method bypasses these problems. A *photometric stereo method* determines surface orientations from three images of the same scene from the same point under three different light conditions (Horn, Woodham, and Silver 1978; Woodham 1979; Ikeuchi 1981; Ikeuchi, et al. 1986). Since the spatial configuration between the TV camera and the object is the same for all three images, there is no disparity between the images. Therefore, no matching operations are needed and surface orientation can be determined very rapidly. Also the photometric stereo system can determine surface orientation at each pixel and can compute a dense surface-orientation map. Unfortunately, however, the photometric stereo system cannot produce an absolute depth.

This paper proposes a method to determine an absolute depth map based on the binocular use of two photometric stereo systems. A pair of surface-orientation maps can be determined by a dual photometric stereo. The surface-orientation maps obtained are segmented into isolated regions of the constant surface orientations using a geodesic dome. Comparing the left regions with the right regions gives pairs of corresponding regions. Matching depends on the size of the regions, average surface orientations, and mass-center positions of the regions. Finally, differences in mass-center positions of corresponding regions give the depths of the regions.

This system has the following features:

1. Since the features for the matching operations are the mass centers of regions, the number of matching operations is small. Thus, the system can determine a depth map rapidly.
2. Since the distributions of the surface orientations within each region are known from the surface-orientation map, we can convert the discrete depth information at the mass-center positions into a continuous depth map over the image using a simple integration operation.

Figure 1 shows the setup of our dual photometric stereo. A pair of TV cameras is surrounded by three light sources. The left TV camera's image plane is parallel to the ground plane and the right TV camera's image plane is inclined so that the optical axes of the TV cameras intersect at the origin of the spatial coordinate system, where the spatial origin is under the left TV camera. The objects are located near the spatial origin.

2. Segmentation on Surface Orientations

Segmentation is necessary for our stereo system, but segmentations of images based on brightness values are not suitable for our region-based stereo system for the following reasons:

1. Since the brightness value of a pixel depends on the viewer direction, a brightness value at a pixel in the left image is usually different from a brightness value at the corresponding pixel in the right image. The difference between two observing directions causes the difference between two brightness values at the corresponding pixels. Thus, a region obtained using brightness segmentation in the left image does not usually overlap with the corresponding region in the right image.
2. It is impossible to have two TV cameras with identical characteristics. A gray value observed always depends on the characteristics of the TV camera. Thus, it makes little sense to compare absolute values observed by the one camera with absolute values observed by another camera. For example, a region in the left image based on an absolute threshold will be different from a region in the right image, even though the two regions correspond roughly to the same physical surface patch, and the threshold value used is the same.

The above two defects of brightness-based segmentation occur because observed brightness is not a characteristic of an object surface but is a measure depending on both the viewer direction and the characteristics of the observing system. Thus, even though a pair of pixels in the left and the right images corresponds to the same physical point on an object, they may not have the same brightness values. Therefore, brightness-based segmentation is not a suitable basis for the binocular stereo.

We propose instead to segment an image into regions based on intrinsic characteristics of object surfaces. Intrinsic characteristics, such as surface orientations, are properties of the object surface itself (Barrow and Tenenbaum 1978). Intrinsic characteristics are independent of both the viewer direction and characteristics of the observing system. We should estimate the same value of intrinsic characteristics, provided that the pair of observing pixels on the left and right images correspond to the same physical point. A segmentation method based on an intrinsic characteristic is stable and has high reliability.

Images are segmented into isolated regions based on surface orientations. There are several kinds of intrinsic characteristics, such as surface orientation, albedo, and color. Surface orientation is the most easily obtained among the intrinsic characteristics. We propose to segment images into isolated regions based on surface orientations obtained by the photometric stereo method for our region-based stereo. The segmentation consists of three steps: (1) segmentation based on shadows, (2) segmentation based on orientation discontinuities, and (3) segmentation based on orientation classification.

2.1. Shadows (Coarse Segmentation)

Self-shadows, projected shadows, and mutual illumination disable the photometric stereo system in places around objects:

Projected shadow
A 3-D object projects shadows onto other objects below it. The photometric stereo system cannot determine surface orientations in the projected shadow regions of any one of the three lights. Since our three light sources are arranged in a triangle, undetectable regions due to the projected shadows exist entirely around the upper object.

Self-shadow
The photometric stereo system also cannot determine surface orientations of steep surface regions. Steep surface areas are self-shadowed. In any place where any one of the three light sources casts a self-shadow, the system cannot determine surface orientations. Self-shadows often occur near the object boundaries. The limiting angle of detectable surface orientation can be controlled. A photometric stereo uses a 3-D lookup table to convert a triple of brightness values into surface orientation. If we blank out entries of the lookup steeper than some

limiting angle, the system cannot determine surface orientations steeper than the limit angle. Since steep areas exist near the object boundaries, then the system cannot determine surface orientations near the object boundaries.

Mutual illumination

Mutual illumination occurs due to the light reflected by a near object. This indirect illumination changes the lighting condition at the region. The brightness triple obtained has no surface orientation at the corresponding entry of the 3-D lookup table, because the triple does not occur under the usual lighting conditions. The photometric stereo system cannot determine surface orientation of mutually illuminated surface regions. This mutual illumination occurs at the peripheral area of an object, which is a convenient characteristic for segmentation.

A surface-orientation map obtained by a photometric stereo system can be easily segmented into isolated regions corresponding to objects using the disabled regions around objects. Consider the following binary image. Each pixel is a 1 if the surface orientation can be determined at the pixel. Each pixel is a 0 if the surface orientation cannot be determined at that pixel. Regions of 1's correspond to object regions, and regions of 0's correspond to disabled regions in the scene. Thus, we can easily segment the map into isolated object regions to check the connectedness of regions of 1's.

2.2. Orientation Discontinuity (Coarse Segmentation)

Orientation discontinuities are also used for more stable segmentation. If two objects overlap each other and surface orientations are determined near the occluding boundaries (usually surface orientations cannot be determined there due to self-shadowing or a projected shadow), surface orientations are not continuous over the occluding boundaries. It rarely occurs that two overlapping objects have the same surface orientation on both sides of an occluding boundary. Thus, the orientation discontinuity divides what might otherwise be a connected region into two divided regions.

Orientation discontinuity may be measured by the following formula:

$$s = f_x^2 + f_y^2 + g_x^2 + g_g^2, \tag{1}$$

where (f, g) denotes a surface orientation using the stereographic plane and f_x is a partial derivative of f with respect to the x direction of the image plane and so on. Regions where s is larger than some threshold are considered to be places of orientation discontinuity.

2.3. Orientation Classification (Fine Segmentation)

A finer segmentation is necessary for more precise depth maps. Our stereo system can determine depth values at the mass centers of regions. Determining depth values inside an object requires finer regions. Thus, a finer segmentation is necessary to extract finer candidate regions.

A tessellated sphere may divide surface orientations into classes. Since surface orientation has two degrees of freedom, surface orientation can be represented by a point on the Gaussian sphere. We can divide the Gaussian sphere into cells using a tessellated sphere. A surface orientation can be assigned to a class, based on the cell that contains the point corresponding to the surface orientation.

A labelling operation can be applied to the cell-number map, after a surface orientation map is converted into a cell-number map. Each pixel of an image has a cell number on a tessellated sphere corresponding to the surface orientation there. Then, a labelling operation is applied to the cell-number map to extract connected regions. Each isolated region consists of surface patches having the same cell number.

A geodesic dome will be used for tesselating a sphere. A geodesic dome is obtained by projecting edges of a polyhedron to the surface of the circumscribing sphere with respect to the center. For example, a geodesic dome having 12 cells is obtained from a dodecahedron. Regular tessellation methods on the sphere only exist for 4, 6, 8, 12, 20 divisions, because tetrahedra, hexahedra, octahedra, dodecahedra, icosa-

hedra are the only regular polyhedra. Finer tessellations are obtained by division of the regular tessellation into smaller triangles (Wenninger 1979). We will use geodesic domes obtained by division of a dodecahedral geodesic dome for finer segmentation.

This finer segmentation using geodesic domes is done hierarchically. At the first stage, regions obtained by coarse segmentation using shadows and orientation discontinuity regions are divided into subregions using a dodecahedral geodesic dome. Figure 2A shows a hyperbolic surface and an elliptic surface. Figure 2B shows a needle map of a hyperbolic surface and a needle map of an elliptic surface. Figure 2C shows the segmented result using a dodecahedral geodesic dome. Each pixel on a connected region has the same cell number over the region. The resulting region map is again divided using a one-frequency, dodecahedral, geodesic dome. (See Fig. 2D.) This refining operation is applied iteratively using a higher geodesic dome until the necessary resolution is achieved. Fig. 2E gives the segmented result using a two-frequency, dodecahedral, geodesic dome.

