Military Technical College
Kobry El-Kobbah,
Cairo, Egypt

6th International Conference
on Electrical Engineering
ICEENG 2008

Real time tracking in 3D space by image processing

By

Khalid A. S. Al-Khateeb* Mat Kamil Awang** Othman O. Khalifa*

Abstract:

A robotic vision system has been designed and analyzed for real time tracking of maneuvering objects. Passive detection using live TV images provides the tracking signals derived from the video data. The calibration and orientation of two cameras is done by a bundle adjustment technique. The target location algorithm determines the centroid coordinates of the target in the image plane and relates it to the aim point in the object plane. The stereoscopic images provide the information, from which the range, r of the object can be determined. The azimuth, θ and elevation, φ of the target with respect to a certain origin are determined by correlating the x-y displacements of the centroid in the image plane with the angular displacement of the target in the object plane. The servo drive signals for both the robot motion and the angular positioning of the cameras are derived from the image processing algorithm that keeps the centroid of the target image in the center of the frame and the target in line with the axis of the optical system. Hence, the spherical coordinates of the target are defined and updated with every TV frame. The time development of the centroid in successive TV frames represents the real time trajectory of the target path. A non-linear prediction technique keeps the target within the aim zone of the tracking system. In order to minimize the image processing time, i.e. kept within the demand of real time operation, one TV frame time, an image segmentation process is made to subtract nearly all redundant background details.

Keywords:

Robotic Vision, Real Time Tracking, Image Processing, Camera Calibration.

---

* Department of Electrical and Computer Engineering, Faculty of Engineering, International Islamic University Malaysia (IIUM)

** Security, Intelligent Applications and Multimedia, Telekom Research and Development, Malaysia

1. Introduction:

Stereoscopic vision is a popular research area especially in computer and robotic vision. This is demonstrated by the variety of problems that have been studied by researchers in recent years [1-14].

This paper deals with the calibration and orientation of the cameras in real time robotic stereoscopic vision, which is part of a more sophisticated system that relates to automatic object detection, tracking and identification for surveillance (ADTIS). The ADTIS is a platform under development at Telekom Research and Development. It consists of various modules, such as the TMEye, which is Viewer, Recording, Playback, Telemetry control, Alarm management, and Data-base function. The stereoscopic vision system under study comprises three main parts; camera-calibration, stereo imaging, and image processing.

In the camera calibration module, a polynomial model is used to compensate for image distortion induced by the optical system and the surrounding environment. The stereo-scoping imaging module adopts an epipolar line based algorithm for matching point searching. There are a number of methods for matching conjugate points, which can be implemented by various techniques, namely Gray-Level Matching, Correlation Methods, Edge-Matching or, Interpolation [9].

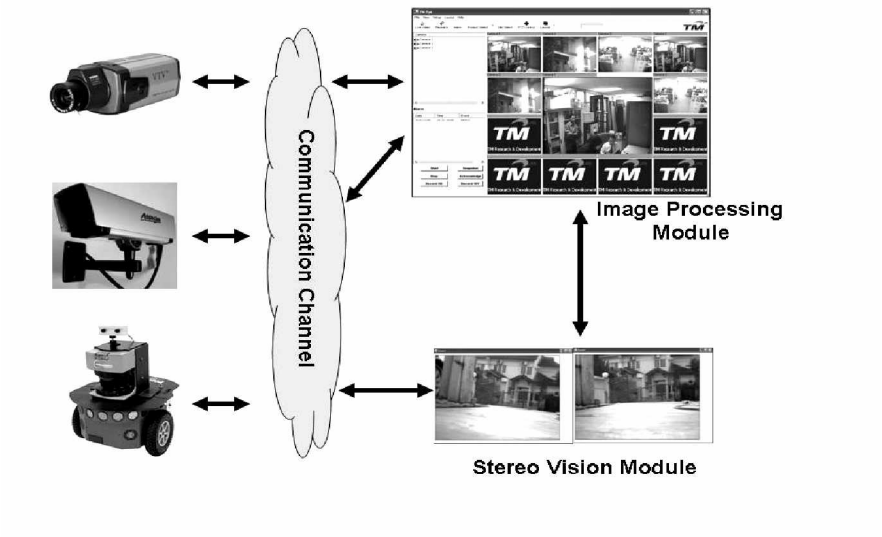


Figure 1*: Simplified Illustration of the Automatic Object Detection, Tracking and Identification System (ADTIS)*

The video image processing module is responsible for object detection, object tracking, and activity recognition. Object tracking is carried out by using a centroid tracking algorithm, background subtraction and non-linear prediction.

