Proceedings of the RAAD 2011
20th International Workshop on Robotics in Alpe-Adria-Danube Region
October 5-7, 2011, Brno, Czech Republic

An Active Stereo Vision Control System Based on PTZ Cameras for Robust Robotic Perception

Gigel Macesanu[a], Sorin Mihai Grigorescu[a] and Florin Moldoveanu[a]

[a] *Department of Automation, Transilvania University of Brasov, Romania*
E-mail: {gigel.macesanu, s.grigorescu, moldof}@unitbv.ro
URL: www.rovis.unitbv.ro

Abstract. In robotics, the most common approach to 3D reconstruction and scene understanding is through stereo vision. In order to keep track the 3D *positions and orientations* (pose) of objects of interest over a large Field of View (FOV), the orientations and focal lengths of the stereo camera system must be adjusted, that is, the pan, tilt and zoom (PTZ) of both vision sensors. In this paper the practical implementation of a 3D object reconstruction and tracking system, using two PTZ cameras in a stereo configuration, is presented. Firstly, the stereo platform is calibrated and the intrinsic and extrinsic parameters of the sensors determined. Secondly, the 2D image streams are processed in order to segment and classify the objects present in the environment. Based on the implemented methods, the objects of interest can be reconstructed in a virtual 3D space through the epipolar geometry constraints. The feedback mechanism for controlling the orientation and zoom of the two cameras are derived directly from the 3D poses of the objects of interest. As performance evaluation, we present a stability analysis of the proposed active vision system, taking into account the time-delay introduced by the image processing system.

Keywords. Active vision, 3D reconstruction, stability analysis

1. Introduction

Conventional robot vision systems use stereo cameras with a fixed orientation and zoom, this fact leading to physical limitations in the scene understanding process. In situation in which non stationary objects should be observed and tracked, the fixed cameras are replaced with systems that able to adapt their viewpoint and zoom along with the poses of the objects. Thus, the visual perceptual capabilities of an autonomous robot are adapted according to the imaged environment, evaluating the camera's extrinsic parameters (i.e. pose and zoom) modifications. Using such an active vision system, a robotic platform can track objects of interest and reconstruct their poses in a virtual 3D space.

The accuracy of a robotic platform depends directly on the accuracy of the visual sensor. An alternative to increasing the accuracy of these subsystems, that is, of the vision sensors is to use a visual-feedback control loop. The concept of visual feedback control has been heavily investigated in the computer vision community (Fitzpatrick, 2003), (Welke et al., 2009). The idea of using such feedback information at the image processing level can be found in older papers such as (Shirai and Inoue, 1973) where the effect of video feedback is described for robotic positioning tasks. This process, in which the visual data is used to control robotic platforms, is also encountered under the name of *visual servoing* (Corke, 1996).

In literature, there are a number of 3D object reconstruction methods, which can be classified into two different groups (Esteban and Schmitt, 2004): *active* and *passive methods*. The active methods use laser, or structured light systems, to obtain 3D data. The passive methods approaches use digital cameras which acquire images from different points of view. The main advantages of using the passive approach are related to the acquisition price and the operation simplicity. The 3D reconstruction model is constructed using the triangulation method, which represents the process of finding coordinates of a 3D point based on its corresponding stereo image points, as well as with the intrinsic parameters of the cameras (e.g. *focal length*, *optical centre*, etc.) (Cyganek and Siebert, 2009).

The main goal of the research presented in this paper is to reconstruct a 3D scene using the information acquired from a stereo camera system and to automatically adapt the orientation of the stereo camera according to the pose of the imaged

object of interest. The resulted 3D model can be used by an autonomous robot to navigate in unknown environments, avoid obstacles, or grasp objects if the robot is equipped with a redundant manipulator. In this paper we also propose a stability analysis of the active stereo vision architecture, taking into account the time-delay introduced by the image processing software.

The rest of the paper is organized as follows. In Section II, the stereo camera calibration and image rectification process are presented. The camera control system is described in Section III, followed in Section IV by its stability analysis. Finally, conclusions are stated in Section V.

2. Stereo camera calibration and rectification

One important problem to be solved in obtaining the position of an object of interest is the estimation of the distance between the camera and the object. This distance is determined using a calibrated stereo camera. In order to determine the object's 3D position, the images used must be rectified.