There may be accidental errors due to mutual illumination or shadows, causing incorrect surface orientations at that area, but this does not cause errors in matching regions. This error does not occur randomly on the left and right needle maps, but occurs in the same way on the maps, because both maps are obtained under the same light source conditions. Note that the two TV cameras share the common light sources in Fig. 1. Thus, that area will be segmented in the same way in both the left and right images. This will give the correct matching at the area, even though the surface orientations obtained there are not correct.

3. Camera Model (Orthographic Projection)

A camera model is necessary to convert a disparity value between two images into a depth value. Figure 3 shows our camera configuration. The left image plane is perpendicular to the spatial z axis, while the right image plane is inclined with respect to the z axis, so that the two optical axes intersect with each other at the origin of the spatial coordinate system. Orthographic projection is used as the camera model. Since the distance between the TV camera and the object is far compared with the size of the TV camera's field of view, we will use the orthographic projection instead of the perspective projection as the camera model. Disparity between images is due only to the angle between two image planes. For example, the images of a physical point at the spatial origin have the same coordinates in the left and the right image planes. If the point moves towards the viewer, the image of the point on the left image plane keeps the same position while the right image of the point moves to the left on the right image plane. This gives the disparity with this orthographic projection.

For the left TV camera

$$\begin{aligned} u^l &= \mathrm{f}^l x + e^{ul}, \\ v^l &= \mathrm{f}^l y + e^{vl}, \end{aligned} \tag{2}$$

where f^l is a conversion constant from the spatial coordinate into the left camera coordinate; (x, y) is the spatial coordinate system; (u^l, v^l) is the left camera coordinate system whose origin exists at the center of the camera plane; and e^{ul}, e^{vl} are error terms (see Fig. 3).

The right TV camera gives

$$\begin{aligned} u^r &= \mathrm{f}^r(\mathrm{a}x + \mathrm{b}z) + e^{ur}, \\ v^r &= \mathrm{f}^r y + e^{vr}, \end{aligned} \tag{3}$$

where a, b are trigonometric constants determined from the angle between two image planes and e^{ur}, e^{vr} are error terms.

Thus, the relationship between image coordinates and depth z is

$$z = \mathrm{a}' u^l + \mathrm{b}' u^r + \mathrm{c}', \tag{4}$$

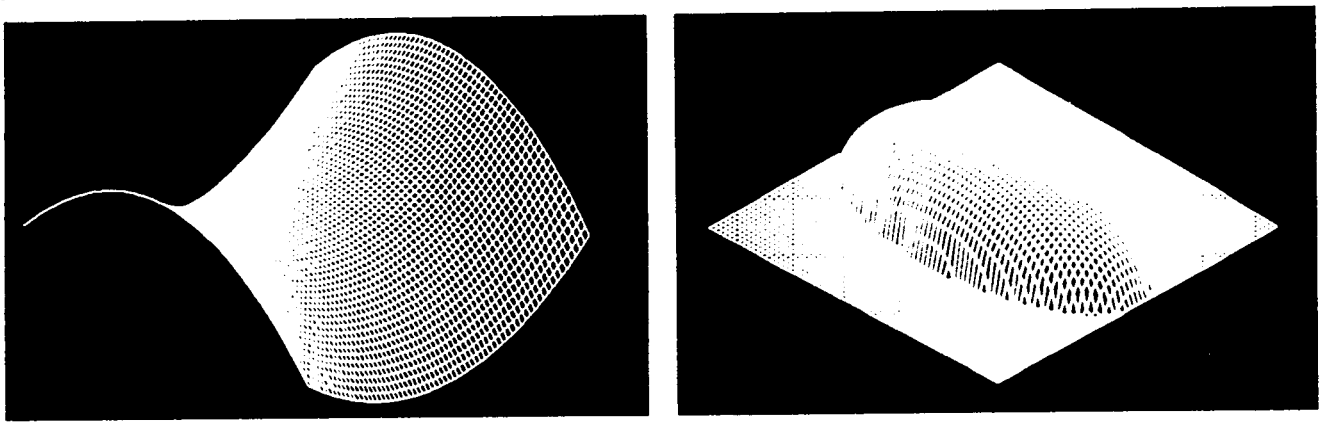
where

$$\begin{aligned} \mathrm{a}' &= -\mathrm{a}/(\mathrm{f}^l \mathrm{b}), \\ \mathrm{b}' &= 1/(\mathrm{f}^r \mathrm{b}), \\ \mathrm{c}' &= ((e^{ul}\mathrm{a})/\mathrm{f}^l) - (e^{ur}/\mathrm{f}^r)/\mathrm{b}. \end{aligned}$$

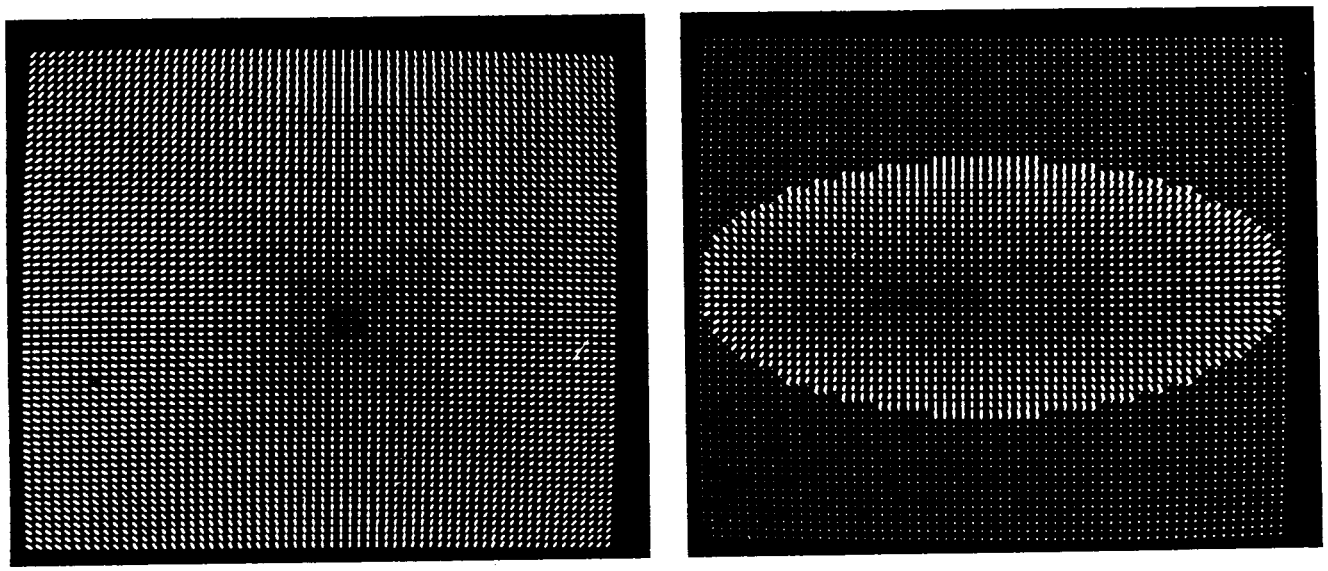
The constants a′, b′, c′ may be determined by means of a calibration. Note that the two optical axes of the two TV cameras intersect with each other at the origin of the spatial coordinates.