2. Stereoscopic Vision

The word "stereo" originates from Greek "stereos" which means solid. Stereo vision enables us to see objects as solid forms in three dimensional space (width, height and depth), or simply x, y and z. It is the added perception of the depth dimension that makes stereo vision important in robotic navigation.

In this application, two cameras, each captures a separate view and the two images are processed independently by the same processor. The processing result is the interpretation of depth from which the position of the object can be determined. This is rather like the human eyes, where the separate images of the left eye and right eye are sent to the brain and processed simultaneously. In this case however, although the left and right images from the two cameras are sent simultaneously, the processing is not simultaneous.

According to De Souza and Kak [4], robotic navigation can be grouped into 3 broad categories;

- .Map-Based Navigation; the robot depends on a user-created geometric model or a topological map of the environment.
- Map-Building-Based Navigation; the robot depends on sensors to construct a geometric or a topological model of the environment.
- Mapless Navigation; no explicit representation of the space in which navigation is to take place, but rather to recognize objects found in the environment to avoid or to track by generating motions based on visual observations.

The system described in this paper is the implementation of a real time robotic vision in a mapless navigation. This application poses the challenge [2], that the tuning of the camera position and orientation to ensure parallelism is more difficult to perform than the stereo vision itself.

The calibration process involves the estimation of the object location from the position and orientation of the two cameras, i.e. to find the coordinates of the object relative to a fixed origin in a general coordinate system. The parameters that are related to the cameras are the principal point or image centre, the focal length and the distortion

coefficients. The calibration method is a two-stage technique. The first is to compute the image distortion parameters and, the second is to determine the position and orientation of the cameras with respect to the general coordinate system.

3. Disparity between left and right images

The first task in developing a stereoscopic vision for robotic applications is the determination of the disparity between the left and right images. Figure 2 shows a simple illustration for the basic geometry of a stereoscopic arrangement.

In general, the two cameras are assumed rigidly attached together so that initially their optical axes are parallel and separated by a distance *b*, representing the base line, which is parallel to the x-axis.

In order to find the coordinates (x, y, z) of an object, represented by a point in the general coordinate system and measured relative to an origin midway between the lens centers, let the image-plane coordinates in the left and right images be (x_l, y_l) and (x_r, y_r), respectively.

$$\frac{x_l}{f} = \frac{x + b/2}{z} \text{ and } \frac{x_r}{f} = \frac{x + b/2}{z} \tag{1}$$

$$\frac{y_l}{f} = \frac{y_r}{f} = \frac{y}{z}$$

Where;
f is the distance from the lens centre to the image plane in both cameras and *b* is the distance between the lens centers. Hence x, y, and z can be found [9] by solving the following equations:

$$x = b\frac{(x_l + x_r)/2}{x_l - x_r}, \qquad y = b\frac{(y_l + y_r)/2}{y_l - y_r},$$

$$z = b\frac{f}{x_l - x_r} \tag{2}$$

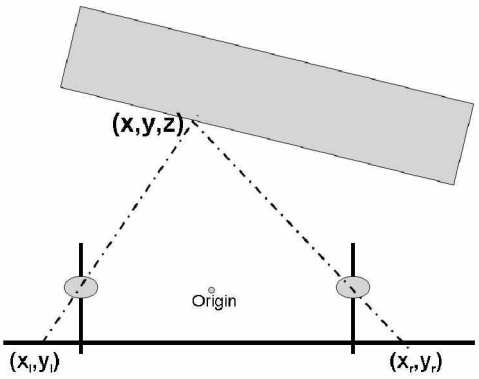


Figure 2: *Camera geometry for stereographic vision*

The difference in the image coordinates $(x_l - x_r)$ is the disparity between left and right images. The left and the right images will become more dissimilar when the separation between them increases.

