

ROBUST POSE DETERMINATION FOR AUTONOMOUS DOCKING

J. S. Goddard
Oak Ridge National Laboratory^a
Oak Ridge, TN 37831-6011
(615) 574-9034
E-mail: sgo@ornl.gov
Fax: (615) 574-6663

W. B. Jatko
ORNL
Oak Ridge, TN
(615) 574-5863
E-mail: emp@ornl.gov

R. K. Ferrell
ORNL
Oak Ridge, TN
(615) 574-5730
E-mail: rkf@ornl.gov

S. S. Gleason
ORNL
Oak Ridge, TN
(615) 574-8259
E-mail: sg6@ornl.gov

ABSTRACT

This paper describes current work at Oak Ridge National Laboratory to develop a robotic vision system capable of determining the position and orientation (pose) of designated objects from their geometry. A method has been developed that connects 2-D point features from a single image into higher-order shapes and then matches these features with corresponding features of a polyhedral model. Pose estimates are made from these matches using a closed-form point solution based on model features of four coplanar points. Rotations are represented by quaternions, which are four component numbers, that simplify the calculations in determining the least-squares solution for the coordinate transformation. This pose determination method, including image acquisition, feature extraction, feature correspondence, and pose calculation, has been implemented on a real-time system using a commercial camera and image processing hardware. Experimental results from this implementation are given for relative error measurements.

I. INTRODUCTION

A major advantage of robotic manipulators is their ability to function in hostile environments impractical or unsafe for humans. In many applications, such as bomb disposal or ordnance handling, the benefits of replacing humans with a mechanical counterpart are obvious. The

dexterity of mechanical hands to grasp and manipulate objects is quite advanced and improving rapidly. Object recognition and localization present one obstacle for future use of robots in unstructured hazardous environments. The ability to "find" a designated item is a subject of much continued research. It would be highly beneficial for the robot to locate an unknown object and determine its position and orientation (pose) in a world coordinate frame based solely on visual cues.

Current work at Oak Ridge National Laboratory to develop a robotic vision system capable of determining the pose of designated objects from their geometry is detailed in this paper. Valid pose estimates will be used to provide destination coordinates to a self-directed robotic arm. Both accuracy and robustness are required for robotic pose estimation where lighting and occlusion of objects are not closely controlled. An estimate of uncertainty in the measurement is also needed.

This system remotely measures the six degrees-of-freedom position and orientation of a target object with respect to a vision sensor. Point features from the target are extracted from a single image and then combined into higher-order geometric features for identification and correspondence with a polyhedral model using a newly developed method. A closed-form algorithm is applied to these geometric features for the actual pose calculation. Robustness and measurement quality estimates have been

^a Managed by Martin Marietta Energy Systems, Inc., for the U. S. Department of Energy under contract DE-AC05-84OR21400.

incorporated into the design to give good accuracy as well as to prevent the reporting of estimates with large position errors. A complete solution has been implemented including calibration and operator interface.

Section II of this paper presents a summary of previous work related to this area of pose measurement. A description of the problem and the theoretical background for the solution are given in Section III. Section IV describes the system implementation and Section V provides experimental results.

II. RELATED WORK

Much work has been published relating to single-camera position determination. Most of this work has used point features in the image to derive the position. Methods of this class were the first to be studied and as a result have been more extensively developed than model-based methods.¹ The most common assumption is the perspective projection of a 3-D object onto a 2-D image plane through a pinhole camera model.² Both single-image and stereo methods have been reported although single-image techniques have by far the greatest number of solutions. One reason is that point correspondence with an object from a single image is easier than determining correspondences between two images and the object, as required in stereo.

The general framework is, given N corresponding points in the object and in the image, to solve for the relative pose between the camera and the object. The minimum N to give a finite number of solutions is three. Under certain conditions, as many as four solutions are possible with three points. Four coplanar, noncollinear points give a unique solution. Four or five noncoplanar, noncollinear points give up to two solutions. For N greater than five noncollinear points, the result is unique and consists of an overdetermined set that can be solved using least-square methods.³ In general, as N increases, the accuracy of the results increases. These overdetermined solutions are also used for camera calibration in which internal camera parameters, including lens distortion and focal length, are measured along with the external parameters. One of the best known methods of this type is that given by Tsai.⁴ For direct pose measurement, however, three- and four-point coplanar targets have been more commonly used for pose determination. The National Aeronautics and Space Administration has described the use of a "T"-shaped target with three feature points that was used in spacecraft docking experiments.⁵ Fischler and Bolles describe a geometric solution with only three points.⁶ Also described is a method with four

