High Accuracy Depth Measurement using Multi-view Stereo

Trina D. Russ and Anthony P. Reeves
School of Electrical Engineering
Cornell University
Ithaca, New York 14850
tdr3@cornell.edu

Abstract

A novel scheme for depth extraction is achieved using a multiple view ring camera system. The ring camera method captures a series of images of scene from a set of camera locations arranged in a circular ring. Tracking of scene features through this sequence realizes circular feature trajectories. The recovery of depth can be obtained with this method by determining the diameter of the circular trajectory. Depth obtained using the ring camera strategy is shown to be more accurate and more robust than binocular methods. In addition, associated with this method is a trajectory confidence measure which provides a good and reliable indication of depth accuracy.

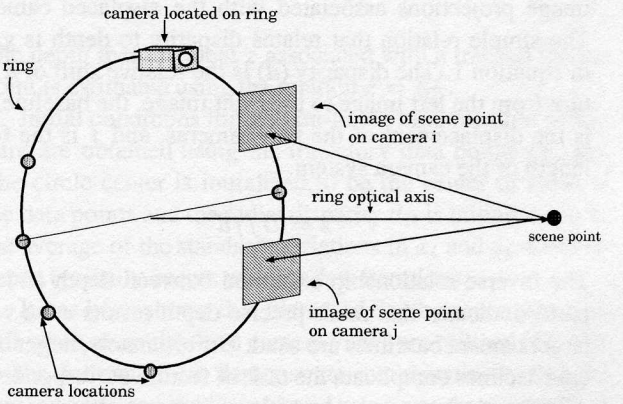

Figure 1: The Ring Camera Experimental Organization

1 Introduction

Depth maps from stereopsis methods that are both accurate and dense are difficult to achieve because of three fundamental problems associated with the depth extraction process: (a) feature correspondence, (b) occlusion, and (c) non-constant image brightness[3]. In this paper, we address the problem of accurate feature correspondence using a multiple baseline stereo system. The multiple baseline system allows the tracking of image features through a sequence of images captured at closely located positions, thus, enabling the use of moderate to large baselines without introducing additional ambiguity in the feature correspondence process. In addition, the utilization of multiple views typically provides a system which is more robust to image noise and less sensitive to occlusion.

In contrast to typical multiple baseline systems which simply displace a camera laterally [11, 13], we introduce the ring camera method for accurate depth measurement which acquires a sequence of images captured from locations arranged in a circular trajectory. Selected tracked features trace a circular trajectory whose diameter directly corresponds to binocular disparity. This work differs from previous research in multi-view depth measurement systems in two main characteristics: (a) the use of a geometric based confidence measure and (b) the large number of camera views used (more than 10). The geometric based confidence measure, based on multiple camera locations, is used to evaluate the confidence of the resulting depth measurement. Utilization of a large number of camera views permits robust accurate statistical depth measurement methods to be employed and increases the sensitivity of the confidence measure. Additional notable features of the ring camera method include the ability for anisotropic depth measurement, comparison of different feature types using depth confidence, and utilization of camera views placed along two dimensions.

For this system depth is reconstructed from a set of images without motion; obtained using either a moving camera and static scene or an array camera. In addition, we assume that the camera system has an optical axis that is orthogonal to the image plane and that the optical axis is fixed for all camera locations. The organization of the ring camera is shown in Figure 1. In this paper results are presented to demonstrate several advantages of this system which include: increased accuracy in comparison to binocular systems, increased robustness to image noise in comparison to a two view system, and a geometric based confidence measure which reliably determines depth uncertainty.

The remainder of this paper is organized as follows. First, closely related multiple viewpoint depth extraction methods are briefly discussed. Second, the ring camera method for measuring depth is described. Finally, results are presented to demonstrate the accuracy, robustness, and performance of the geometric confidence measure in the ring camera system.