2.1. Stereo geometry

The standard model of a stereo camera is illustrated in Fig. 1 (Hartley and Zisserman, 2004). A real word point $P(x, y, z)$ is projected onto the image planes of a stereo camera as the homogeneous 2D image points:

$$\begin{cases} p_l = (x_l, y_l, 1), \\ p_r = (x_r, y_r, 1), \end{cases} \tag{1}$$

where, p_L and p_R represents the projection of a P point which has the (x_l, y_l) and (x_r, y_r) coordinates. The projected points are situated on the left I_L and right I_R images, respectively. The O_L and O_R represent the camera's optical centres. The plane formed by the optical centres and the P point represents the epipolar plane. This plane intersects the image plane in p_l and p_r. The image, or the principal plane, is located at a distance f from the optical centre of the camera, where f is the focal length. The z axis of the coordinate system attached to the optical centre is referred to as the principal ray, or optical axis. The intersection between the image plane and the principal ray at the image centre is known as the principal point, c_l and c_r. In order to determine de distance Z from the stereo camera to point P the distance b between the optical centres of the two cameras and the projection points p_l and p_r has to be known. Knowing these parameters, the 3D position of P with respect to the camera can be obtained. The 3D position of P is determined using the following equations (Huang, 2005):

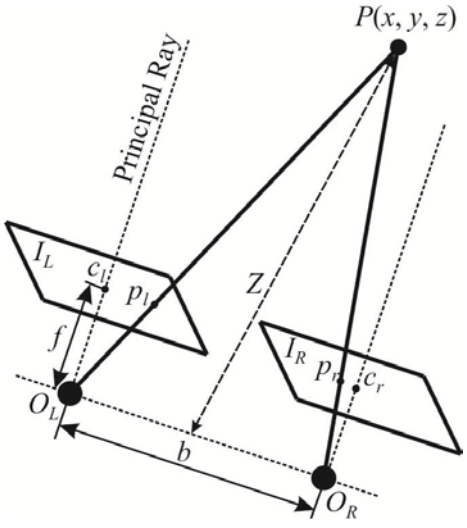

Fig. 1. Stereo camera geometry

$$x = x_l \cdot \frac{b}{d}, \tag{2}$$

$$y = y_l \cdot \frac{b}{d}, \tag{3}$$

$$z = f \cdot \frac{b}{d}, \tag{4}$$

where, d represent the disparity of the projected point P and is equal with:

$$d = x_l - y_r . \tag{5}$$

Using the above equations we are able to compute the 3D position of a point, based on its 2D Cartesian position and camera parameters.

2.2. Stereo calibration and rectification

The process of stereo camera calibration is supposed to determine the internal and external camera parameters. The internal parameters, that is *intrinsic parameters*, are obtained based on the Zang method (Zang, 2000). The resulted intrinsic calibration matrix has the follow form:

$$K = \begin{vmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{vmatrix}, \tag{6}$$

where, f_x and f_y are the horizontal and vertical camera's focal length, whereas c_x and c_y are the focal points.

Other computed parameters are the distortion coefficients. In this case we compute two types of distortions: *radial* and *tangential*. Radial distortions appear as a result of the shape of lens, whereas tangential distortions arise from the assembly process

of the camera. These parameters are computed using the method proposed by Brown (Brown, 1971).

While the above parameters are specific for each camera, and contain internal camera's specifications, the extrinsic parameters show the relative position of the cameras. The extrinsic parameters contain two vectors: rotation and translation and has the form:

$$E = \begin{vmatrix} R_{3x3} & T_{3x1} \\ 0_{1x3} & 1 \end{vmatrix}, \quad (7)$$

where, R_{3x3} represent the rotation matrix and T_{3x1} is the translation matrix. These vectors contain the position of the left camera relatively to the right camera. First parameters of the translation vector represent the baseline *b* (see Fig. 1).

In our implementation we calculate the intrinsic and extrinsic matrices using a calibration chessboard table, imaged in a number of 25 frames.

The rectification process transforms each image plain in such a way that pairs of conjugate epipolar lines become collinear and parallel to one of the image axes (Fusiello et al. 2000). This process is needed because the real stereo system isn't perfectly aligned, since the two cameras almost never have exactly coplanar, row-aligned imaging planes.

The process of rectification is realized based on the simplified method developed by Zhang (Zhang, 2000) who use only the rotation matrix, R_{3x3}, and the translation matrix, T_{3x1}, obtained from the calibration process. The returned values from the rectification process include two matrices that rotate the left camera about the centre of projection so that the position of epipols to be at infinity and the epipolar line becomes horizontal (see Fig. 1) (Bradski and Kaehler, 2008).