coplanar points which provides a unique solution with redundancy. A fast, closed-form solution for a four-point coplanar target is given by Abidi and Chandra.³ Their approach does not require that the lens focal length be known as long as the plane of the target is not parallel to the image plane of the camera. This restriction is removed if the focal length is known. Redundant measurements are combined through averaging or by calculating the median. No explicit estimate of measurement error is provided. This method produces as its output the coordinates of the target features with respect to the camera reference. Horn gives a method for determining the least-square error transformation between two coordinate reference frames given the coordinates of N points in each frame.⁷

Quaternions are used to represent rotations and simplify the calculation of the least-square error transformation. Where one set of coordinates is known to high accuracy, as in a geometric model, the distance squared error between the model points and the transformed points may be used as an indication of measurement error.

A second category of solutions uses more complex geometric shapes and features to determine the pose. Known collectively as model-based methods, these approaches generally attempt to match features in an image with those of the geometric model. Lowe describes a solution in which a 3-D model is projected to 2-D and then matched with the image.⁸ In an iterative procedure, the position and orientation that give the optimum match between the features in the model and the image are calculated.

Much of this work treats the correspondence between points in the image and the model as a separate problem that is not addressed. The correspondence may be trivial for a small number of points in a plane. However, with a larger number of points on a 3-D object where some of the points may be obscured, the problem becomes much more difficult.

III. THEORY OF OPERATION

The problem of pose determination is shown in Figure 1 along with the separate 3-D reference frames for the target object and camera. A pinhole lens camera model is used where the lens center is the camera reference point. The desired result is the coordinate transformation between the camera and the object.

This transformation can be defined as a composition of three rotations and three translations and is denoted by a 4-by-4 matrix T ,

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{D} \\ \hline 0 & 0 & 0 & 1 \end{bmatrix}, \quad (1)$$

where \mathbf{R} is a 3-by-3 rotation matrix and \mathbf{D} is a 3-element displacement vector. The conversion between the target and the camera coordinates of a single point is a simple matrix multiplication,

$$\mathbf{X}_c = \mathbf{T}\mathbf{X}_o. \quad (2)$$

\mathbf{X}_c and \mathbf{X}_o are 4-element vectors consisting of the 3-D point coordinates in the form $(x \ y \ z \ 1)^T$. The decomposition of \mathbf{T} into the individual rotation and translation matrices is

$$\mathbf{T} = \mathbf{R}_r \mathbf{R}_p \mathbf{R}_t \mathbf{S}. \quad (3)$$

The roll, pan, and tilt rotation matrices represent rotations about the x , y , and z axes, respectively, while \mathbf{S} is the translation matrix. These individual rotation angles are not unique and, in general, are dependent on the order of multiplication.

From one intensity image of the target, the 2-D feature coordinates in the image plane can be extracted. These coordinates, along with the geometric model of the target and intrinsic camera parameters, provide the necessary information for pose calculation. Camera calibration performed off-line prior to the pose measurement is used to determine the intrinsic camera parameters such as pixel size and focal length. The relation between the 3-D camera coordinates and the 2-D image is based on the assumed perspective projection vision model.

These relations are, given image coordinates $(x_i \ y_i)$ and

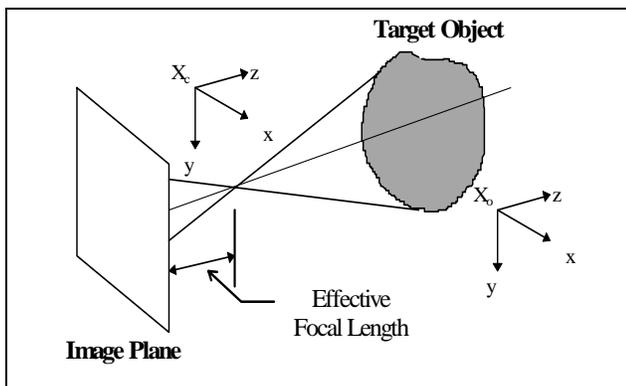


Figure 1. Perspective model for pose determination

camera coordinates $(x_c \ y_c \ z_c)$,

$$\begin{aligned} x_i &= x_c \frac{f_e}{z_c}, \\ y_i &= y_c \frac{f_e}{z_c}, \end{aligned} \quad (4)$$

where f_e is the effective focal length.⁹

While the calculation of the image coordinates is straightforward given the 3-D coordinates and the camera parameters, the inverse operation of determining the 3-D coordinates from the 2-D image plane coordinates is more difficult since the correspondence of 2-D to 3-D coordinates is not unique.