This camera model is used for two purposes. First, it converts a disparity value into a depth value. A hori-

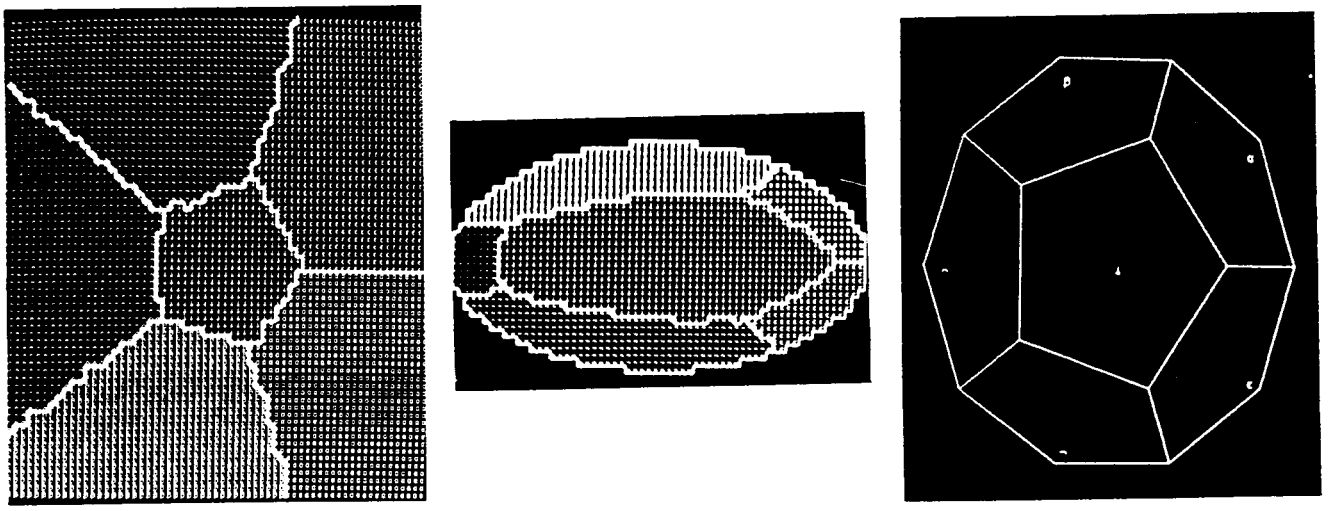
Fig. 2. Segmentation schema. A. A hyperbolic surface and an elliptic surface. B. A needle map of a hyperbolic surface and a needle map of an elliptic surface. C. Segmented result using a dodecahedral, geodesic dome. D. Segmented result using a one-frequency, dodecahedral, geodesic dome. E. Segmented result using a two-frequency, dodecahedral, geodesic dome.



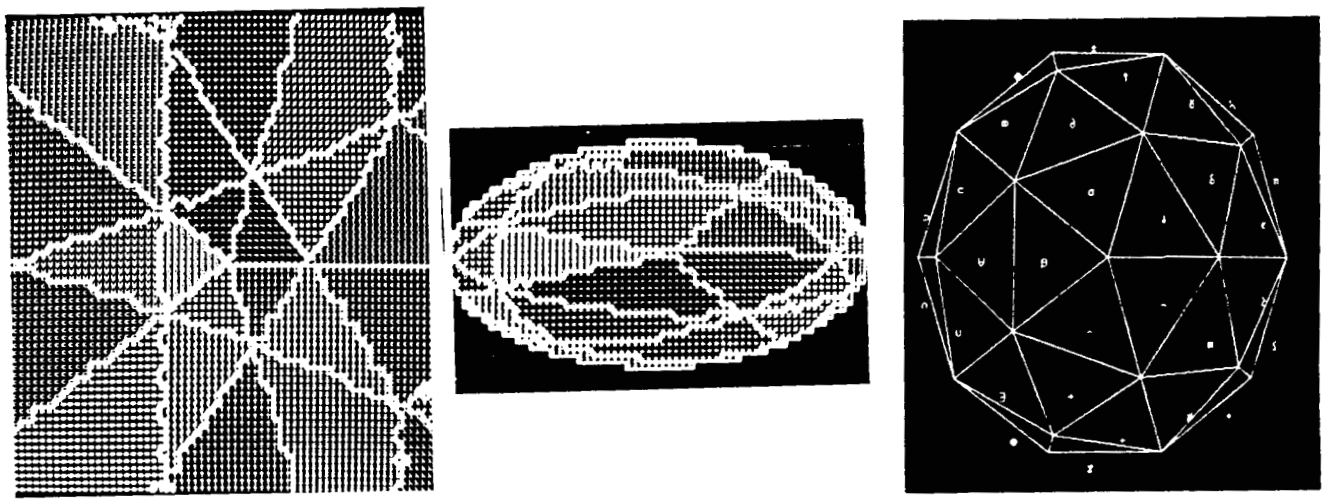
A



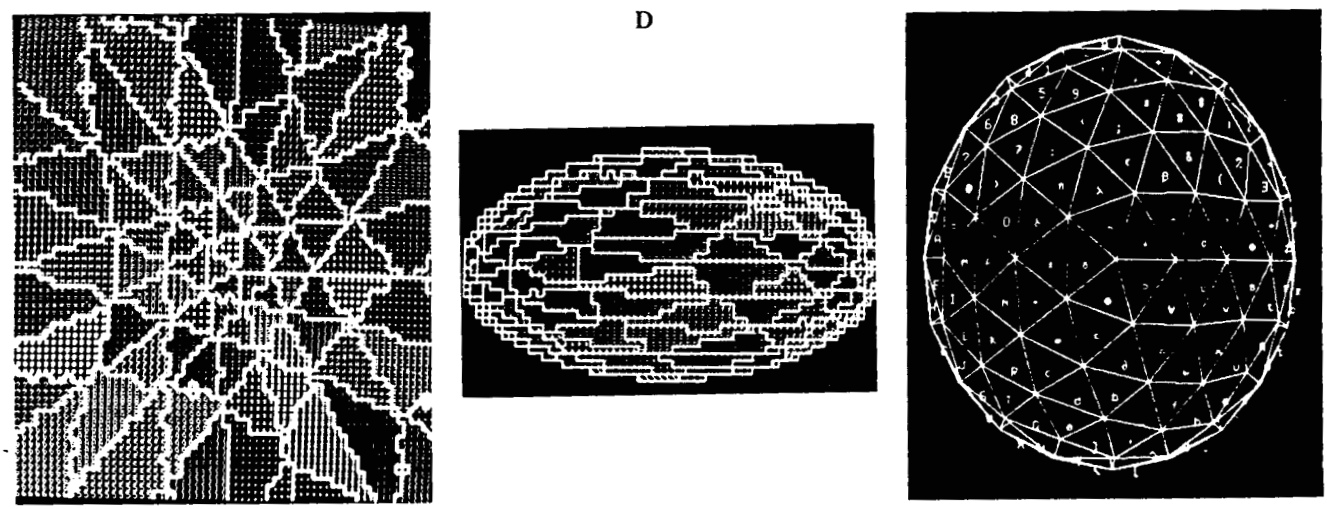
B



C



D



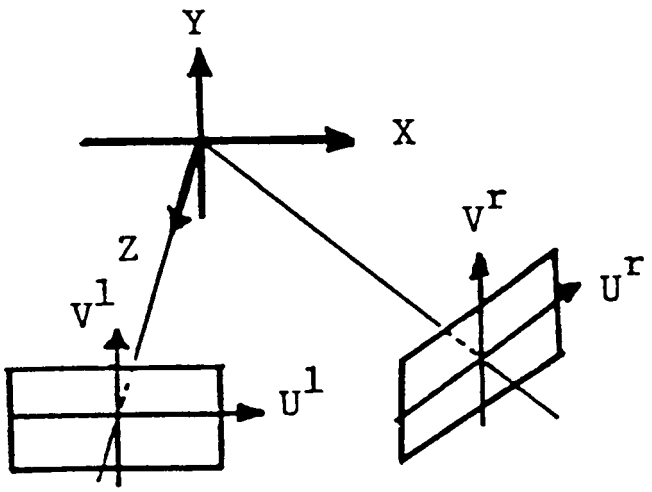
E

zontal difference in a region's mass-center positions gives a disparity value. This disparity value can be converted into a depth value using the camera model. Second, it constrains the possible area within which a candidate match can exist. Under this orthographic projection, two corresponding regions have the same y coordinate values in the left and right images. This constraint is the epipolar-line constraint (Barnard and Fishler 1982).

4. Region Matching

Matching of regions depends on characteristics of the regions. Since corresponding regions in images are projections of the same physical surface patch, corresponding regions have similar physical characteristics. Our system uses the following three characteristics to constrain possible candidate pairs: vertical mass-center

Fig. 3. Camera model.



positions, average surface-orientation directions, and area.

The matching operation also follows the coarse-to-fine strategy. At first, coarse-segmented regions, which have been obtained using the smoothness filter, are matched. Then, fine segmented regions using a geodesic dome are matched, based on the result of the coarse matching. The coarse-segmented result gives the disparity limits for the fine-segmented regions.

