Depth Measurement by Motion Stereo

RAMAKANT NEVATIA

Image Processing Institute, University of Southern California, Los Angeles, California 90007

Communicated by A. Rosenfeld

Received June 15, 1975

Depth information can be of great value in the analysis of scenes of three-dimensional objects. A stereo pair of images allows extraction of such depth information. Use of a number of progressive, closely spaced views has the advantage of reliability over the use of just two views for stereo, and potential savings of computational effort. Such progressive views are naturally available in many applications; such as for a moving robot vehicle, pictures taken at different times in the flight of an aircraft, and for industrial parts placed on a slowly moving conveyor belt. Experimental results on laboratory scenes, including average error and speed of execution, are presented.

1. INTRODUCTION

A major difficulty in the analysis of pictures of a three-dimensional scene is the lack of direct depth information in a single picture. The projection of three-dimensional objects onto a two-dimensional image plane necessarily results in loss of information about one dimension. Techniques for segmenting a scene usually rely on discontinuities of surface attributes such as brightness or color. However, it is a depth discontinuity that most naturally corresponds to the boundaries of objects in three-dimensional space.

Humans presented with a photograph of a three-dimensional object are able to infer depth relations in the picture with little difficulty. Among the many cues believed to be used by humans are texture gradients, shadows, occlusion, and size of familiar objects. Depth inference from such cues is complex and poorly understood; some computer implementations that work on a limited variety of scenes are described in [1, 2]. Much effort has been spent on segmentation without using depth. Most of this work has been limited to scenes of polyhedral objects. Effectiveness of depth information in segmenting complex scenes has been demonstrated [3].

Several depth measurement techniques have been proposed; a laser ranging technique is described in [4]. Usually, these techniques require use of non-pictorial data or implements and are not always suited for unconstrained environments. Use of stereo is attractive in that it needs only pictorial information about the scene.
2. PRINCIPLES OF MOTION STEREO

In the image of an opaque object, each point of the image corresponds to one point of the object, assuming an ideal pinhole camera as the imaging device. Further, with each image point is associated a ray along which the corresponding object point must lie. The distance of the object point along this ray cannot be determined from a single monocular view in any direct way. If another image of the same object point is available, with the camera in a different position, then the object point is constrained to lie along two known rays in space and so must lie at their intersection (e.g., see points $P_1$ and $P_2$ in Fig. 1).

The following two operations need to be performed to determine the three-dimensional positions of points on an object using a pair of pictures of the same scene.

1. Determine the point pairs in two images which correspond to images of the same object point. This has been called the correspondence problem [5]. Note that the image of a small area on the object may be different in the two images because of changes in perspective and changes in surface reflectivity with the imaging angles. The translation distance between a region in one image and the corresponding region in the other image is known as the “disparity” of the region for a given stereo pair.

2. Determine the three-dimensional positions of points whose images have been identified in the two pictures. These positions can be determined by simple triangulation, as shown in Fig. 1, if the relative positions and orientations of the cameras for the two views are known.

Use of stereo has received some attention recently, with emphasis on determining corresponding regions in two images efficiently and reliably [6–9]. The common approach has been to use two images of a scene, separated by a chosen angular distance. The error in depth measurement depends directly on the choice of this angle (see Appendix). A larger angle will give more accurate information, but the disparities increase proportionately. The increase in disparities results in larger searches for determining correspondences, and increased likelihood of errors.
DEPTH MEASUREMENT BY MOTION STEREO

It is suggested that if intermediate views between two stereo images are used, searching of corresponding regions can be conducted more efficiently and more reliably. This issue is discussed in detail and some results are presented later in this paper. The work described here differs from the other stereo implementations in use of such intermediate views.

A progression of views of a scene is available if the observer is moving with respect to the environment, hence the use of the term "motion stereo." This occurs naturally for a mobile robot updating its model of the environment as it moves or in aerial photographs taken at successive periods in the flight of an aircraft.

The experiments described in this paper were conducted at the Stanford Artificial Intelligence Laboratory Hand/Eye Project. The objects were placed on a turntable which can be rotated in precisely determined increments. A progression of views was obtained by rotating the turntable (in increments of \( \frac{1}{3} \)° each), with the camera being held stationary. This procedure is equivalent to the camera being moved around the objects. The total separation between the extreme views varied between 5° and 10°.

The imaging device used is a vidicon camera, with a 2-in. focal length lens, and a 0.8 × 0.6 in. imaging surface. An image covering the entire imaging surface is digitized into 333 × 256 point samples. The brightness levels for each sample range between 0 and 63 units.