As discussed above, a minimum of four coplanar points in the image are needed to perform the inverse perspective mapping. However, for a single four-point target in a scene with other objects and clutter, the problem becomes one of identifying the target points and determining the point correspondence between the object model and the image feature points. This approach is not robust, however, in that the loss of a single point will prevent the pose calculation. Additional points are needed for redundancy. Also, the use of object points in a single plane depend entirely upon perspective to recover pose. This is not a problem as long as the object dimensions are comparable to the distance of the object from the camera. At longer distances, this perspective effect is reduced, and the calculation of pose becomes more sensitive to noise. As a result, pose measurement based on a model of a polyhedral object has been developed.

The model of the object is represented as surface faces, each of which is a convex quadrilateral with four vertices. Adjacent faces then share two vertices. From an image of the target object, points are extracted representing the vertices of the faces of the object that are visible from the camera. These image points are then matched to the target vertices based on convexity and adjacency constraints. The convexity constraint states that the image of a convex polygon under perspective projection remains a convex polygon. Likewise, adjacent polygons remain adjacent. For N target points, there are $N!$ possible correspondences to the image points. To reduce the computation required to test this large number of possible trials, feature points in the image are grouped into sets of convex quadrilaterals. From these sets, candidate sets are further selected that match the adjacency of faces in the geometric model. The image points are matched first to the model faces and then to the vertices within each face. These individual face points are then used for the pose

computation. Extraneous points at random locations that are not part of the target may also be detected. The extraneous points will not, in general, provide a geometric fit to the model and will be rejected by the geometric constraints. However, it is possible for an erroneous matching set to be generated. Error checking performed during the subsequent pose calculation will then flag the mismatch.

After the point correspondence is made, the pose is calculated separately for each identified face. A direct algorithm developed by Abidi and Chandra is used that determines a unique measurement for the four coplanar points in each face, assuming a perspective vision model.³ This algorithm calculates the 3-D coordinates of the points relative to the camera using only the image point locations and the dimensions of the face points from the target model. The next step is to determine the transformation matrix \mathbf{T} . While the pose calculation method is exact, errors in determining the image feature point locations due to noise or other factors will prevent the exact determination of \mathbf{T} . This error can be represented as

$$e_k = X_{c,k} - sRX_{o,k} - D \quad (5)$$

for each point k , where s is a scale factor and \mathbf{R} and \mathbf{D} are the rotation matrix and displacement vector respectively. The optimum transformation is found when the sum of the squares of the errors is minimized. Horn shows that the optimum translation is

$$D_{opt} = \frac{1}{n} \sum_{k=1}^n X_{o,k} - sR \left(\frac{1}{n} \sum_{k=1}^n X_{c,k} \right), \quad (6)$$

which is the difference between the centroid of the camera coordinates and the scaled and rotated centroid of the object coordinates.⁷ Quaternions, which can be thought of as complex numbers with three imaginary parts, are used to represent rotations. Horn further shows that the unit quaternion \hat{q} that maximizes

$$\sum_{k=1}^n \left(\hat{q} \overset{\circ}{X}_{o,k} \hat{q}^* \right) \cdot \overset{\circ}{X}_{c,k} \quad (7)$$

represents the optimum rotation. The solution is an eigenvector problem requiring the solution of a quartic polynomial equation for the eigenvalues and then finding the maximum eigenvector.

The sum of squared errors is used as an error measure to indicate the difference in geometric shape between the calculated points and the transformed model face points. An error value above an expected level due to random noise is an indicator of an incorrect point assignment as well as a measurement quality indicator.