4.1. Constraints on Region Matching

The following four constraints are used to reduce the search space for matching efficiency: (1) surface orientation constraint, (2) area constraint, (3) epipolar-line constraint, and (4) parent-children constraint. The first three constraints reduce the search space between the left and right regions, and the fourth constraint reduces the search space in the coarse-to-fine direction.

Surface orientation constraint

If a region in the left image corresponds to a region in the right image, then the average surface orientation of the region in the left image should be similar to the average surface orientation of the region in the right image. Note that this criterion is based on the global coordinate system. Namely, each surface orientation in an image coordinate system is converted into the surface orientation in the global coordinate system. This is possible because our camera model is the orthographic projection.

Area constraint

If two regions correspond to each other, the two regions should have similar areas. Again this criterion is based on the global coordinate system. Since we know the average surface orientation of the region, we can convert the area size on the image plane into the area size on the plane perpendicular to the average surface orientation. This converted area size is independent of the local coordinate system. The resulting area size will be used for comparison.

Epipolar-line constraint

Our camera model projects a physical point onto the left and right image planes so that y coordinates of the two corresponding image pixels are the same. Thus, the corresponding mass-centers should have the same y coordinate in the image planes. Usually segmentation results and camera positions contain noise. Thus, the corresponding mass centers do not always have the same y coordinate. However, they have nearly the same y coordinate. The search process searches the corresponding mass center around the y coordinate within some limits.

Parent-children constraint

Regions are divided into finer subregions as we go along. We will call a region the *parent region* and the subregions the *children regions*. If a left parent region and a right parent region correspond to each other, a left child region from the left parent region should correspond to one of the right children regions from the right parent region: that is, a subregion's search area for matching is limited to the subregions of the parent's partner.

4.2. Matching Operation

Roughly speaking, the matching operation compares left and right region lists using (1) *orientation constraint*, (2) *area constraint*, (3) *epipolar-line constraint*. First, left region lists and right region lists are gener-

ated so that each list consists of regions having a common surface orientation (common cell number on the geodesic dome). Second, list pairs between left lists and right lists are generated so that each list pair has left and right lists whose regions have the same surface orientation (common cell number on the geodesic dome). Third, a matching algorithm establishes a region pair among regions in a list pair so that the left region and the right region of the region pair have the similar area size and position of mass-center.

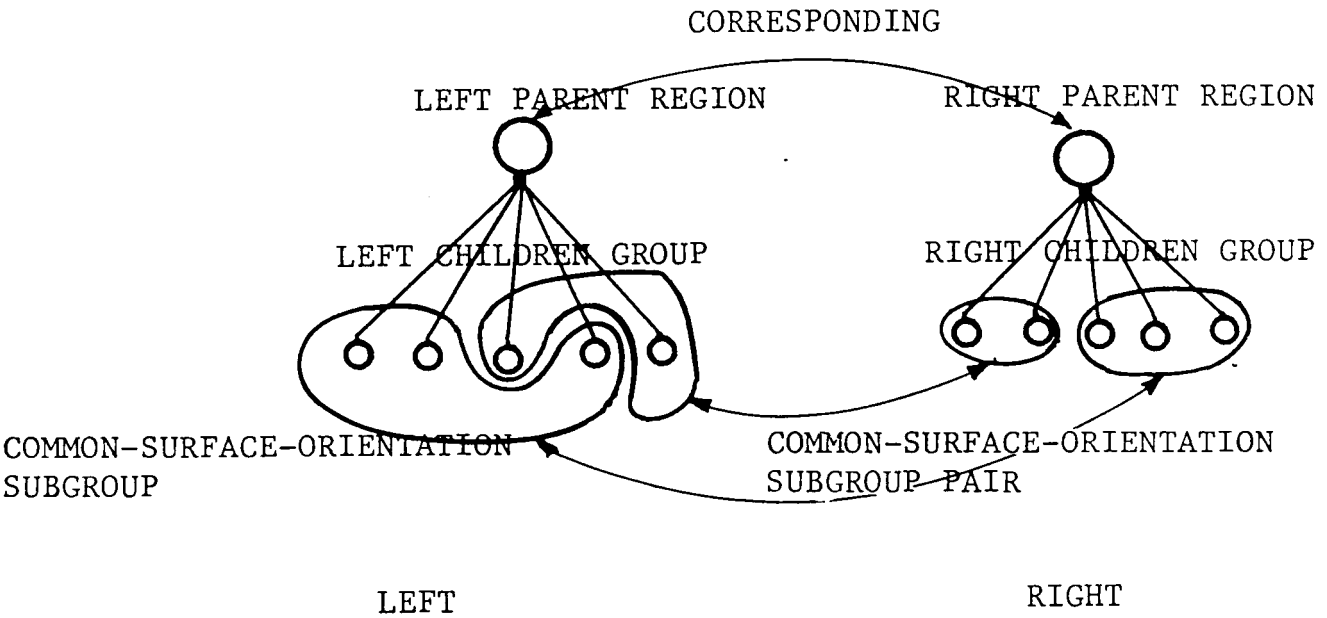
This matching algorithm follows the coarse-to-fine strategy. At the coarsest stage, region matches will be established between the left and the right region maps using the smoothness filter. Then, region matches will be refined between the left and the right region maps by a dodecahedral geodesic dome using a *parent-children constraint.* These region matches will be refined iteratively between region maps obtained by a one-frequency, dodecahedral, geodesic dome; by region maps obtained by a two-frequency, dodecahedral, geodesic dome; and so on, until the desired accuracy is obtained. The following nine steps show the region-matching procedure.

1. *Making children groups*
   Children groups will be generated from region maps. Each children group consists of regions sharing a common parent region. In particular, the first stage has only one children group, which consists of all regions segmented by the smoothness filter. This is a preparation step for utilizing *the parent-children constraint.*
2. *Making common surface-orientation subgroups*
   Each children group will be divided into common surface-orientation subgroups (CSO subgroups) so that each CSO subgroup consists of regions having common surface orientation: that is, each region in a CSO subgroup shares both the same parent and the same surface orientation. This step is due to *the surface orientation constraint.*
3. *Generating CSO subgroup pairs*
   CSO subgroup pairs are searched among the left CSO subgroups and the right CSO subgroups whose parents are known to be corresponding regions from the previous iteration. (At the very beginning, all the regions are regarded as sharing the same parent. See Step 1.) If a left CSO subgroup and a right CSO subgroup have the same surface orientation, these two subgroups are registered as a subgroup pair. See Fig. 4.
4. *Deciding a target subgroup pair*
   One subgroup pair is selected for matching. Note that one subgroup pair contains a left CSO subgroup and a right CSO subgroup.
5. *Deciding a target CSO subgroup*
   The subgroup having fewer regions is selected for matching between the left CSO subgroup and the right CSO subgroup of the target subgroup pair. For the sake of explanation, let the left CSO subgroup contain fewer regions than the right CSO subgroup.
6. *Deciding a target region*
   The largest region among the left CSO subgroup is selected for matching. This region is called the *target region.*
7. *Matching the target region with candidate regions*
   Each region keeps the following five properties for matching: surface orientation, parent region, area size, horizontal disparity, and vertical disparity. Surface orientation, parent region, and horizontal disparity reduce candidate regions. Area and vertical disparity give a criterion to determine the most likely region.

   A region corresponding to the target region is searched among the right CSO subgroups. Note that the regions in the right CSO subgroup have the same surface orientation and the same parent region as the target region. Thus, this stage can be regarded as applying the surface-orientation constraint and the parent-children constraint to the search process.

   Only regions in the right CSO having horizontal disparity within some limit are considered as candidates. Since our cameras have optical axes intersecting at the ground level, the negative horizontal disparity never occurs. This gives the negative limit ϵ_1 of horizontal disparity. Also our system has a height limit in measurement. Namely, a region that is too high cannot be observed by the photometric stereo

Fig. 4. Common surface-orientation subgroup pairs.



system due to distortion of the orthographic camera model. Thus, this height limit gives the positive horizontal disparity ϵ_2.

Since a previous coarse stage gives depth information over a region, a disparity value d can be predictable. This d may contain an error term ϵ_3, where ϵ_3 can be determined as a function of the angular difference between the previous segmentational dome cell and the present cell. Regions whose horizontal disparity satisfies the following condition are considered as candidates.

$$\{x|(\max x_r - \epsilon_1, x_r + d - \epsilon_3) < x < (\min x_r + \epsilon_2, x_r + d + \epsilon_3)\},$$

where x_r is the mass center of the target region and d is the observed disparity obtained in the previous stage along the coarse-to-fine strategy. (In the coarsest stage, d is set to zero and ϵ_3 is set to infinity.)