The orientation and position of the camera with respect to the table is known precisely by a calibration procedure [10]. For a different position of the turntable, a new equivalent camera calibration is easily computed. Note that this setup allows accurate information about relative camera positions for different views and the triangulation problem becomes trivial. For an actual moving observer, the camera transforms may be difficult to obtain and could be a major source of errors. Some work has been performed on determining relative camera positions from the stereo disparities, making some assumptions about the scene [7].

3. CORRESPONDENCE MEASURES

The following are two major approaches to finding corresponding areas in stereo images.

(a) Use of Features

"Interesting" features are located in one image and the other image is searched for corresponding features. This approach is potentially efficient as only few correspondences need be determined, but may be limited to situations where invariant features exist. In scenes of polyhedral objects, vertices remain invariant over a range of views. Useful, invariant features for more general scenes are not known. In particular, edge elements at the extremities of smooth surfaced objects will change in different views. This approach has not been used for the experiments described here.
Correlation of Regions

This approach is to find "similar" regions in two images, without an analysis of the contents of these regions. The choice of regions is based on the likelihood of finding corresponding regions for them, and not necessarily on their expected importance for the analysis of the scene.

Similarity measures need to be defined for evaluating the correspondence of two image regions. The ideal comparison is to determine whether the two sub-images are formed by the same part of the surface of an object. However, this is difficult and similarity of the two-dimensional images is measured instead. These measures do not account for perspective changes.

In the following, similarity measures are defined on two regions of the same shape and size; rectangular windows have been used in the experiments described. \( X_{ij} \) and \( Y_{ij} \) are used to denote the pixels at row \( i \) and column \( j \), measured relative to chosen points in the two image windows to be compared.

(i) Cross-Correlation

The cross-correlation coefficient of two picture windows is defined to be

\[
\alpha = \frac{E(X_{ij}Y_{ij}) - E(X_{ij})E(Y_{ij})}{S_d(X_{ij})S_d(Y_{ij})}
\]

where \( E \) stands for the average value over the window and \( S_d \) is the standard deviation.

The value of the cross-correlation coefficient, \( \alpha \), is unity for a perfect match. This measure is insensitive to absolute brightness levels and contrast in the two windows, and is useful if the two views were obtained under different illuminations [6].

(ii) Mean Square Difference

Normalized mean square difference is defined as

\[
\text{M.S.D.} = \frac{\sum (X_{ij} - Y_{ij})^2}{\sum X_{ij} \sum Y_{ij}}.
\]

The sums are computed over the entire regions being matched. This measure is sensitive to scaling of brightness levels and is useful only if the images to be compared are obtained under similar illumination, and surface reflectivities do not change rapidly with the viewing angle. This measure was found to be adequate for the experiments described here and requires less computation than the cross-correlation coefficient.

4. SEARCH STRATEGIES

The problem of finding correspondences is equivalent to finding best (most similar) matches for picture windows in one image with picture windows in another image. An exhaustive search for the best matches that compares each
window in one image to each window in a second image is clearly prohibitive, particularly since each comparison itself is expensive. Some strategies to limit the search space are described below.

*Search Along a Line*

A point in one image defines a ray along which the corresponding object point must lie. The image of this ray in the other view of a stereo pair is a straight line. Thus, the search for correspondence with the image of a point need only be along a line. To account for errors in the knowledge of the relative orientations and positions of the two cameras, the search needs to be conducted in a narrow band. It is assumed that the depth of points in a chosen window is nearly the same and that the above observation applies to the entire window.

The search distance along the line can be limited if the limits on the range to the object from the camera are known. Further improvements can be obtained by searching on a coarse grid, followed by a local “hill climbing” procedure. Variations of these methods are described in [6, 7]. Savings obtained by such methods are at a potential cost of missing the global optimum, as the form of the function being optimized is not known.

Barnea and Silverman used an interesting variation which does not require the complete computation of similarity measure for each potential match [11]. Sums of the differences are computed for one point of the window at a time, and those matches where the differences build up fast are not considered further.

*Use of Progressive Views*

Availability of several views offers further alternatives for searching for correspondences. Assume that the search for disparities between the two extreme views of a series of \( k \) views can be limited to a band \( n \) pixels long and \( m \) pixels wide, as discussed above. The search for correspondences between two adjacent views can then be limited to a band \( n' \) pixels long and \( m' \) pixels wide, where \( n' = (n/k - 1) \) and \( m' = (m/k - 1) \). Let the successive views be called view 1, view 2, ..., view \( k \).