2 Previous Work

In binocular systems with two horizontally displaced cameras, depth is typically recovered from the two-dimensional image projections associated with the displaced cameras. The simple relation that relates disparity to depth is given in equation 1. The disparity (d) is the relative shift of a feature from the left image to the right image, the baseline (B) is the displacement of the two cameras, and f is the focal length of the camera system.

$$z = Bf/d \tag{1}$$

The inverse relationship exhibited between depth and disparity indicates that more precise depth results arise when larger camera baselines are used. Unfortunately, larger camera baselines complicate the task of feature correspondence. Additional phenomenon which complicate the correspondence problem include lack of image texture, regularly repeated image patterns, occlusion, and photometric distortions/lighting affects. To address these problems a number of approaches have been investigated in the literature. For instance, integration approaches have been studied where the process of feature matching is integrated with various other stages of the stereo process such as surface interpolation [6, 15, 9]. Area-based methods have sought better correspondences by investigating various techniques for selecting an appropriate window size for a given feature [8, 4, 10]. In addition, multiple view systems have been employed to facilitate feature correspondence which have displaced a single monocular camera [11, 13, 5, 1] or utilized camera array systems [2, 7, 17, 12, 16]. The method presented here involves a feature tracking method using multiple sequential views.

Feature tracking methodologies are generally employed to sequences generated by the motion of a single monocular camera. The most common situation involves the lateral displacement of a single camera to generate a sequence of views. These systems typically generate a number of disparity and baseline estimates (d_i, B_i) associated with a single feature which are integrated to determine the feature's depth. Matthies, Kanade, and Szeliski [11] and Hanmandlu, Shataram, and Sudheer [5] integrate different depth estimates using a Kalman filtering strategy. In [13] Okutomi and Kanade perform feature matching using a sum-of-squared differences (SSD) operator with respect to inverse distance (disparity) which exhibits a unique matching minimum when the SSD results from the multiple stereo pairs are summed. Baker, Bolles, and Marimount [1] have implemented a single camera system which allows linear motion and reconstructs depth utilizing epipolar geometry and Kalman filtering.

Array camera systems which use multiple cameras have also been introduced to improve the tradeoff between estimation accuracy and matching difficulty. Kanade et. al. [7] use the model presented in [13] to design a five camera video-rate multiple baseline array camera system with vertical and horizontal camera translations. Nakamura and Matsuura et. al. [12] use a 3×3 array of cameras to resolve occlusion by introducing occlusion masks which represent occlusion patterns in real scene. Cox et. al. [2] develop a maximum likelihood formulation of the N-camera stereo problem. Zitnick and Webb [17] introduce a system of four cameras which are horizontally displaced and analyzes potential 3D surfaces to resolve the feature matching problem. Tsai [16] introduces two similarity metrics in an array camera system to obtain stronger evidence (peaks) for correct disparity measurement, thereby providing increased robustness to image noise.

The ring camera approach is a multi-view system that is designed to improve depth accuracy by exploiting the placement geometry of a larger number of camera positions and a geometric based confidence measure. Most depth extraction systems focus on known controlled motion in one direction, while few have considered using a large number of camera positions in both the X and Y directions. To investigate the ring camera we have implemented a simple prototype system which involves moving a camera in a circular trajectory and reconstructing the depth from the sequence of images.

3 The Ring Camera

The depth estimation algorithm used in this paper requires the acquisition of a sequence of images captured at equidistant locations around a circle. For small distances between image capture positions feature correspondence is facilitated by tracking features through the sequence of images. Each tracked feature traces out a circular trajectory through the sequence which provides a direct relationship to disparity calculation. In Figure 2 images acquired from two views of the ring camera system which correspond to the left and right camera positions for a typical binocular system are shown. Overlayed on these two images are the circular trajectories obtained from tracking three chosen features with our ring camera system. The trajectories in Figure 2 are obtained from a sequence of N images captured at locations θ_i for $i = 1 \ldots N$. Camera location θ_i and θ_{i+1} are separated by an angle increment of $\Delta\theta = 360°/N$. For the ring camera system the *radial disparity* (d_R) is defined as the diameter

of the best fit circle to the feature trajectory. In Figure 2, it is denoted by $d1$, $d2$, and $d3$. The radial disparity corresponds directly to binocular disparity when the degenerate case of just two views obtained from camera locations θ and $\theta + 180°$ is considered.