The area constraint and the epipolar constraint are used to measure similarity between the target region and a candidate region among the right CSO subgroups. The evaluation function is as follows:

$$\left(1 - \frac{\Delta \text{ area}}{\text{area}}\right) \times \left(1 - \frac{\Delta y}{\Delta y \text{ dist\_limit}}\right)$$

where *area* is the area size of the target region, Δ *area* is the difference between the area size of the target region and that of the candidate region; *y dist_limit* is the limit of allowable vertical disparity, and Δy is the vertical disparity between the mass-center position of the target region and that of a candidate region.

7.1. *Success.*
If a region corresponding to the target region is found among the regions of the right CSO subgroup, the region pair is registered. Then, the target region is deleted from the left CSO subgroup and the corresponding region is deleted from the right CSO subgroup.

7.2 *Failure.*
If a corresponding region cannot be found in the right CSO subgroup, only the target region is deleted from the left CSO subgroup.

8. *Applying all the left region*
   Operations 6–7 will be applied to all the regions of the left common surface-orientation subgroups until the left subgroups are exhausted.
9. *Applying all the region pairs*
   Operations 5–8 will be applied to all the subgroup pairs until the subgroup pairs are exhausted.

5. Iterative Smoothing Operation

An *iterative smoothing operation* obtains a precise depth map. A coarse depth map by the region matching method will be smoothed into a precise depth map obtained by this iterative method. This iterative method is not a simple integration but a smoothing operation based on surface orientations and differences in surface orientations between corresponding pixels.

An iterative equation is constructed to satisfy the following conditions:

1. Observed (p, q) should agree with the first derivative of z with respect to the image axis, x, y, respectively. This is the definition of (p, q). Thus,

$$s = (z_x - p)^2 + (z_y - q)^2 \tag{5}$$

   should be zero everywhere in the resulting precise depth map.
2. A pair of corresponding pixels on the left and right images should have the same surface orientation, because the pixels must represent the same physical point. Thus,

$$\begin{aligned} d^p &= (p^r(ax + bz + c, y) - p^l(x, y))^2, \\ d^q &= (q^r(ax + bz + c, y) - q^l(x, y))^2 \end{aligned} \tag{6}$$

   should be zero, where (p^l, q^l), (p^r, q^r) are the surface orientations on the left and right surface-orientation maps. Here a, b, c are parameters determined from the spatial relationship between the two cameras.

By using Eqs. 5 and 6, the following equation is obtained.

$$e = \int (d^p + d^q)dxdy. \tag{7}$$

We will find the functional z that minimizes e.

From the Euler-Lagrange equation,

$$\begin{aligned} \Delta^2 z = (p^l{}_x + q^l{}_y) + (\lambda b) \\ \times ((p^r - p^l)p^r{}_x + (q^r - q^l)q^r{}_x). \end{aligned} \tag{8}$$

Using these approximation equations,

$$\Delta^2 z_{i,j} = k(\bar{z}_{i,j} - z_{i,j}), \tag{9}$$

$$\bar{z}_{i,j} = (z_{i+1,j} + z_{i-1,j} + z_{i,j+1} + z_{i,j-1})/4, \tag{10}$$

we have

$$\begin{aligned} z^{n+1} = z^n - \lambda(p^l{}_x + q^l{}_y) - (\lambda\rho b) \\ \times ((p^r - p^l)p^r{}_x - (q^r - q^l)q^r{}_x), \end{aligned} \tag{11}$$

where λ, ρ, b are constants.

6. Experiment

6.1. System Set-Up

In our system, the left TV camera is roughly 2 m above the ground, and the TV camera's image plane is parallel to the ground plane. The right TV camera is located 30 cm to the right of the left TV camera, and that TV camera's image plane is inclined by 7 degrees so that the optical axes of these TV cameras intersect at the origin of the spatial coordinate system, where the spatial origin is under the left TV camera. See Fig. 1.

Parameters for conversion from disparity values into depth values will be determined experimentally. Namely, parameters a′, b′, c′ in Eq. 4 will be determined by the least-squares method. A checkerboard pattern is observed by the TV cameras. The addresses of feature points (intersection points) on the TV images will give u^l and d, while the checkerboard's height is substituted into z. The least-squares fitting gives a′, b′, c′ from these observed values.

Fig. 5. Angular accuracy.

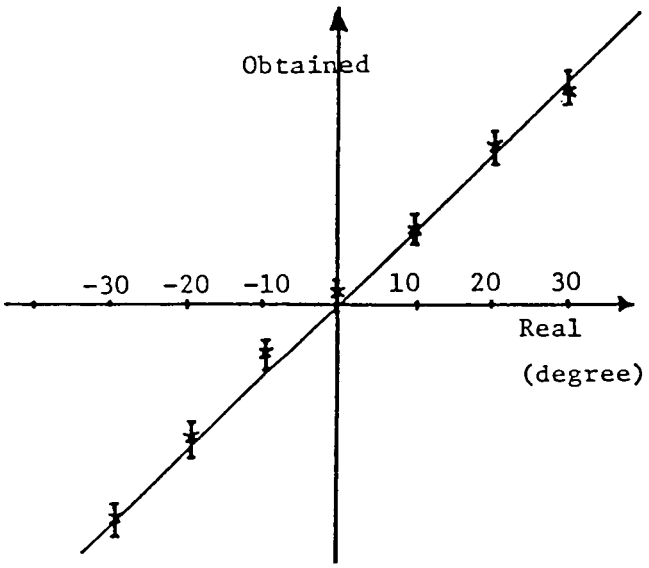
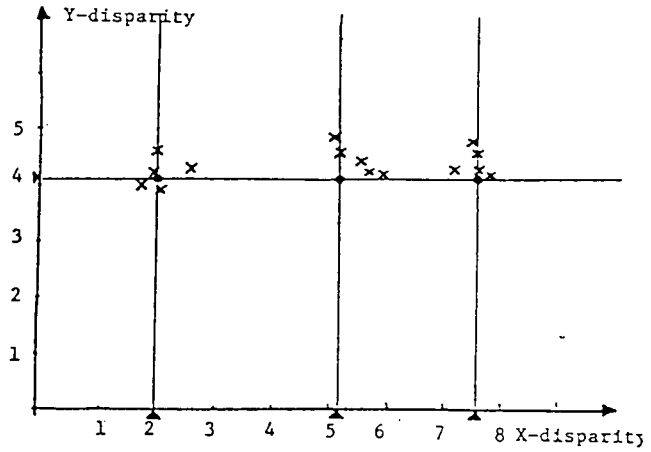


Fig. 6. Accuracy in depth measurement.



The lookup tables of the left and the right photometric stereo are obtained using a calibration sphere. The following operations will be executed in both the left and the right cameras so that the left and right lookup tables are obtained. First, the occluding boundary of the sphere is extracted after turning on all three light sources. Second, three pictures of the sphere after turning on one of the three light sources are taken from the camera. Third, a brightness triple at each pixel is sampled. Fourth, the surface orientation at the pixel is registered at a boxel, corresponding to the triple, of the 3-D lookup table. These operations give the left and right lookup tables.

(p, q)'s on the right lookup table are converted into (p, q)'s on the left camera coordinate system. Since surface orientation of the right lookup table is expressed on the right camera's coordinate system, it is converted into the surface orientation in the left camera's coordinate system. This conversion constant is obtained by the angle between the left and right image planes.

The visible area from the left camera is always visible from the right camera. A photometric stereo system can observe areas where three brightness values are higher than some threshold. The limiting angle is the radius of the spherical circle circumscribing this area. The limiting angle of the left photometric stereo is less than 50 degrees, while the angle between the left and right image planes is 7 degrees. Thus, the visible area from the left camera is always visible from the right camera. Note that the left and right cameras share the same lighting system.

6.2. Accuracy of this System

The accuracy of this system depends on the accuracy of the photometric stereo system, because our segmentation is based on surface orientations by the photometric stereo system. The accuracy of the photometric stereo system is measured by determining surface orientations of a white board inclined at various angles. Figure 5 shows both the angles obtained experimentally and the true angles. It also shows that the photometric stereo system can determine surface orientations within a 3-degree error.