Disparities between extreme views can be determined by chaining through the intermediate views. Consider the matching of a particular, chosen region in view 1. Assume that a match has been found for this region between view 1 and view \( i \). To determine the best match for the same region in view 1 with a region in view \( (i + 1) \), the search need be conducted only in an \( n' \times m' \) band, centered at the center of the region of best match in the previous view (view \( i \)). This process begins with matching of all regions of interest in view 1 with regions in view 2 and continues until the last view has been considered.

An alternative is to match views \( i \) and \( i + 1 \) at each step, where \( 1 < i \leq k - 1 \), and sum the disparities. However, for the former method, only crude disparities need to be found at all but the last step and the errors of matching do not accumulate; this is the method used for experiments reported in Section 6. Only integer-valued disparities are computed at all but the last step, where a real valued number is obtained by parabolic interpolation along the directions of the two axes.
Assuming exhaustive searches in the specified bands, the total number of matches examined for all steps will be \((k - 1) \times n'm'\) which is equal to \((nm)/(k - 1)\). This is a saving of a factor of \((1/k - 1)\) compared to the direct search between the extreme views. For example, by using 11 intermediate frames, a saving of a factor of 10 can be obtained.

Actually, the savings realized are not so large, as the width of the band to be searched is not directly proportional to the angle between the two views. If this width were independent of the angle, the search times would be the same. However, even in this case, the use of intermediate views has a strong advantage of increased reliability. Two different regions in one view are sometimes similar to the same region in another view. Choosing one region based solely on marginal differences in the similarity measure will occasionally lead to the wrong choice. Using intermediate views as described here, the search at each step is confined to a small area, less than half the size of the chosen window in our examples, and the likelihood of finding an erroneous similar region in another part of the image is reduced substantially.

Fig 2. Stereo views of a rock and two examples of computed correspondences.
Use of Context

Search for correspondence to a region may be speeded up by using disparities of neighboring regions. Large regions of a scene are often continuous in depth and hence the disparities vary continuously. Hannah has made some use of such context [12]. In motion stereo, the information obtained from a subset of views could be used to predict features in the succeeding views. Efficient use of context in stereo remains an important problem for further research.

5. RESULTS

This section describes experimental results using progressive views obtained by moving objects on a turntable, as described in Section 2. In all examples, two successive views were separated by a 0.5° rotation of the turntable.

An image was decomposed into nonoverlapping regions of a fixed size (8 X 8 or 16 X 16 pixels). Only those regions whose variances exceeded a fixed threshold
were examined for correspondences. The search strategy was as described in Section 4, under Use of Progressive Views, using normalized mean square difference as the similarity measure. (The results using cross-correlation were similar, except for longer run times.) The search for correspondences between two views, $\frac{1}{2}^\circ$ apart, could be confined to a $4 \times 8$ pixel window.

Figures 2a and 2b show two views of a rock, $5^\circ$ apart (9 intermediate views, not shown in the figure, were also used). Small ($8 \times 8$ pixels) windows, outlined in white, show two examples of the corresponding regions discovered by the program in two views. Figures 3a and 3b show similar results for a cup (an ordinary styrofoam coffee cup covered with textured paper). Here the two views are $6.5^\circ$ apart and the matching window size is $16 \times 16$ pixels.

The example of the cup is used to test the accuracy of stereo depth measurements. For the cup placed upright on a horizontal table, the contours of points on the surface of the cup at fixed heights above the table are circular. Points in Fig. 4 are obtained by projecting the computed three-dimensional points on the surface of the cup onto the horizontal table. Points lying in different ranges of height above the table are grouped together. Each height range is 0.5 in. wide; the cup being approximately 4 in. tall. A least mean square circular arc is fit to the points in each group. The quality of stereo measurements can now be judged by the departure of these points from the circular arc (the effect of including points in a height range, instead of just at a single height, is small compared to other errors discussed below). The radii of the arc are of less importance, as they are strongly affected by the accuracy of camera calibration. Further, the centers of the circles are ill-defined, as only a small fraction of the circle is used for the least mean square fit.

For these experiments, the average error of the points from the fitted circular arcs was 0.075 in. To compute the expected error, assume the maximum error due to spatial quantization of the image to be one half of a pixel. Assume the root mean square error to be one half of this value. For the cup at a distance of 40 in. from the camera, and for a stereo separation of $6.5^\circ$, an error of 0.25 pixels in determining correspondences corresponds to an error of 0.1 in. in range measurement (see Appendix). This is in excellent agreement with the experimental results. Further, the computed radii of the circular contours were in good agreement with the actual radii.

These results indicate that the methods described here are useful for measuring depth to an accuracy of 2.5 mm at a distance of 1 m, with the resolution of the imaging device used (as described in Section 2).