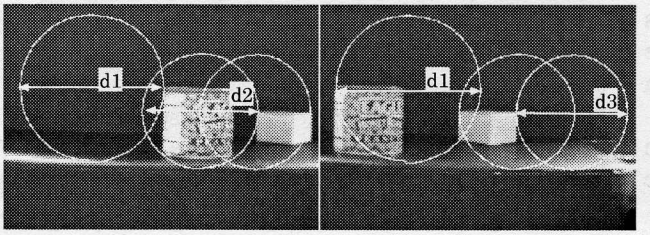

Figure 2: Shows a stereo right and left camera view with the ring trajectory super-imposed

The ideal trajectory for correctly tracked image locations is a circle. The deviation from a circular path for a given trajectory can be attributed to several sources of error. The major sources of error considered include camera positioning error (x^c, y^c), and tracking alignment error (x^t, y^t). In the system, the camera is positioned by hand to pre-marked locations, thus camera positions may not perfectly fit a circle and result in camera positional error. Non-constant image brightness, lens imperfections, quantization effects, and tracking algorithm limitations indicate that the tracking of a feature region will not be precise and thus, a small amount of error in tracking is accounted for. For this system, tracking is achieved using a standard sum-of-squared differences algorithm. A feature is matched by minimizing the error associated with an image shift (x_i^d, y_i^d) that lies within a given search region. The error function is given by

$$e(\tilde{x}, \tilde{y}_i; x_i^d, y_i^d) = \sum_{j,k \in W} [I_1(\tilde{x}_1 + j, \tilde{y}_1 + k) - I_i(\tilde{x}_{i-1} + j + x_i^d, \tilde{y}_{i-1} + k + y_i^d)]^2 \quad (2)$$

and the ith trajectory point is given by

$$\begin{pmatrix} \tilde{x}_i \\ \tilde{y}_i \end{pmatrix} = \begin{pmatrix} \tilde{x}_{i-1} \\ \tilde{y}_{i-1} \end{pmatrix} + \begin{pmatrix} x_i^d \\ y_i^d \end{pmatrix}. \quad (3)$$

The measured trajectory locations $(\tilde{x}_i, \tilde{y}_i)$ are related to the true locations (x_i, y_i) by equation 4 which accounts for system errors.

$$\begin{pmatrix} \tilde{x}_i \\ \tilde{y}_i \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} + \begin{pmatrix} x_i^c \\ y_i^c \end{pmatrix} + \begin{pmatrix} x_i^t \\ y_i^t \end{pmatrix} \quad \forall i \in \{1, \ldots, N\} \quad (4)$$

Assuming system errors have a Gaussian distribution, a least squares regression involving all views achieves a more precise estimate of the *radial disparity* (d_r) than can be achieved using only one stereo pair. Circle-fitting is performed using the Levenberg-Marquandt non-linear least squares algorithm [14]. Given N data points $(\tilde{x}_i, \tilde{y}_i)$ which approximate a circle, the radial disparity(circle diameter (d_R)) and circle center (x_c, y_c) are determined to minimize the following χ^2_{circle} error function where σ_i is the standard deviation of the ith data point. The standard deviation associated with the ith data point is currently unknown, thus σ_i is currently set to one so that all points are considered equal in the circular fit algorithm. We minimize the mean squared error which indicates the average squared deviation of a trajectory point from the best fit circle.

$$\chi^2_{circle} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\frac{1}{2} d_R - \sqrt{(\tilde{x}_i - x_c)^2 + (\tilde{y}_i - y_c)^2}}{\sigma_i} \right)^2 \quad (5)$$

The depth measurement z associated with a tracked feature point is estimated using the relation $z = \frac{Bf}{d_R}$.

Initial conditions for the non-linear minimization procedure are obtained using the trajectory data points $(\tilde{x}_i, \tilde{y}_i)$. The circle-center is initialized to be the center of mass of the data points and the radial disparity d_R is initialized to be the average of the standard deviations in $\tilde{x}_i$ and $\tilde{y}_i$. Convergence typically occurs within a few iterations.

It has been shown that increased depth accuracy can result when more than two camera views are utilized [13, 11]. In the following we demonstrate mathematically that increased accuracy can be achieved using a multi-view system formulated in a non-linear least squares framework.