IV. IMPLEMENTATION

Camera Calibration

To extract an accurate 3-D position from 2-D image coordinates, the camera's internal geometric and optical characteristics must be known. The parameters needed for pose computation are effective focal length, pixel size, lens distortion, and optical center. Camera calibration computes a camera's intrinsic and extrinsic parameters based on some number of points whose world coordinates (x_w, y_w, z_w) are known and whose image coordinates (x_i, y_i) are measured. To perform the calibration, a specialized target with a large number of precisely spaced points is imaged. Next, the coordinates of the target points in the image must be found to subpixel accuracy. These image coordinates are correlated with the world position of those target points. This information is then provided to the camera calibration routine, which calculates the camera parameters based on the target data and the camera model. Once the image points have been correlated with world coordinate points, these data are provided to the camera calibration program, and the focal length, radial distortion, and the optical center are calculated for the camera and lens arrangement. This calibration process uses a well-known model for camera calibration developed by Tsai.⁴ A public domain software implementation of Tsai's method was adapted for use in this application. An automated procedure was developed to match the image points with the target points for use by the calibration software.

Pose Calculation

A complete system has been implemented to demonstrate and test the pose determination method described above in Section III. Figure 2 shows a block diagram of the principal hardware elements. The major items include a target object, a camera, and an image acquisition and processing computer. Six infrared (780 nm wavelength) light-emitting diodes (LEDs) are arranged at the corners of the front faces on the target.

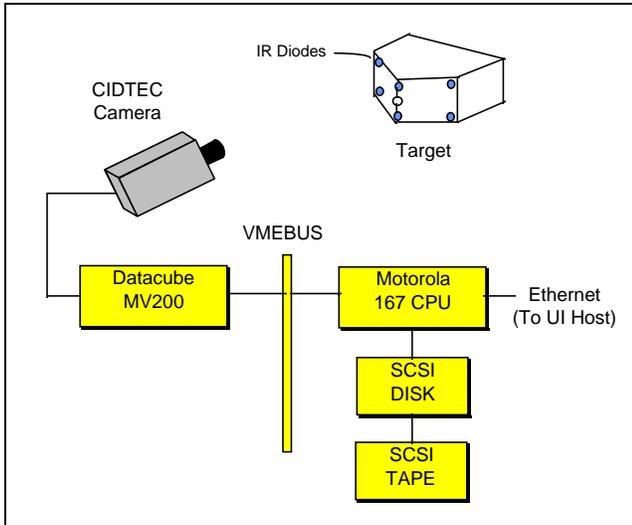


Figure 2. Hardware implementation

A commercial camera, a CIDTEC 2250a, is used for viewing the target. A 6-mm focal length lens is used for a wide-angle field of view with a 780 nm infrared pass filter to block out visible light. The image resolution is 512 by 512 square pixels at 30-frames/s progressive scan. The camera image is acquired by a Datacube MV200 image processing VME board. This board along with the main central processing unit (CPU), a Motorola MVME167, SCSI disk, and VME chassis, comprise the Datacube MaxTd development system. System software includes the Lynx real-time operating system, Motif X-windows, and Datacube imageflow software. Application software written in C controls the MV200 and programs the processing elements on the board for each pose calculation cycle.

Approximately 100 ms is required to perform the full pose calculation, which includes both point correspondence and pose calculations for two faces. An option is included in the program that can be selected from the main menu to display a live digital image from the camera. The result is a pattern of LEDs from the target. Overlaying this image are rectangles that outline the LED regions as detected by the program. These regions are identified and connected according to the geometry of the target as described above. Overlay lines show this connectivity as well as indicate that a valid pose has been calculated. A continuous update and display of the calculated pose with translation parameters in millimeters and angles in degrees are presented on the local screen.

The functional breakdown for the pose determination software is shown in Figure 3. During image acquisition

the MV200 digitizes the camera data to 8 bits per pixel and stores an entire frame in a local memory buffer. The data are then thresholded to segment the light points and are passed through a morphological filter that removes some artifacts from the image such as single pixel points and small holes within a larger object. Blob analysis is next performed, with the result that the centroid of each LED image is found to subpixel accuracy. Because the LED image is roughly elliptical, filtering is also used at this point to eliminate shapes arising from other sources. The centroid values from the remaining objects are then used to establish point correspondence and connectivity between points from the actual geometry of the 3-D target points.

Two faces are extracted from the six points of the target, and a pose calculation based on the four-point coplanar algorithm is subsequently made from each face. While these values could be combined using several different methods, the result giving the lowest error measure is reported as the pose estimate. This pose value is sent to the user interface computer as requested using a command-response protocol developed for this application based on Unix sockets. Some robustness is built into this implementation in that one light and some patterns of two lights may be obscured without preventing pose determination since one face can still be identified.