The illustration of the disparity effect is shown in Figures 3(a) and 3(b). The distance between the lens centers of the left and right cameras was fixed to illustrate the dissimilarities between left and right images. It also illustrates the effect of illumination when the two cameras are placed at a relatively large distance apart. Illumination variation can pose a problem in determining matching conjugate points between left and right images. This problem however will be ignored for the time being, because it is insignificant in tackling the issue of camera calibration and orientation.

The selection of the parameter *b* can play a critical role in designing the stereoscopic robot vision system.

Figure 3a: *Image captured by the left camera*

Figure 3b: *Image captured by the right camera*

4. Image distortion compensation

Image compensation may be performed according to the two major causes of distortion. The first is due to the camera lens system itself and the second is due to the medium surrounding the cameras. It can be minimized by calibrating the cameras in the intended environment using the method proposed by Gremban [7]. An arbitrary point of coordinates (x, y, z) in the general coordinate system has a corresponding point in the TV monitor image defined by (x_k, y_k). Assuming *I* and *J* to be the order of the polynomials for x_k and y_k, respectively, then *I* and *J* can be chosen to be 3 to achieve reasonable compensation without taxing the processing time too heavily. The calibration coefficients α_{ij} and β_{ij} are given by:

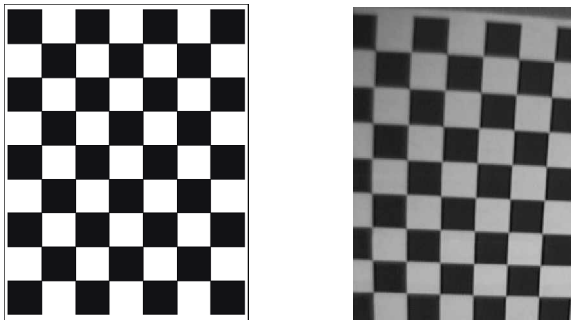
$$x = \sum_{i=0}^{I}\sum_{j=0}^{J}\alpha_{ij}x_k^i y_k^j, \qquad y = \sum_{i=0}^{I}\sum_{j=0}^{J}\beta_{ij}x_k^i y_k^j \tag{3}$$

The calibration coefficient is a function of z/f due to the effect of the perspective, which can be found by using a polynomial of the form shown below.

$$\begin{aligned}\alpha_{ij} &= \sum_{k=0}^{2} c_{\alpha k}\left(z/f\right)^k \\ \beta_{ij} &= \sum_{k=0}^{2} c_{\beta k}\left(z/f\right)^k\end{aligned} \tag{4}$$

A simplified calibration configuration was used, which basically utilizes a black and

white checker board, as in Fig. 4. The images were captured for three different positions at varying distances between the cameras and the board.



(a) (b)

Figure 4: *black and white checker board and one of the images used for the calibration.*

The method for finding the calibration coefficients [2] yielded the results shown below;

$$\mathrm{A}_R = \begin{bmatrix} c_{\alpha_{00}} = 0.002 & c_{\alpha_{01}} = -0.771 & c_{\alpha_{02}} = 8.864 \\ c_{\alpha_{10}} = -0.009 & c_{\alpha_{11}} = 2.506 & c_{\alpha_{12}} = -23.266 \\ c_{\alpha_{20}} = 0.001 & c_{\alpha_{21}} = -0.021 & c_{\alpha_{22}} = -0.651 \end{bmatrix}$$

$$\mathrm{B}_R = \begin{bmatrix} c_{\beta_{00}} = 0.001 & c_{\beta_{01}} = -0.426 & c_{\beta_{02}} = 0.088 \\ c_{\beta_{10}} = -0.001 & c_{\beta_{11}} = 0.037 & c_{\beta_{12}} = -1.801 \\ c_{\beta_{20}} = -0.007 & c_{\beta_{21}} = 2.395 & c_{\beta_{22}} = -16.700 \end{bmatrix}$$