Consider the variance of our error function,

$$\mathbf{Var}(\chi^2_{circle}) = \mathbf{Var}\left(\frac{1}{N} \sum_{i=1}^{N} \left(\frac{\frac{1}{2} d_R - \sqrt{(\tilde{x}_i - x_c)^2 + (\tilde{y}_i - y_c)^2}}{\sigma_i} \right)^2 \right). \quad (6)$$

Define the random variable W_i which is a function of the random variables $\tilde{x}_i$ and $\tilde{y}_i$ by

$$W_i = \left(\frac{\frac{1}{2} d_R - \sqrt{(\tilde{x}_i - x_c)^2 + (\tilde{y}_i - y_c)^2}}{\sigma_i} \right)^2 \quad (7)$$

The variance then becomes

$$\begin{aligned} \mathbf{Var}(\chi^2_{circle}) &= \mathbf{Var}\left(\frac{1}{N} \sum_{i=1}^{N} W_i \right) \\ &= \frac{1}{N^2} \mathbf{Var}\left(\sum_{i=1}^{N} W_i \right) \\ &= \frac{1}{N^2} N \mathbf{Var}(W_i) = \frac{\mathbf{Var}(W_i)}{N}. \end{aligned} \quad (8)$$

Therefore, the variance in our error function decreases as N increases which indicates improved disparity and depth estimation with a multi-view system.

4 Ring Camera Depth Accuracy

An experiment was conducted to measure the relationship between depth measurement accuracy and the number of camera views used. To avoid errors and issues associated with precise camera calibration, the experiment recovers the depth of a set of locations that lie on a planar surface. This allows the matching of recovered depth values to a plane which provides a calibration independent method for gaging depth accuracy. For example, depth is recovered by approximating the camera focal length f and using the well known equation $z = Bf/d_R$. Depth values are then fit to a surface using a planar least squares algorithm given accurately located x and y positions [14]. The chi-squared error function given below is minimized to determine the best parameters $\{a_1, a_2, a_3\}$ which approximate the data (x_i, y_i, z_i) where σ_i is set to 1, because the standard deviation of the data is unknown.

$$z(x, y, ; \mathbf{a}) = a_i x + a_2 y + a_3 \tag{9}$$

$$\chi^2_{plane} = \sum_{i=1}^{N} \left(\frac{z_i - a_1 x_i - a_2 y_i - a_3}{\sigma_i} \right)^2 \tag{10}$$

Depth accuracy is measured by taking the square-root of χ^2_{plane} divided by the number of features N (eqn. 11). Therefore, the accuracy corresponds to the average deviation of a given feature from the best fit plane in the units of centimeters.

$$Depth\ \ Variation\ \ = \sqrt{\frac{\chi^2_{plane}}{N}} \tag{11}$$

Image data was acquired at $N = 32$ equally spaced image locations and tracking was achieved using all 32 image locations to formulate a circular trajectory. The use of all camera locations in feature tracking avoids the confounding effect of the tracking algorithm and the "feature correspondence" problem, when fewer image locations are

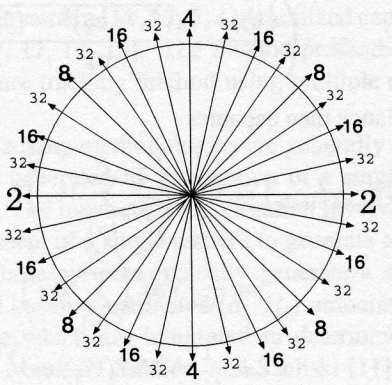

Figure 3: Camera Views

utilized for disparity measurement. A different number of camera locations was simulated by selecting an equally spaced subset of circular trajectory points and using these points in the non-linear circle fitting procedure. The positions of these camera locations are shown in Figure 3. In Figure 3, locations marked less than or equal to N are included in the experiment where N views are studied. In our experiments, we investigate the effect of using $\{2, 4, 8, 16, 32\}$ camera locations on depth estimation accuracy.