2.4. Stereo correspondences

Another important process in scene reconstruction is represented by the correspondences problem. The idea is to find the same point from the left image into the right images. If the camera geometry is known, the correspondence problem is reduced to the fact that points in one image can correspond only to points along the same scan-line in the other image (Ogale and Aloimonos, 2005). Thus, using the geometry resulted from the calibration process and the rectified images we can compute correspondences between the two views.

In this paper, the method of *Block Matching* (BM) has been used for estimating camera-objects distances. The block matching algorithm is based on using small windows to find matching points between the left and the right rectified stereo images. The match is based by computing a *Sum of Absolute Differences* (SAD) (Kisacanin et al., 2009). This algorithm is used for measuring the similarity between image blocks and works by computing the absolute difference between each pixel in the original block and the corresponding pixel in the block being used for comparison.

The process of block matching using SAD can be divided into three distinct stages:

- *pre-filtering* – the input images are normalized, in order to enhance textures and to reduce lighting differences, these are done by using a 7x7 window;
- *correspondence matching* – in this stage are search correspondences using a sliding SAD window. The search is done along the horizontal epipolar lines;
- *post-filtering* – represent the process of eliminate bad correspondences matches.

The BM algorithm is applied on rectified images, thus, correspondences are search along the same row in both images of the stereo pair. The interval in which the correspondent point is search has a finite distance, with its low value called *minimum disparity*, while it's high value is named *maximum disparity*. The interval between the minimum and maximum value is the so-called *horopter*, defined as the 3D volume that is covered by the search range of the stereo algorithm (Cyganek and Siebert, 2009), (Bradski and Kaehler, 2008).

3. Stereo camera control

The block diagram of the proposed stereo control architecture is presented in Fig. 2. The robotic head is composed from two digital Pan-Tilt-Zoom cameras which compose the artificial vision system. Based on these cameras, stereo images are acquired and used for analytical stereo calibration and image rectification. At the same time images are used for object recognition. Both results, calibration with rectification and object recognition are used as inputs for the 3D scene reconstruction.

3.1. 2D object recognition

The process of object of interest detection begins with an images filtering. The process of filtering involves the shifting of a filter mask, $w(i, j)$ over the input image. At each point (*x*, *y*), the response of the filter at that point is calculated using a predefined relationship (Gonzalez and Woods, 2002). In our implementation a 3x3 mask was used. The object of interest is detected using color segmentation. This involves a color representation conversion, from RGB (*Red*, *Green*, *Blue*) format to HSI (*Hue*, *Saturation*, *Intensity*) color model. Using the conversion other colours than the desired are rejected. The segmentation process is followed by the detection of the object's contour in the 2D image plane using the chain-code border following method (Bradski and Kaehler, 2008).

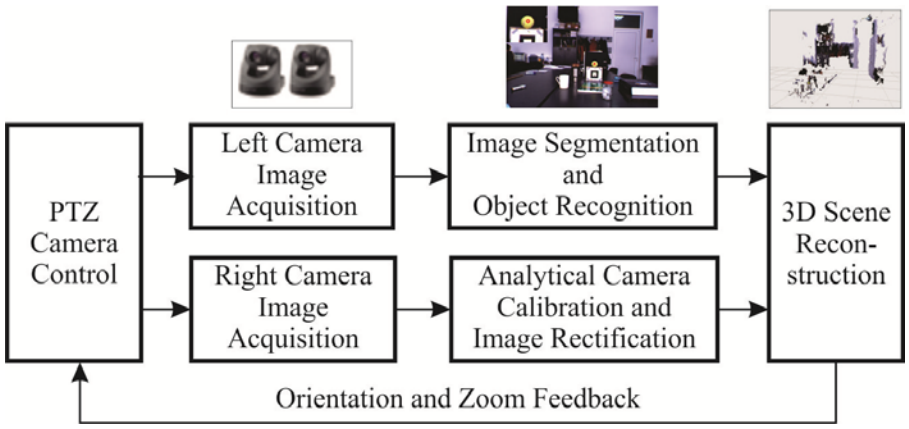

Fig. 2. The architecture of the stereo control system

The object of interest contour recognition is performed using invariant Hu moments that are invariant to object rotation, translation and scaling. Using the coordinates of the central of gravity for each object, detected in the left and the right image and the epipolar geometry, presented in previous section, we are able to reconstruct the 3D position of the object. Using the 3D position of an object of interest we can perform an active vision control that involves sending back to the PTZ camera control, as feedback information, the current object position. This fact leads to adapting the cameras parameters in such a way to obtain an optimal visualization. Using only the correspondent points, extracted from the block matching algorithm, we can reconstruct the whole viewed scene. This robotic active vision system can perform an object of interest proper visualization and whole scene reconstruction.