The average computation time for correlating a $16 \times 16$ pixel region over $6.5^\circ$ of stereo views is nearly 10 sec (on a PDP-10, KA-10 processor). The cup images had about 50 interesting regions and required a total of 8 min to process. The issue of computation time is further discussed in the next section.

6. SOME PROBLEMS

A major difficulty of stereo depth measurement is that certain parts of an object do not appear in both (or all) stereo views, because of different occlusions.
Of course, the depth of such regions cannot be measured by stereo. The difficulty is in discovering that a particular region does not match with any other in a stereo pair and preventing a false match. The numerical value of the similarity measure (such as the correlation coefficient) is not always adequate for distinguishing a false match, i.e., some false region matches result in higher similarity measures than that for the match of two other regions that are in fact matched correctly. Note that we are still asserting that the similarity measure works well if a correct match does exist.

Unfortunately, the regions of an image most likely to be subjected to such occlusions are those near the edges of an object. These are precisely the regions of most interest for segmentation of the scene into objects. As an example, Fig. 5 shows two views of a scene of two rocks (5° apart). Figure 6 shows a map generated by projecting the computed points on the surface of the rocks in three...
dimensions onto the horizontal plane. Though the points group into two clusters as expected, some points fall between the two clusters, making the segmentation more difficult. Hannah [7] presents some methods for detecting regions that have no matches, using two views. Use of intermediate views may make the detection of such regions easier as parts of an object gradually disappear from the camera view.

A related problem is to account for occasional, erroneous matches. The human visual system is believed to use global correspondences to correct for errors of local correspondences. Julesz [12] has suggested a "cooperative" stereo model for the human system. No attempt has been made at the solution of these problems in this work.

Excessive computation time can be a serious problem for any real-time application of a stereo system. Currently available faster hardware processors (e.g., a DEC PDP 11/45) can reduce the execution times by a factor of 5. Computations using most of the execution time are simple arithmetic operations on small integers, and hence special purpose, inexpensive array processors, e.g., a Signal Processing Systems Model SPS-41, may be used. Also, processing of each region in an image is independent of the other regions and could be performed using multiple parallel processors.

7. APPLICATIONS AND CONCLUSION

Feasibility of motion stereo has been demonstrated here. Motion stereo is more attractive than conventional two-camera stereo, when a series of progres-
DEPTH MEASUREMENT BY MOTION STEREO

sive views is conveniently available, because of the smaller likelihood of gross errors in the former case. Strategies for finding correspondences in stereo images are still in an evolutionary phase, and it is difficult to estimate the computational efficiency of motion stereo compared to two-image stereo. At least for one form of searching, motion stereo is more efficient (Section 4). Multiple views are naturally available for the environment of a moving robot (such as a future Mars rover). Such views are also conveniently available for many indoor, industrial applications. Parts to be assembled or inspected frequently come down on a slow moving conveyor belt, providing a succession of views useful for motion stereo. Aerial (or orbital) photography is another instance where progressive views are readily available.

Further research into the problems of occlusion and speedup of computation times is necessary before the use of these techniques in a practical system. However, even a limited amount of stereo capability used in conjunction with techniques of analysis of monocular images is likely to make the task of machine perception of three-dimensional scenes much easier.

APPENDIX: ERROR ANALYSIS

The error in range measurement of a point corresponding to an error in disparity measurement is derived here. For simplicity, consider a point on the optical axes of the two views. Let \( b \) be the base line or the distance between the camera centers of the two views; let \( R \) be the distance of the point from the lens centers; and let \( \theta \) be the angle subtended by \( b \) at this point (see Fig. 7). From simple geometry,

\[
b = 2R \sin \left( \frac{\theta}{2} \right).
\]

For small \( \theta \), \( R = b/\theta \), With fixed \( b \), \( |dR/d\theta| = (R/\theta) \) or,

\[
\Delta R = R \times |\Delta \theta| / \theta. \tag{1}
\]

This analysis strictly applies for points on the camera axes only, but is a good approximation for other points whose range \( R \) is large compared to \( b \). For an error of one pixel in disparity measurement, \( \Delta \theta \) is 0.001212 radians, for our imaging system described in Section 2. For objects at a distance of 40 in. and a stereo angle of 6.5°, Eq. (1) gives an error, \( \Delta R \), of 0.4 in. For a 0.25 pixel error, the estimated error is 0.1 in.

ACKNOWLEDGMENTS

The author wishes to thank Dr. Thomas O. Binford of Stanford University for suggesting the problem of motion stereo and to Dr. Lynn Quam for the use.
of his display programs. This research was supported by Advanced Research Projects Agency of the Department of Defense under Contracts SD-183 and F0606-72-C-0008.

REFERENCES