The trajectory depth uncertainty measure is based on how well a circle is fit to the feature trajectory data. The chi-squared error for the circle least squares fit χ^2_{circle} gives a notion of this, but it requires normalization so circles of different radii can be directly compared. The following normalization is used as an indication of depth uncertainty where N is the number of points in the disparity trajectory, and d_R is the diameter of the best fit circle.

$$Depth\ \ Confidence = \frac{\sqrt{\chi^2_{circle}/N}}{\frac{1}{2}d_R} \tag{12}$$

This measure gives the ratio of the mean circle error relative to the circle radius.

4.1 Data Acquisition and Scene Description

Data is acquired using a simple experimental set up which consists of attaching a Sony CCD video camera to a mechanism which mechanically constrained the motion of the camera to a single plane. An image sequence was obtained by manually displacing the camera in a circular trajectory. Scenes are created by placing objects on a stage located approximately 56cm from the camera and capturing images to be combined in a sequence. The ring camera baseline distance used for the experiments is $15.24cm$.

The planar scene consisted of a textured flat surface shown in Figure 4. The surface was inclined to the camera with respect to the x (horizontal) axis by an angle of approximately 2°. Depth measurements were made at 126 regularly spaced location in the leftmost camera location image shown by the white dots in Figure 4. The samples were organized in a regular grid pattern with a spacing of 7 pixels.

4.2 Accuracy With Respect to Number of Views

The features shown in Figure 4 were tracked using an image window size of 11 and a search region size of 21. The graph given in Figure 5 shows the planar depth variation (eqn. 11) for 126 depth measurements versus the number of camera views used in circle trajectory fitting. The graph shows that the variation in depth improves as the number of cameras used increases. An improvement of 62% is achieved by increasing the number of views from 2 to 32 views. Planar

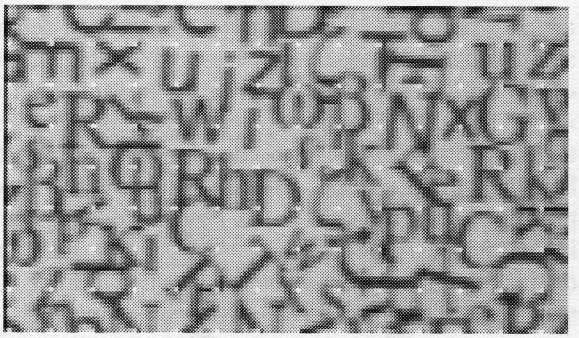

Figure 4: Scenes and Selected Features: Letter Scene with Tracked Features

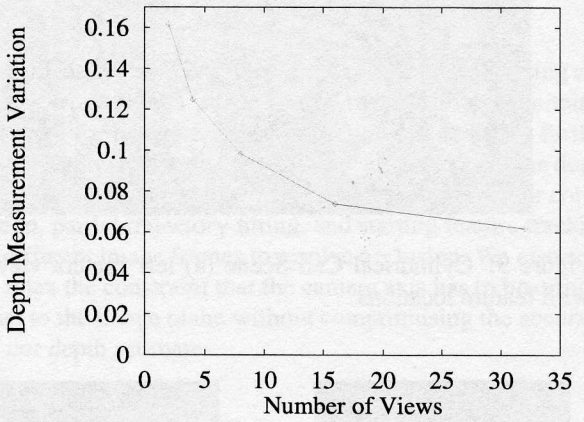

Figure 5: Depth Accuracy (cm) vs. Views

surface reconstructions using 2, 8, and 32 camera views are shown in Figure 6. These surfaces vary in intensity according to depth and give additional insight to the benefit of a multi-view system. When two views are used the reconstructed surface resembles a small step edge rather than a single planar surface. The introduction of additional views (8 and 32) allows us to see the true planar surface with a higher degree of accuracy. Therefore, additional views using the ring camera method produce more accurate depth estimates.

4.2.1 Accuracy Prediction from Depth Confidence

An important novel feature of the ring camera system is the circularity confidence measure which can be used to predict depth measurement accuracy. To demonstrate the ability of the confidence measure in predicting depth measurement accuracy the previous experiment is repeated, but with a smaller 7×7 tracking window. The error in surface planarity is very large ($\approx$ 6cm) because the tracking algorithm was unable to track some of the feature locations. The feature points are then ranked according to their confidence measure and the experiment is repeated using the most confident subsets of the tracked locations. Figure 7 shows the results of selecting the most confident 20%, 40%, 60%, and 80% of the features. The rejection of the least confident 20% of the features allows all the incorrect trajectories to be removed, thus, depth accuracy is comparable to the first experiment. Furthermore, as additional less confident features are removed consistency of the selected subset continues to improve.