The accuracy in depth is measured by determining the known height of blocks. Figure 6 shows the measured disparity values and the real disparity values. The horizontal axis denotes x disparity and the vertical axis denotes y disparity. The x disparity gives depth information. Even though the ideal case gives 0 disparity in y direction, our system gives 4.1 mean disparity due to a vertical tilt between the two TV cameras. A 0.7 pixel disparity error occurs in both directions. This error corresponds to 1 cm error in depth. Note that since a mass-center position is deter-

Fig. 7. Input scene.

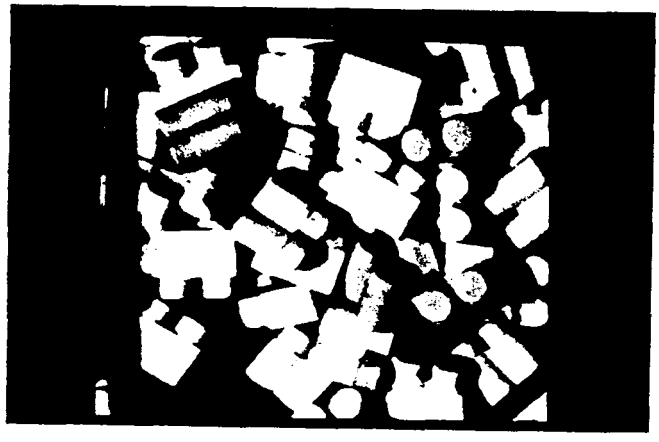
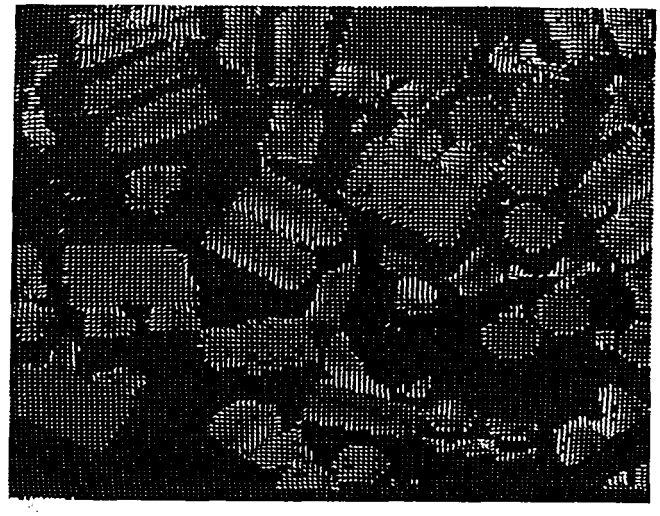
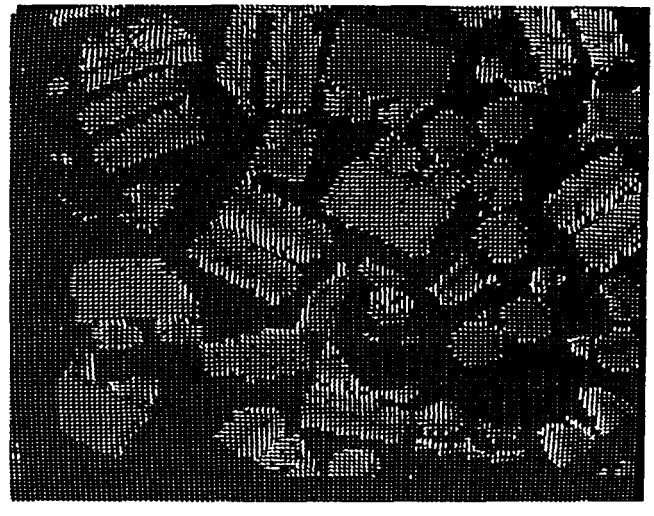


Fig. 8. A pair of surface-orientation maps.





mined by a subpixel resolution, a disparity is also determined by a subpixel resolution.

Our earlier discussion assumes that epipolar lines are parallel to scan lines on the image planes. However, real epipolar lines are not always parallel to scan lines. The maximum difference of an epipolar line from a scan line occurs at the corner of the image, provided that the scan line and the epipolar line intersect each other at the center of the image plane and can be approximated by $(dx)/z \tan \gamma$, where x is the physical length of the observable area, z is the distance of the point from the TV camera, d is the size of the image planes, and γ is the angle between the two image planes of the TV cameras. In our case, $x = 15$ cm, $z = 200$ cm, $d = 64$ pixels, $\gamma = 7$ degrees. Thus, the maximum relative difference of an epipolar line from a scan line gives 0.6 pixels. This is the same order as the observed measurement error, so we can ignore the nonparallel effect.

6.3. Experiment 1: Bin of Parts

The system was applied to the scene shown in Fig. 7. Surfaces of each object have the Lambertian property. Three pairs of images were taken under three different light sources using the left and right TV cameras. The pair of photometric stereo systems converted the three pairs of images into a pair of surface-orientation maps, shown in Fig. 8.

The smoothness filter applied to these surface-orientation maps converts them into the segmented region maps shown in Fig. 9. A mass-center position, an area size, and a mean-surface orientation are calculated for each isolated region. These characteristics of regions reduce the matching possibilities between the left regions and the right regions.

Comparing the left regions with the right regions gives corresponding region pairs with similar characteristics. Horizontal disparities in mass-center positions give the depth values at mass centers using the camera model. Figure 10 shows the resulting depth map. Since this segmentation by the smoothness filter

Fig. 9. A pair of segmented region maps.

Fig. 10. A depth map.

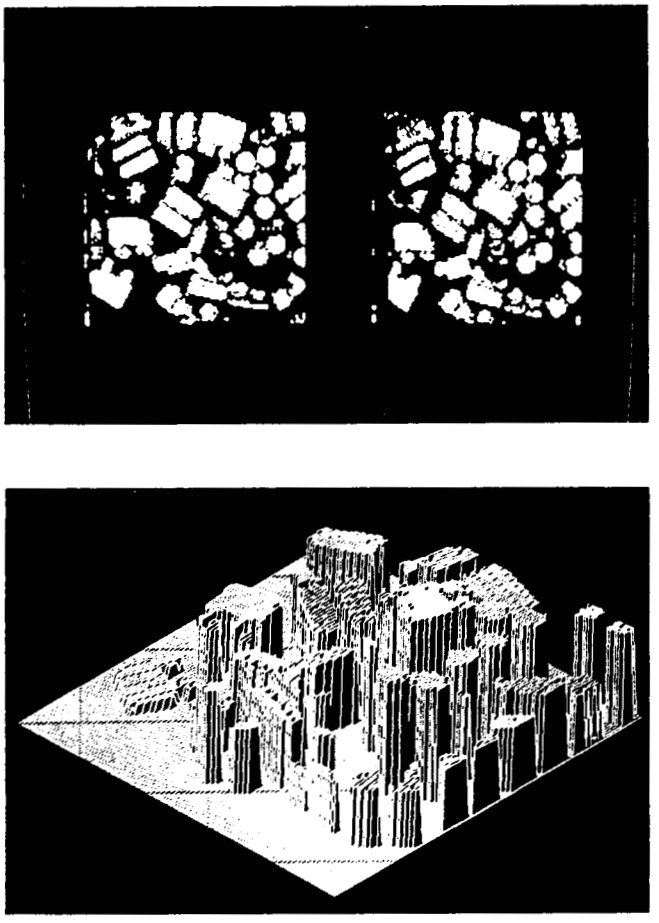
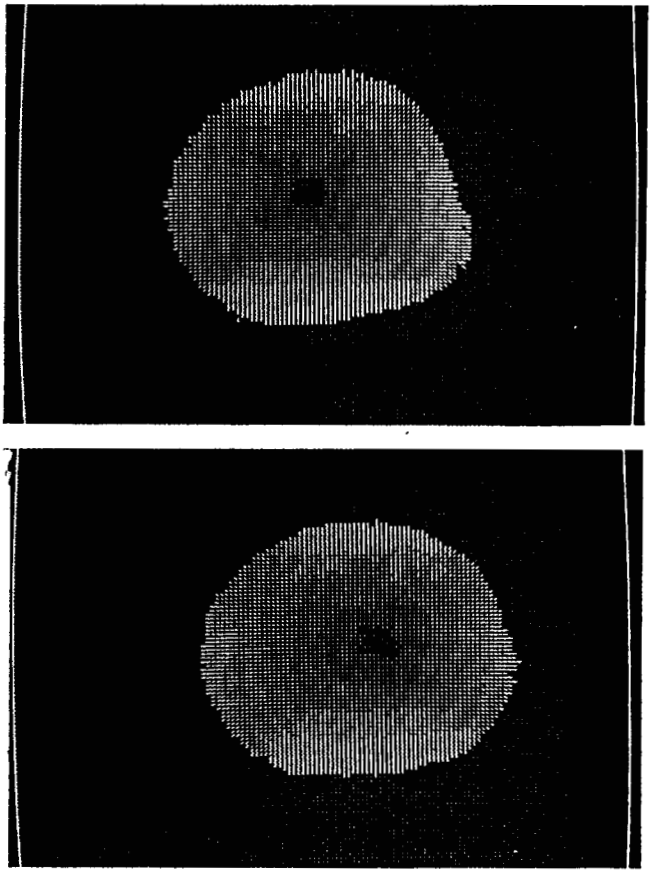


Fig. 11. A pair of surface-orientation maps of a sphere.



gives precise enough regions for this scene, the system is stopped at this stage of iteration.