$$\mathrm{A}_L = \begin{bmatrix} c_{\alpha_{00}} = 0.003 & c_{\alpha_{01}} = -0.672 & c_{\alpha_{02}} = 9.546 \\ c_{\alpha_{10}} = -0.052 & c_{\alpha_{11}} = 2.756 & c_{\alpha_{12}} = -20.394 \\ c_{\alpha_{20}} = 0.001 & c_{\alpha_{21}} = 0.131 & c_{\alpha_{22}} = -1.951 \end{bmatrix}$$

$$\mathrm{B}_L = \begin{bmatrix} c_{\beta_{00}} = 0.009 & c_{\beta_{01}} = -0.347 & c_{\beta_{02}} = 0.008 \\ c_{\beta_{10}} = -0.023 & c_{\beta_{11}} = 0.213 & c_{\beta_{12}} = -5.801 \\ c_{\beta_{20}} = -0.034 & c_{\beta_{21}} = 2.005 & c_{\beta_{22}} = -3.700 \end{bmatrix}$$

A_R and B_R are the Distortion Matrices for the R camera.

A_L and B_L are the Distortion Matrices for the L camera.

These elements are used in equation (8) for bundle adjustment.

5. Compensating for camera position and orientation

The other important parameters in stereoscopic vision, which must also be compensated, as represented in Fig. 5, are the position and orientation of the cameras.

One of the popular camera calibration techniques is the Tsai method [12]. This method is suitable for a wide range of applications, which can deal with coplanar and non-coplanar points. It also offers the possibility to calibrate internal and external parameters separately. Although this option can be quite useful since it gives the possibility to fix the internal parameters of the camera, when known, and carry out only pose estimation. It is probably more sophisticated than may be required. Hence, the bundle adjustment technique is found to be quite adequate for the purpose of this work.

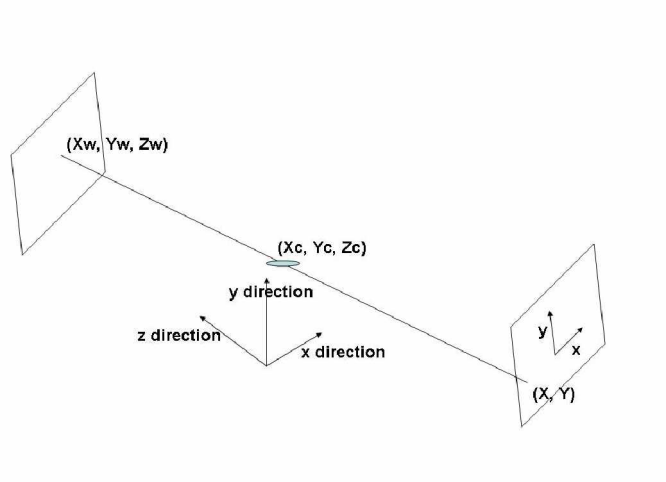


***Figure 5**: camera position and orientation*

The transformation from the general (X_w, Y_w, Z_w) to image (X_p, Y_p, Z_p) co-ordinates involves the translation and rotation operation as given by;

$$\begin{bmatrix} X_p \\ Y_p \\ Z_p \end{bmatrix} = R \begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} + T \tag{5}$$

where R and T are the 3D matrices, which transform the general coordinate system to the camera co-ordinate system, defined by:

$$R = \begin{bmatrix} r_1 & r_2 & r_3 \\ r_4 & r_5 & r_6 \\ r_7 & r_8 & r_9 \end{bmatrix} \quad T = \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix} \tag{6}$$

where:

$r_1 = \cos\phi \cos\gamma$

$r_2 = \sin\delta \sin\phi \cos\gamma - \cos\delta \sin\gamma$

$r_3 = \sin\delta \sin\gamma + \cos\delta \cos\gamma \sin\phi$

$r_4 = \cos\phi \sin\gamma$

$r_5 = \sin\delta \sin\phi \sin\gamma + \cos\delta \cos\gamma$

$r_6 = \cos\delta \sin\phi \sin\gamma - \cos\gamma \sin\delta$

$r_7 = -\sin\phi$

$r_8 = \cos\phi \sin\delta$

$r_9 = \cos\delta \cos\phi$

(δ, ϕ, γ) are Euler's angles of camera coordinate system that rotates relative to the general coordinate system and (T_x, T_y, T_z) are the 3D translation parameters from the general to the image coordinates.