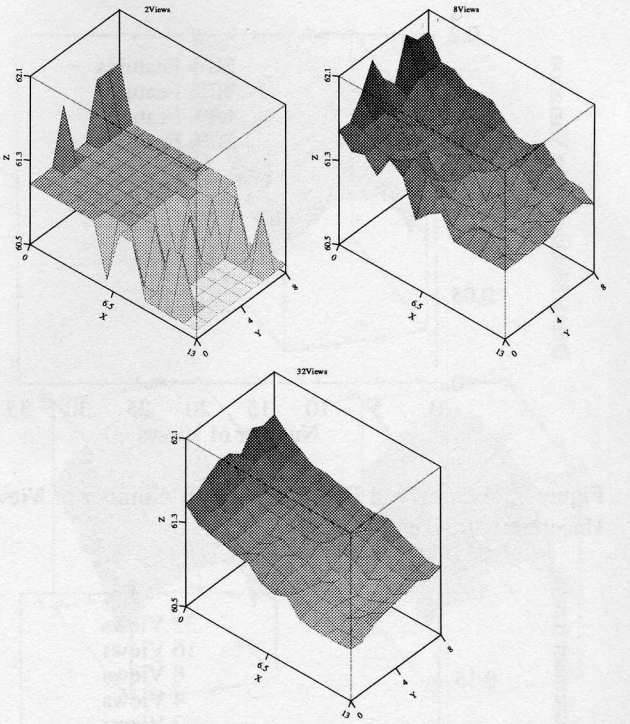

Figure 6: Planar Surface Reconstructions: (a) 2 Views (b) 8 Views (c) 32 Views

4.2.2 Robustness With Respect to Image Noise

The robustness of the system to noise is evaluated by adding different levels of random Gaussian noise to the letter scene shown in Figure 4 and repeating the experiment on the modified sequence with a window size of 17. Figure 8 shows the depth variation versus the standard deviation of Gaussian noise added to the sequence and the signal level of the image is approximately 100. The results in Figure 8 demonstrate that the system is able to achieve more accurate results than a 2 view system for the noise levels considered. Furthermore, in addition to being more accurate, using a larger number of camera locations results in a system that is less sensitive to noise. For example, the accuracy of the system changes very little when 32 views are utilized, while a notable change is accuracy is realized for two and four views.

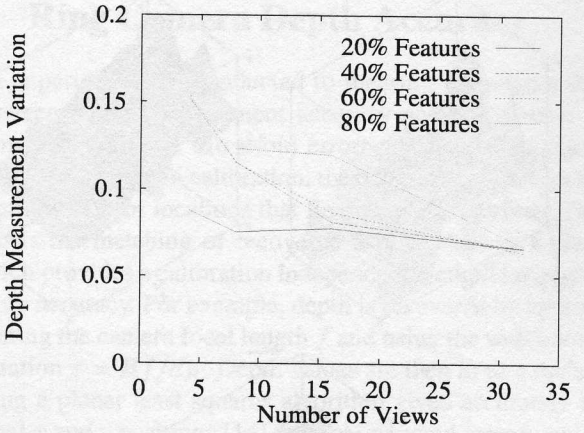

Figure 7: Accuracy in Depth cm versus Number of Views: Uncertainty Ranked Features