3.2. Camera orientation system

During the robot navigation the position of the objects of interest can change. The goal of the proposed active vision system is to maintain the image object in the middle of the 2D image plane, as well as within the image boundaries. These conditions can be satisfied using a system with the camera's *Field of View* (FOV) automatically adjustable, in order to cover a wider scene as possible. The scene where the position of the objects of interest can be too far or too close from the vision sensor can't be analysed using conventional pan and tilt cameras. This problem can be solved by adding an extra *Degree of Freedom* (DoF) to the camera, that is, of the zoom control system. This extra control system aims at controlling the focal length through with the environment is sensed.

The active vision system has to minimize three errors, one for each axis of the coordinate system. The errors which have to be compensated by the active vision system are illustrated in Fig. 3, where the position of a real world point, $pt_{\text{int}}(x, y, z)$ in the 3D Cartesian space, along with its 2D location in the image plane $pt_{proj}(x, y)$ is presented. The 3D point $pt_{ref}(x_r, y_r, z_r)$ represents the reference (desired) position for the object of interest position. The x_r and y_r coordinates represent the 2D central coordinates, while the z_r is the distance, or depth, from the middle point on the baseline between the two cameras and the imaged object.

The position error can be defined as the difference between the current position of the objects of interest in 3D space and the reference position:

$$e = pt_{\text{int}} - pt_{ref}, \tag{8}$$

where, $e = |e_x, e_y, e_z|$, represent the position error vector, $pt_{\text{int}}(x, y, z)$ and $pt_{ref}(x_r, y_r, z_r)$ represent the real position and the desired position of the object of interest.

The goal of the visual controller is to minimize the e_x and e_y errors by adjusting the pan and tilt values of the cameras and e_z by controlling the focal length of the cameras.

4. Stability analysis

To evaluate the performance of the system we analyse the stability of the proposed control structure.

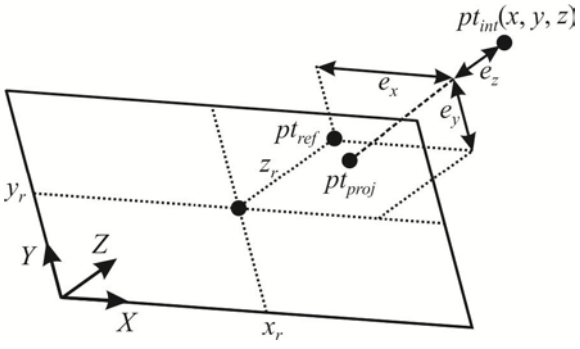

Fig. 3. The position error for a 3D point

The analysis supposed to study all three moving elements: pan, tilt and zoom. Because all of them are identical as structure, only their role in the system is different, we analyse only the pan (the other two are identical). In this sense we have used the *Nyquist criterion* with the gain margin k_g and phase margin γ for the frequency response method (Dorf and Bishop, 2010).

In the Nyquist criterion, the locus of open-loop transfer function represents a measure of the system's relative stability. The objective of this analysis it to determine the gain and phase margins in order to observe the system's stability reserve.

The open-loop transfer function, in the frequency domain of the active vision system, with its block diagram presented in Fig. 4, is defined as:

$$G(j\omega) = \frac{1.7}{1+0.24\,j\omega} \cdot e^{-s\tau}, \tag{9}$$

where, τ represent the time-delay introduced by the image processing component (Corke, 1996). The transfer function is composed from a DC (*direct current*) motor with its controller and an additional controller for pixels to angles conversions. The DC and its controller are modelled as *first order lag element* (PT1) and the conversion system as a *proportional element* (P). The parameters of the controllers were chosen in order to obtain a zero steady-state error and an optimal rise time.

The analysis is performed in an open-loop loop manner on the transfer function, without considering the time-delay. The goal of the approach is to determine the open-loop stability and to observe the closed-loop system's evolution when a time-delay is introduced. For this purpose, the phase margin and its associated gain crossover frequency has to be determined.