Graphical User Interface

A graphical user interface (GUI) was developed for a simulated docking demonstration in conjunction with the pose calculation software. It is implemented on a Sun SPARCstation 10/30 using X-windows (X11R5) protocol and Motif (Version 3.0) widgets. Several different techniques for graphically representing the pose information have been implemented on the dashboard simulation GUI. In autonomous docking modes, the GUI

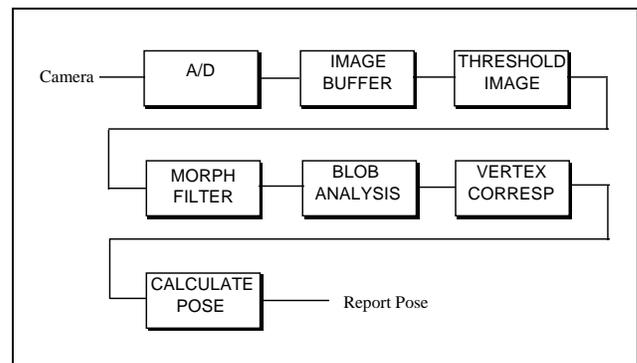


Figure 3. Software functional flow

display is used by the operator to monitor the path of a robotic arm. For manual docking modes, the GUI is intended to improve the operator's ability by providing a supplementary viewpoint. The challenge in this development is to provide an operator with information about six degrees of freedom on a 2-D screen. Position instruments and text displays were implemented on one screen.

V. EXPERIMENTAL RESULTS

A relative accuracy measurement of all six pose parameters has been performed for the implemented system. In all cases, the transformation matrix is calculated with respect to the camera lens center, which is the origin of the baseline reference frame and the target reference frame located at the center of the target. The transformation matrix is then decomposed into the six parameters. A test fixture for the camera and target permitted precise movement of the target independently in each translation and rotation direction. The target used is that shown in Figure 2. Overall target dimensions are approximately three inches high, seven inches wide, and five inches deep.

Initially the camera and target were aligned so that the x and y axes of both frames were parallel and the z axis of the target was collinear with the z axis of the camera. For the translation axes, the procedure was to move the target along one axis by 5 mm and by 15 mm, for various distances from the camera. The rotation angles remained at zero. For each movement approximately 100 pose measurement readings were taken. Similarly, movements of 5, 10, and 20 degrees were made for each rotation with corresponding measurements recorded. In the graphs that follow, the mean calculated values as a function of camera distance are shown along with error bars that give the maximum and minimum readings for each measurement set. Where no error bars are visible, the data variability was not large enough to be shown for the resolution of the graph. Figures 4 to 6 show the relative changes in x, y, and z measurements when the target is moved 5 and 15 mm in the translation axis. These changes are shown as a function of distance of the target from the camera. For both x and y, the changes are within about 5% of the actual movements over the entire range of camera distances. However, the z axis variations become much worse as the camera distance increases. At a distance up to 400 mm, the variation is comparable to the x and y axes. At larger distances, the mean errors become as large as 100%, while a larger variation is seen in the maximum and minimum values. These data indicate that small lateral movements are measured accurately even at relatively large distances while small distances along the optical axis of the camera

are accurate only when the variations are a significant percentage of the camera distance to the target. Figure 7 plots the percentage error in the absolute z-axis measurement as a function of distance from the camera using the closest z-axis measurement as a reference with zero error. This provides an estimate of absolute error in z, which is low compared to the relative errors shown in Figure 6.

Angular measurements are shown in Figures 8 to 10 for pan, tilt, and roll movements of 5, 10, and 20 degrees. These show more variation and mean error than the translation axes at larger camera distances. Similar accuracies, however, are obtained at closer distances. The translation axes reference movements were made using precision micrometer stages that are accurate within a few microns. Rotation axes reference movements, in contrast, are accurate only to about ± 0.5 degree. Some of the variation in rotation axis measurement versus actual movement can be attributed to this error component.

Other sources of error include camera and data acquisition noise, processing of the image to extract the LED centroids, calibration error, and uncertainty in the initial off-line measurement of the distances between LED centroids for use in the geometric model. It is believed that a significant improvement in performance can be achieved by obtaining more accurate centroid estimates in both the image and in the model. Modifications to the model to add points will increase redundancy that can also be used to improve accuracy.