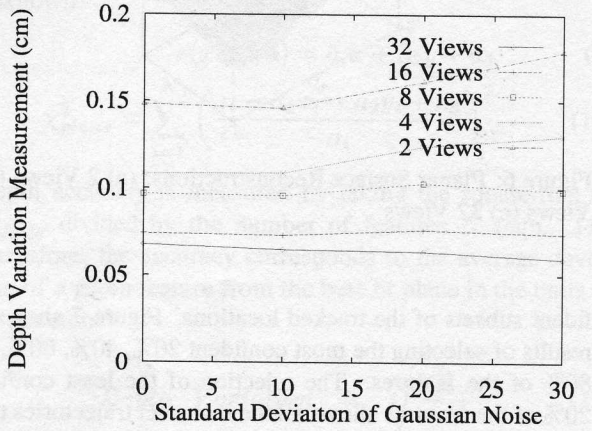

Figure 8: Accuracy vs. Noise Standard Deviation

5 Occlusion Boundary Detection and Curved Surfaces

The feature confidence measure of the ring camera may be used to identify problem regions within the scene. In traditional binocular stereo, confidence is typically determined by examining intensity variances to identify regions of insufficient texture or by employing the left-to-right consistency constraint to identify occlusion regions. For the ring camera, the confidence measure is used to identify problem areas. This is demonstrated for the smooth cylindrical object shown in Figure 9.

The cylindrical can used for this investigation is wrapped with highly textured paper to facilitate feature tracking. An image window size of 21 was used with a tracking search region of 31. The feature locations were set at a grid spacing of 10 pixels. Features were tracked starting from the 9 o'clock position and traversing in the counter clockwise direction. The reconstructed depth maps for 2 and 32 views are shown in Figure 10. A number of incorrect depth measurements may be observed at the external boundary region of the can. In Figure 11, the least confident depth values are set to black. Note these regions correspond, in general, to the incorrect pixels in the depth map and to occluding boundaries in the scene. The left boundary of the can is a problem because feature points are occluded during tracking. This does not occur from the right side of the can because of the ring cameras 9 o'clock starting location. Occlusion at the open top of the can also causes problems.

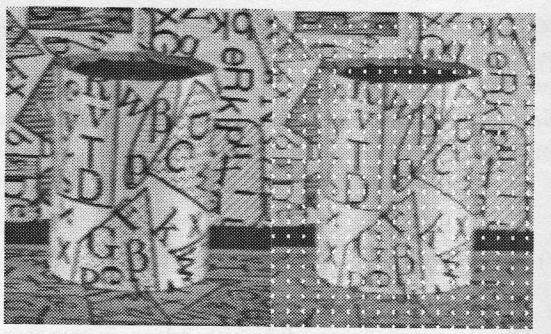

Figure 9: Cylindrical Can Scene (a) left camera view (b) with feature locations

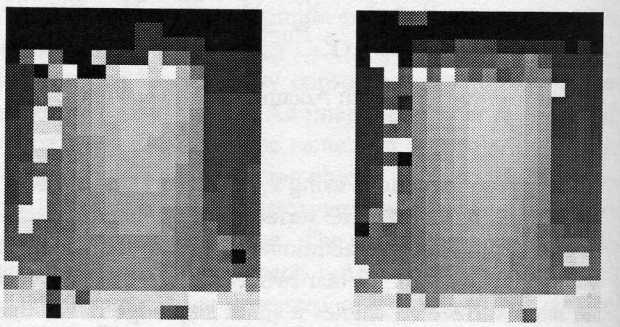

Figure 10: Depth Maps Generated for the Can Scene: (a) 2 Views (b) 32 Views

Figure 12 presents a surface description of the reconstructed can for 2, 8, and 32 camera views. Here we observe the improvement with which we are able to resolve curved surfaces compared to the binocular case.

6 Conclusion

In this paper we have presented a ring camera method for disparity measurement. The method has been shown to have better accuracy and robustness to image noise than two camera views. In addition, the disparity confidence measure is a good indication of depth accuracy and can be used to identify occlusion regions, depth discontinuities, and areas of

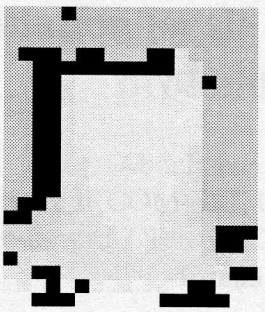

Figure 11: Depth map for 32 views with uncertain locations identified n black.

low intensity variation. The initial results are promising and future work entails utilization of the disparity uncertainty measure for adaptive feature selection. In addition, further investigation will seek to resolve uncertain areas of the depth map through exploration of different window sizes for correlation, partial trajectory fitting, and starting feature tracking at different image frames to resolve occlusion. We also seek to relax the constraint that the camera axis has to be orthogonal to the image plane without compromising the accuracy of our depth estimates.