The calibrated camera coordinates (X_c *and* Y_c) can be derived [2] from equations (3) and (4) as follows;

$$[X_c \quad Y_c]^T = M \left[\left(Z_c / f \right)^2 \quad Z_c / f \quad 1 \right]^T \tag{7},$$

$$\text{where } M = \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} A_1 & A_2 & A_3 \\ B_1 & B_2 & B_3 \end{bmatrix} = \begin{bmatrix} \sum c_{\alpha_0} x^m y^n & \sum c_{\alpha_1} x^m y^n & \sum c_{\alpha_2} x^m y^n \\ \sum c_{\beta_0} x^m y^n & \sum c_{\beta_1} x^m y^n & \sum c_{\beta_2} x^m y^n \end{bmatrix} \tag{8}$$

Equations (5) and (7) are the fundamentals of bundle adjustment in photogrammetry. The Calibration Matrices *M* above were calculated using bundle adjustment technique based on the least-square algorithm [3,6]. The bundle system model uses the parameters of camera calibration coupled with stereo imaging to determine the calibration coefficients.

6. The tracking algorithm

The tracking algorithm for a maneuvering object can be considered as part of the Image Processing Module. Tracking initially begins by detecting and locating the object of interest in a scene. This is done by a background subtraction technique, which identifies a moving object in a static environment from a portion of a video frame that differs significantly from the background. There are many challenges involved in developing a good background subtraction algorithm. The algorithm;

- must be robust against changes in illumination,
- avoids detecting irrelevant non-stationary background objects such as moving leaves, rain, snow, etc.
- reacts quickly to changes in background such as starting and stopping of moving objects.

The algorithm for each pixel in the background subtraction [10] involves the following steps:

1. Warp the pixel of the *key* image into a corresponding pixel of the *reference* image.
2. If the reference pixel has the same color and luminosity as the key pixel, then the key pixel is labeled as background.
3. If the reference pixel has different color and luminosity than the key pixel, then it is either a foreground object, or an occlusion.
4. In the case of stereoscopic cameras, the potential object pixels are verified by warping each of them to the other.

Background subtraction can be represented by;

$$m(\Gamma(x,y),\Theta(x',y')) = \begin{cases} 0 & if\ |\Gamma(x,y)-\Theta(x',y')| \le \varepsilon \\ 1 & otherwise \end{cases} \quad (9)$$

where:

(x,y) is a point in the key image,

$\Gamma(x,y)$ is the key image pixel,

d_x is horizontal disparity,

d_y is vertical disparity,

ε is the set subtraction threshold value,

$x' = x - d_x$ is the corresponding position in reference image,

$y' = y - d_y$ is the corresponding position in reference image,
$\Theta(x', y')$ is the reference image pixel
$m(\Gamma, \Theta)$ is a masking function

In Figures 6 (a) and (b) samples of the background images are used as key image for the background subtraction. The coordinates of the centroid of the object image $C(\hat{x}, \hat{y})$ can be determined [12] as shown below.

$$C(\hat{x}, \hat{y}) = \left(\hat{x} = \frac{\sum\sum_{(i,j)\in TG} i.c(i,j)}{\sum\sum_{(i,j)\in TG} c(i,j)}, \hat{y} = \frac{\sum\sum_{(i.j)\in TG} j.c(i,j)}{\sum\sum_{(i,j)\in TG} c(i,j)} \right) \quad (10)$$

where,
$\hat{x}$ is the target centroid estimate in the x direction
$\hat{y}$ is the target centroid estimate in the y direction
i, j are the image coordinates in the x and y direction
c(i,j) is pixel classifying function

Figure 6 (a) *Background captured by Left Camera,* ***(b)*** *Background captured by Right Camera*

The geometric centre of the object image is used to set the direction of the optical axis of each camera to the aim point of the target. Hence the azimuth and elevation angles of the target with respect to each camera are determined.