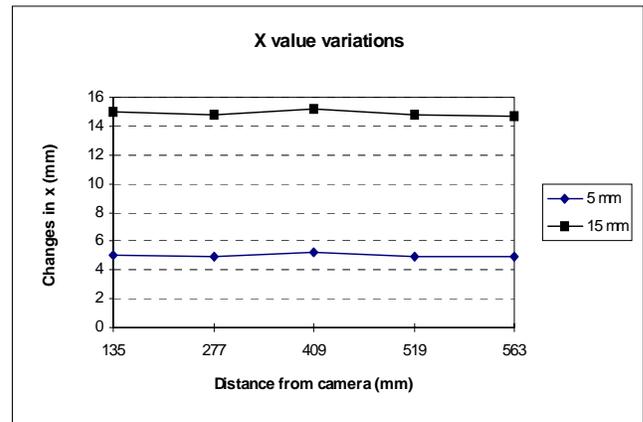


Figure 4. X-axis measurement

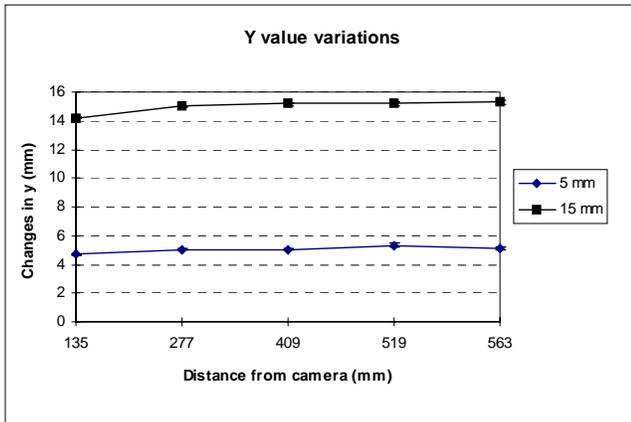


Figure 5. Y-axis measurement

Figure 7. Z-axis absolute error

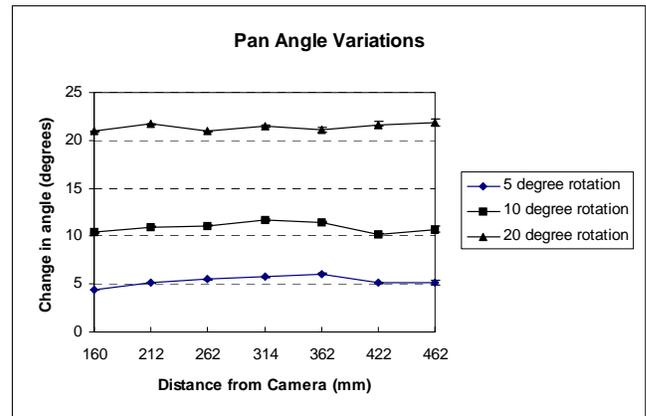


Figure 8. Pan angle measurement

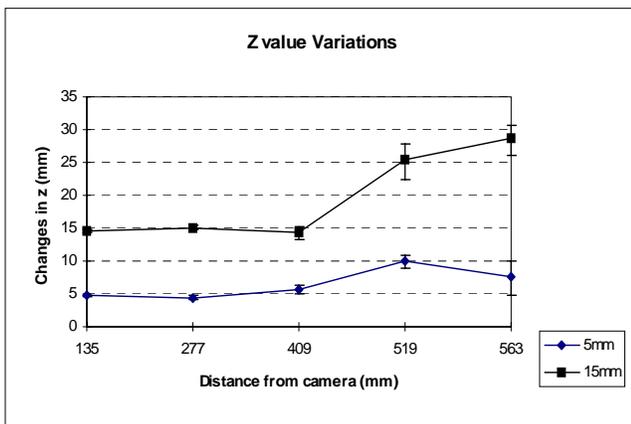


Figure 6. Z-axis measurement

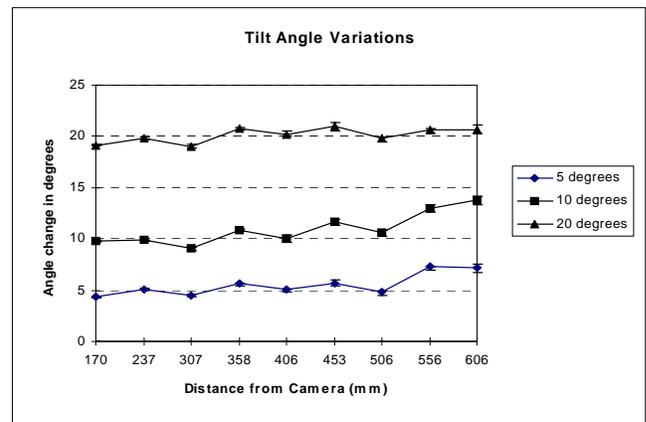
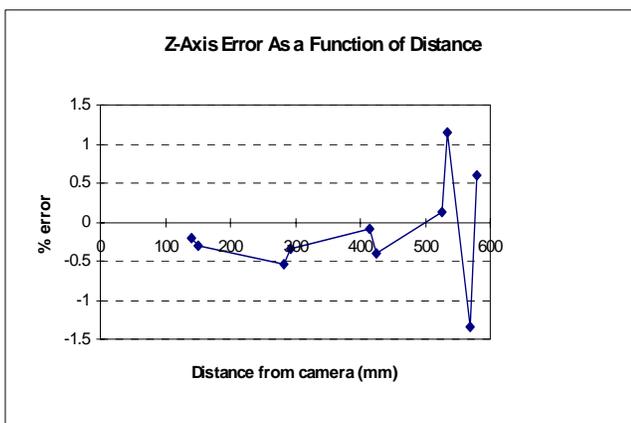


Figure 9. Tilt angle measurement



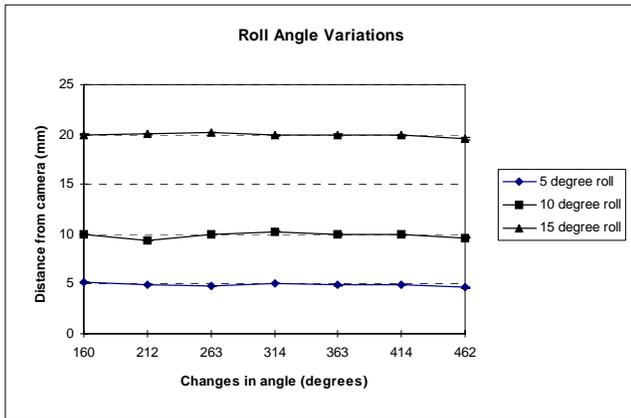


Figure 10. Roll angle measurement

VI. CONCLUSIONS

As part of ongoing development into robust pose determination methods for robotics applications, this paper describes a complete approach for single camera vision using feature points that are part of a 3-D model. This method has been implemented on real-time computer hardware and has demonstrated that fast and reliable pose measurements can be made. Experimental results show that this system can provide reliable, robust measurements of the target object with respect to the camera from six inches away to about seven feet away. This range is equivalent to between 1x and 12x the target size. These distances are strongly dependent on the focal length of the lens used and the brightness of the LEDs and do not necessarily reflect the limits of this method. Relatively accurate lateral and angular measurements were obtained at distances closer than three feet. Larger relative errors were obtained at greater distances although absolute measurements of distance remained accurate. Present work is directed towards increasing the accuracy and robustness of this technique through improved models and alternative algorithms.