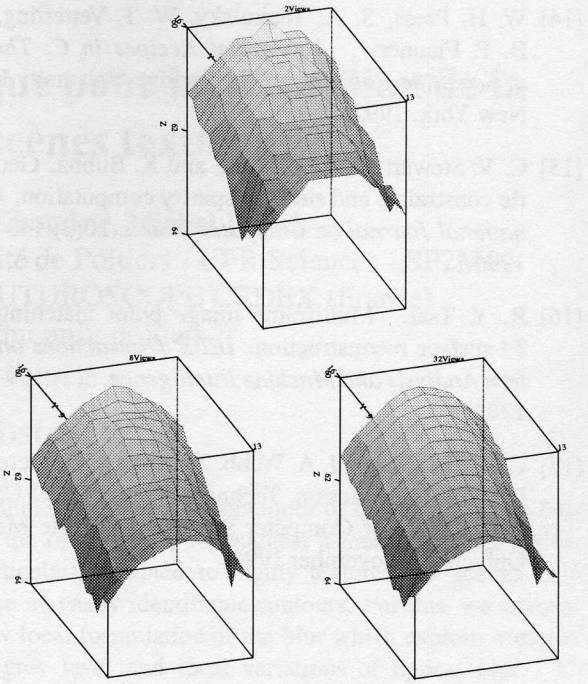

Figure 12: Curved depth surfaces for the front of the can: (a) 2 Views (b) 8 Views (c) 32 Views

References

[1] R. C. Bolles, H. H. Baker, and D. H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal on Computer Vision*, 1:7–55, 1987.

[2] I. J. Cox, S. L. Hingorani, S. B. Rao, and B. M. Maggs. A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding*, pages 542–567, 1996.

[3] U. R. Dhond and J. K. Aggarwal. Structure from stereo-a review. *IEEE Transactions on Systems, Man , and Cybernetics*, 19(6):1489–1510, 1989.

[4] A. Fusiello, V. Roberto, and E. Trucco. Efficient stereo with mulitple windowing. *International Conference on Computer Vision and Pattern Recognition*, pages 858–863, 1997.

[5] M. Hanmandlu, V. Shantaram, and K. Sudheer. Depth estimation from a sequence of images using spherical projection. *Internationl Conference on Robotics and Automation*, pages 2264–2269, 1997.

[6] W. Hoff and N. Ahuja. Surfaces from stereo: Integrating feature matching, disparity estimation and contour detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(2):121–136, 1989.

[7] T. Kanade, H. Kano, S. Kimura, A. Yoshida, and K. Oda. Development of a video-rate stereo machine. *International Conference on Intelligent Robots and Systems*, pages 95–100, 1995.

[8] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9):920–932, 1994.

[9] M. S. Lew, T. S. Huang, and K. Wong. Learning and feature selection in stereo maching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9):869–881, 1994.

[10] J. Lotti and G. Giraudon. Adaptive window algorithm for aerial image stereo. *International Conference on Pattern Recognition*, pages 701–703, 1994.

[11] L. Matthies, T. Kanade, and R. Szeliski. Kalman filter based algorithms for estimation depth from image sequences. *International Journal on Computer Vision*, pages 209–236, 1989.

[12] Y. Nakamura, T. Matsuura, K. Satoh, and Y. Ohta. Occlusion detectable stereo-occlusion patterns in camera matrix. *International Conference on Computer Vision and Pattern Recognition*, pages 371–378, 1996.

[13] M. Okutomi and T. Kanade. A mutliple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4), 1993.

[14] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery'. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York, 1992.

[15] C. V. Stewart, R. Y. Flatland, and K. Bubna. Geometric constraints and stereo disparity computation. *International Journal on Computer Vision*, 20(3):143–168, 1996.

[16] R. Y. Tsai. Multiframe image point matching and 3d surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):159–173, 1983.

[17] C. L. Zitnick and J. A. Webb. Multi-baseline stereo using surface extraction. Technical Report CMU-CS-96-196, School of Computer Science, Carnegie Melllon University, November 1996.