7. Indication of test results

The initial testing of the calibration and camera orientation was done using 2 web cameras with 640 by 480 pixels. The cameras are mounted on a wooden panel, and the distance between them was adjusted manually. The present system gave satisfactory results. However, the accuracy in the determination of the azimuth and the elevation angles is largely affected by the image resolution. The error in determining the range of the target on the other hand, is influenced not only by the resolution that affects the azimuth and elevation angles, but also by the error in the separation distance of the cameras. Hence, optimization is needed before the system can be implemented for the ADTIS. These results are considered encouraging for further development.

The initial tracking experiments are performed by fixing the cameras and moving an object within the field of view (FOV) of the cameras. The tracking performance is found acceptable. The processing takes less than 30 ms to grab a frame, display it on the screen, process the image, identify the moving object and send appropriate signals that move the cameras. The tracking proved to be reliable, with reasonable accuracy.

8. Conclusions:

Camera calibration, orientation and object tracking were developed for a real time robotic application. Based on this preliminary result, future work will be conducted using a proper robot with mounted stereo cameras. The calibration and orientation module will be installed onboard, which will be located within the housing of the mobile tracking robot.

Acknowledgements

The authors wish to thank the International Islamic University Malaysia, Research Centre at Gombak, Malaysia, and Telekom Research and Development Sdn. Bhd. for funding this research work.

References:

[1] A. Bensrhair, M. Bertozzi, A. Broggi, A. Fascioli, S. Mousset, and G. Toulminet, "Stereo Vision-Based Feature Extraction for Vehicle Detection", *IEEE Intelligent Vehicles Symposium., Versailles, France,* (2002)

[2] SW Cheng, HW Hsu & SC Cheng. "A Dynamic Vision System for Tracking and Localization of Underwater Objects", *Underwater Technology, International Symposium on,* pp. 215 – 222 (2004)

[3] R. Delara Jr., E. A. Mitishita, A. Habib, "Bundle Adjustment of Images From Non-Metric CCD Camera Using Lidar Data as Control Points", *Proceedings 20th ISPRS Congress,* (2004)

[4] Guilherme N. DeSouza, Avinash C. Kak. "Vision for Mobile Robot Navigation: A Survey", *Pattern Analysis and Machine Intelligence, IEEE Transaction on,* **Volume 24, No. 2** pp.237-267 (2002)

[5] Luis Falcon-Morales. "Design of Algorithms of Robot Vision Using Conformal Geometric Algebra", *International Mathematical Forum,* **Volume 2, No. 20**, pp. 981-1005. (2007)

[6] S.I. Granshaw, "Bundle Adjustment Methods in Engineering Photogrammetry", *Photogrametric Record*, **Volume 10, No. 56,** pp. 181-207 (1980)

[7] K.D Gremban, C.E Thorpe. "Geometric Camera Calibration Using System of Linear Equation", *Proceedings of IEEE International Conference on Robotics and Automation,* pp. 562-567 (1988)

[8] E. Grosso, M. Tistarelli. "Active/Dynamic Stereo Vision", *Pattern Analysis and machine Intelligence, IEEE Transaction on,* **Volume 17, No. 9,** pp. 868-879, (1995)

[9] B.K.P. Horn. "Robot Vision", MIT Press (1986).

[10] Yuri Ivanov, Aaron Bobick & John Liu. "Fast Lighting Independent Background Subtraction", *International Journal of Computer Vision,* **Volume 32, No. 2.** pp 199 - 207 (2000)

[11] Franck Jung, Didier Boldo. "Bundle Adjustment and Incidence of Linear Features on the Accuracy of External Calibration Parameters", *Proceedings 20th ISPRS Congress,* (2004)

[12] Khalid A. S. Al-Khateeb, Mohd. Akram Dandu. "Real Time Tracking in 3D Space by Robot Vision", ISM07 Sharja, UAE, March (2007)

[13] R.Y. Tsai. "An Efficient and Accurate Camera Calibration technique for 3D Machine Vision", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,* pp. 364-374 (1986)

[14] Jon A. Web. "Implementation and Performance of Fast Parallel Multi-Baseline Stereo Vision", *Image Understanding Workshop,* (1993)