REFERENCES

1. R. J. HOLT and A. N. NETRAVALI, "Camera Calibration Problem: Some New Results," *CVGIP: Image Understanding*, **54** (3), 368-383 (1991).
2. J. S.-C. YUAN, "A General Photogrammetric Method for Determining Object Position and Orientation," *IEEE Transactions on Robotics and Automation*, **5** (2), 129-142 (1989).
3. M. A. ABIDI and T. CHANDRA, "Pose Estimation for Camera Calibration and Landmark Tracking," *Proceedings of the IEEE International Conference on Robotics and Automation* (1990), pp. 420-426.
4. R. Y. TSAI, "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses," *IEEE Journal of Robotics and Automation*, **RA-3** (4), 323-344 (1987).
5. J. TIETZ and L. GERMANN, "Autonomous Rendezvous and Docking," *Proceedings of the 1982 American Control Conference* (1982), pp. 460-465.
6. M. A. FISCHLER and R. C. BOLLES, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, **24** (6), 381-395 (1981).
7. B. K. P. HORN, "Closed-form Solution of Absolute Orientation Using Unit Quaternions," *Journal of the Optical Society of America*, **4** (4), 629-642 (1987).
8. D. G. LOWE, "Three-Dimensional Object Recognition from Single Two-Dimensional Images," *Artificial Intelligence*, **31**, 355-395 (1987).
9. R. C. GONZALEZ and P. WINTZ, *Digital Image Processing*, Addison-Wesley, Reading, MA (1987).