This system runs on a Symbolics 3600 and takes roughly 10 s to get the surface-orientation map, 20 s to get the region map, and 5 s to get the depth map comparing the left and right segmented regions. Overall, about 1 min is necessary to get the final depth map from the three pairs of 128 × 128 brightness images.

6.4. Experiment 2: Sphere

The next example is a sphere. The pair of photometric stereo systems gives a pair of surface-orientation maps (Fig. 11). The smoothness filter converts these surface-orientation maps into a pair of region maps (Fig. 12A). The matching operation between these region maps gives the depth map (Fig. 12B). The segmentation by a dodecahedral geodesic dome gives the region maps as shown in Fig. 13A from the pair of surface-orientation maps and the pair of previous region maps. Using both these segmented results and the coarse depth map in the previous stage, the region matching operation gives the finer depth map shown in Fig. 13B. The same operation is repeated using a one-frequency, dodecahedral, geodesic dome and gives the result as shown in Fig. 14. A two-frequency, dodecahedral, geodesic dome gives Fig. 15.

The iterative smoothing operation described in Section 5 gives the results as shown in Fig. 16 after 300 iterations, starting from the depth map shown in Fig.

Fig. 12. A depth map by the smoothness filter. A. A pair of region maps by the smoothness filter. B. A depth map obtained.

Fig. 13. A depth map by a dodecahedral, geodesic dome. A. A pair of region maps by a dodecahedral, geodesic dome. B. A depth map obtained.

Fig. 14. A depth map by a one-frequency, dodecahedral, geodesic dome. A. A pair of region maps by a one-frequency, dodecahedral, geodesic dome. B. A depth map obtained.

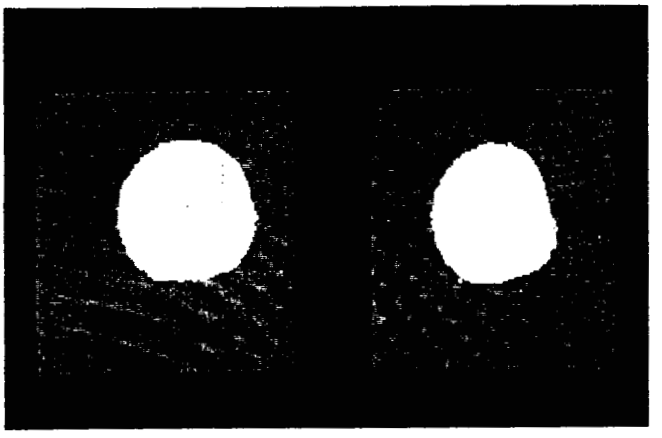
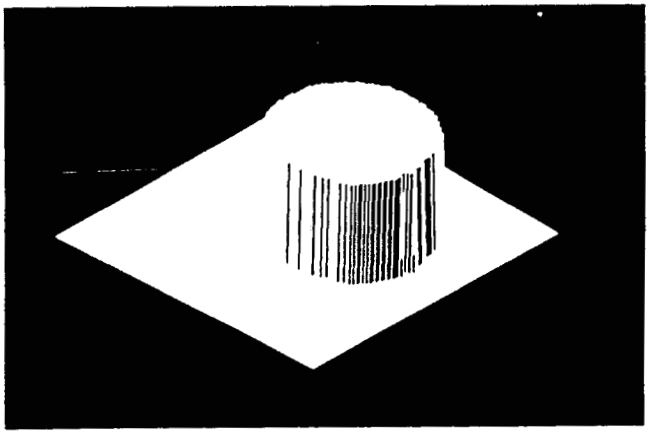


Fig. 12A



B

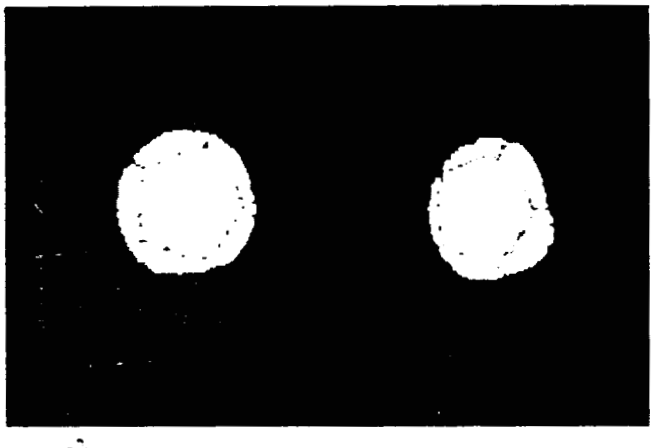
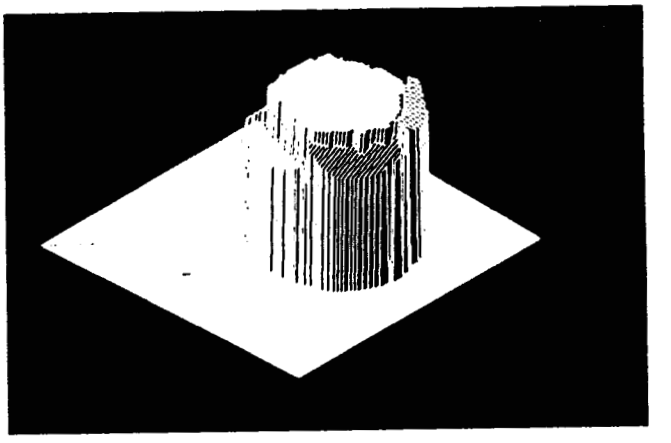


Fig. 13A



B

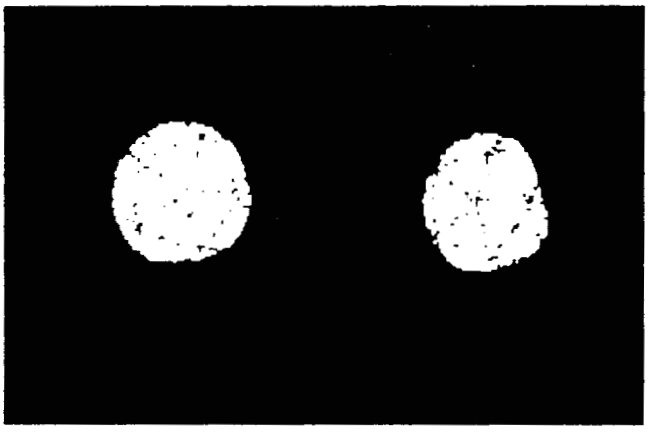
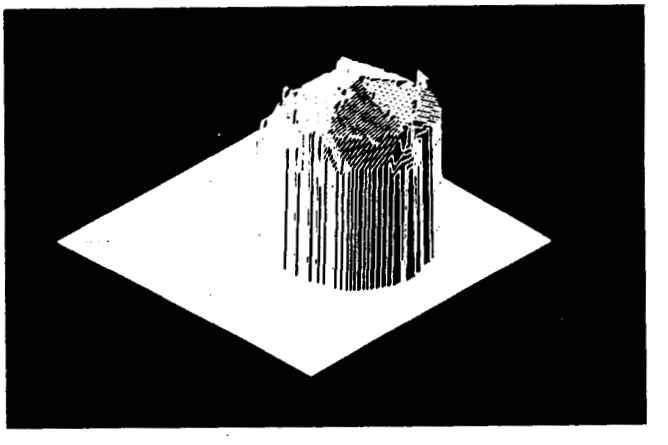


Fig. 14A



B

Fig. 15. A depth map by a two-frequency, dodecahedral, geodesic dome. A. A pair of region maps by a two-frequency, dodecahedral, geodesic dome. B. A depth map obtained.

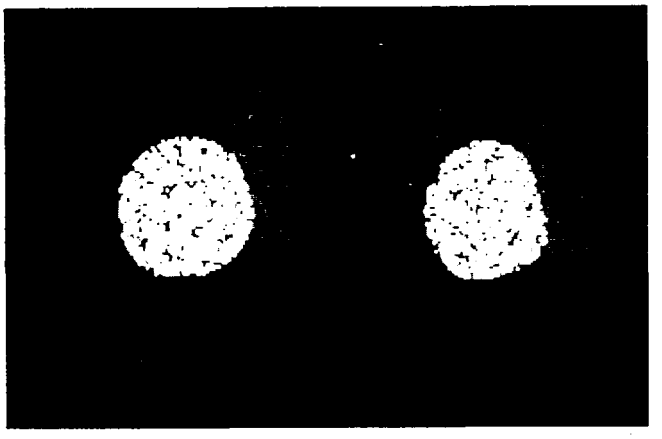


Fig. 15A

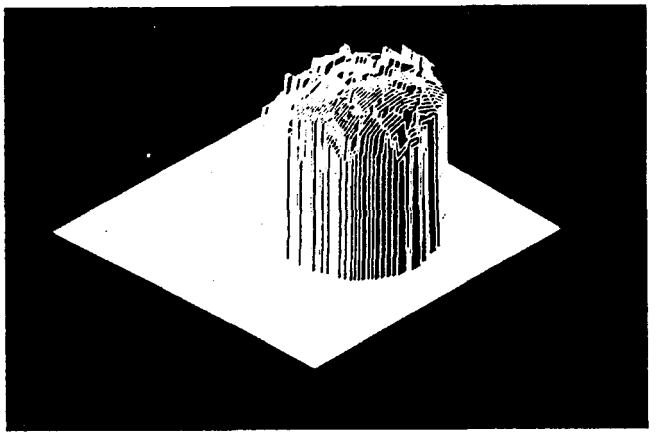
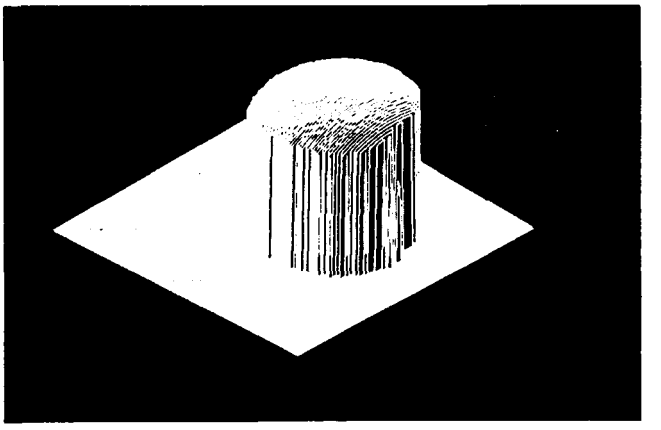
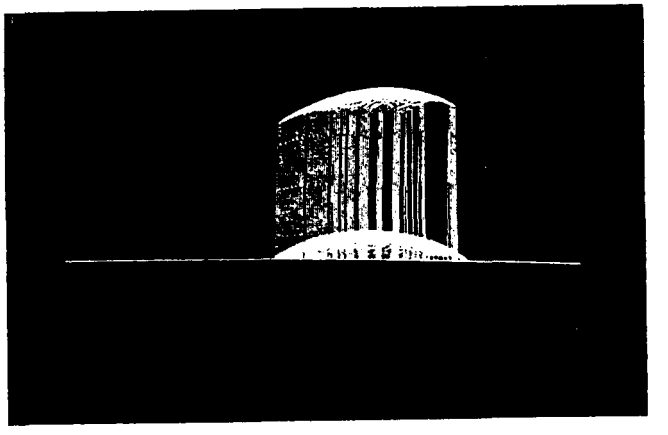


Fig. 16A

Fig. 16. A depth map obtained by the iterative smoothing operation. A. A perspective view of the obtained surface. B. A side view of the obtained surface.



B



B

15. Figure 16B contains two diagrams: the upper shape is the one obtained by this operation and the lower shape is the one obtained by directly integrating surface orientations of the left needle map. The vertical difference is the absolute depth determined by this iterative algorithm. The upper shape's distortion is due to the error by this iterative algorithm.

7. Conclusion

This paper describes a region-based stereo method using a pair of surface-orientation maps. This algorithm consists of the following components:

1. A pair of surface-orientation maps is obtained from a pair of photometric stereo systems.
2. A pair of surface-orientation maps is segmented into a pair of isolated region maps on surface orientations using a geodesic dome.
3. Feature points for stereo matching are mass centers of isolated regions.
4. The area constraint, the mean surface-orientation constraint, and the epipolar constraint improve the efficiency of search operations.
5. The process follows the coarse-to-fine strategy. At the beginning stage, coarsely segmented regions are compared and a coarse depth map is generated. This coarse depth map is used as the input to the next stage. The parent-chil-

dren constraint is applied to make the search operation efficient using this strategy.

We do not segment images using brightness thresholds. Segmentation for binocular stereo should use an intrinsic property of the object surface, such as surface orientation; the segmentation should not be based on a property such as apparent brightness, which depends on the viewer direction. Only an intrinsic property gives a segmentation result independent of the viewer direction. Other possible intrinsic properties suitable for region-based stereo would be color and albedo.

Acknowledgments

B. K. P. Horn, Takeo Kanade, S. A. Shafer, and the referees provided many useful comments, which have improved the readability of this paper.

References

Baker, H. 1981. (Vancouver, B.C.). Edge-based stereo correlation. *Proc. IJCAI-7*:631–636.

Barnard, S. T., and Fishler, M. A. 1982. Computational stereo. *ACM Computing Survey* 14(4):553–572.

Barrow, H. G., and Tenenbaum, J. M. 1978. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, eds. A. Hanson and E. Riseman, pp. 3–26. New York: Academic Press.

Brady, J. M. 1981. *Computer vision.* Amsterdam: North-Holland.

Grimson, W. E. L. 1981. *From images to surfaces.* Cambridge: MIT Press.

Horn, B. K. P., Woodham, R. J., and Silver, W. M. 1978. Determining shape and reflectance using multiple images. AI memo 490. Cambridge: Massachusetts Institute of Technology Artificial Intelligence Laboratory.

Ikeuchi, K. 1981. Determining surface orientations of specular surfaces by using the photometric stereo system. *IEEE Trans. PAMI,* PAMI 2(6):661–669.

Ikeuchi, K., et al. 1986. Determining grasp configurations using photometric stereo and the PRISM binocular stereo system. *Int. J. Robotics Res.* 5(1):46–65.

Marr, D., and Hildreth, E. 1980. Theory of edge detection. *Proc. Royal Society of London B.* 207:187–217.

Marr, D., and Poggio, T. 1979. A computational theory of human stereo vision. AI Memo 451. Cambridge: Massachusetts Institute of Technology Artificial Intelligence Laboratory.

Moravec, H. 1979 (Tokyo). Visual mapping by a robot rover. *Proc. IJCAI 6*:598–600.

Nishihara, H. K. 1984. PRISM: a practical realtime imaging stereo matcher. AI Memo 780. Cambridge: Massachusetts Institute of Technology Artificial Intelligence Laboratory.

Ohta, Y., and Kanade, T. 1985. Stereo by intra- and inter-scanline search using dynamic programming. *IEEE Trans. PAMI* PAMI 7(2):139–154.

Roberts, L. G. 1965. Machine perception of three-dimensional solids. *Optical and electro-optical information processing,* ed J. T. Tippett, pp. 159–197. Cambridge: MIT Press.

Thorpe, C. E. 1984 (April 30–May 2, Annapolis, Md.). An analysis of interest operators for FIDO. *Proc. IEEE Workshop on Computer Vision: Representation and Control:* 135–140.

Wenninger, M. J. 1979. *Spherical models.* New York: Cambridge University Press.

Woodham, R. J. 1979. Reflectance map techniques for analyzing surface defects in metal casting. AI-TR-457. Cambridge: Massachusetts Institute of Technology Artificial Intelligence Laboratory.