The gain crossover frequency is determined using the following equation:

$$\left|G^*(j\omega_g)\right| = 1, \tag{10}$$

where, $G^*(j\omega_g)$ represent the open-loop transfer function without time-delay and ω_g represent the gain crossover frequency. After solving the above equation, a value of $\omega_g = 5.728$ is obtained.

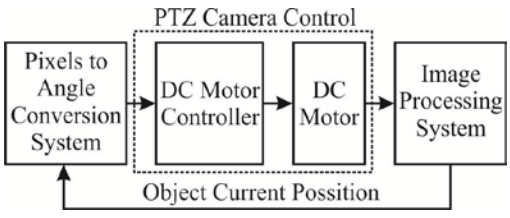

Fig. 4. Block diagram of the active control system

The phase margin for the time-delay free open-loop transfer function can be written as:

$$\begin{aligned}\gamma &= 180^\circ + \arg(G^*(j\omega_g)) = \\ &= 180^\circ - arctg(\omega_g \cdot 0.24) = 126.1^\circ\end{aligned} \tag{11}$$

When a time-delay exists within the system, it affects the phase margin and its associated frequency. The phase margin is the amount of additional phase lag, at the gain crossover frequency, which can bring the system to the stability limit (Ogata, 2002), can be computed as:

$$180^\circ + \arg(G^*(j\omega_g)) - \omega_g \cdot \tau, \tag{12}$$

where τ represent the time-delay.

When the system is at stability limit, the time-delay phase margin is zero:

$$180^\circ + \arg(G^*(j\omega_g)) - \omega_g \cdot \tau = 0. \tag{13}$$

Using the previously equations we can determine the maximum value of the time-delay, τ_{max} after which the overall closed-loop system becomes unstable:

$$\tau_{max} = \frac{[180^\circ + \arg(G^*(j\omega_g))]_{rad}}{\omega_g} = 0.38\,\text{sec} \tag{14}$$

The obtained maximum values of time-delay Eq. (14) indicates that the image processing chain from Fig. 2 should process a pair of stereo images (e.g. image acquisition, segmentation, object recognition and 3D reconstruction) in less than 0.38sec.

5. Conclusion

This paper presents an active stereo vision system for autonomous robots which have to sense the 3D structure of the images environment. The presented algorithms aim at detecting the 2D location of the object of interest in the image plane and at reconstructing their pose in a virtual 3D space. Based on the obtained 3D model the orientation of the camera system can be automatically adapted. The stability of the proposed system is investigated based on the time-delay introduced by the image processing software.

6. Acknowledgments

This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contracts

number POSDRU/88/1.5/S/59321, POSDRU/89/1.5/S/59323.

7. References

Bradski, G. and A. Kaehler. 2008. *Learning OpenCV*. O'Reilly Media, Sebastopol, USA.

Corke, P.I. 1996. *Visual Control of Robots: A High-performance Visual Servoing*. Research Studies Press Ltd., John Wiley & Sons Inc, Great Britain.

Cyganek, B. and J.P. Siebert. 2009. *An Introduction to 3D Computer Vision Techniques and Algorithms*. John Wiley & Sons, Great Britain.

Dorf, R.C. and R.H. Bishop. 2010. *Modern Control Systems*, 12th edition. Prentice Hall, New Jersey, USA.

Esteban, C. H. and F. Schmitt. 2004. Silhouette and Stereo Fusion for 3D Object Modeling. In: *Computer Vision and Image Understanding*, Vol. **96**(2), pp. 367-392.

Fitzpatrick, P. 2003. First Contact: an Active Vision Approach to Segmentation. In: *Proc. Of the 2003 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Las Vegas, USA, pp. 2161-2166.

Gonzalez, R.C. and R.E. Woods. 2002. *Digital Image Processing*, 2nd edition. Prentice Hall, New Jersey, USA.

Hartley, R. and A. Zisserman. 2004. *Multiple View Geometry in Computer Vision*, 2nd Edition, Cambridge University Press, Cambrige, United Kingdom.

Huang Y., S. Fu and C. Thompson. 2005. Stereovision-Based Object Segmentation for Automotive Applications. In: *EURASIP Journal on Applied Signal Processing*, Vol. **2005**(14), pp. 2322–2329.

Kisacanin, B., S.S. Bhattacharyya and S. Chai. 2009. *Embedded Computer Vision*. Springer-Verlag, London, United Kingdom.

Ogale, A.S. and Y.A Kinematics. 2005. Shape and the stereo correspondence problem. In: *Int. Jurnal of Computer Vision*, Vol. **65**(3), pp. 147-162.

Ogata, K. 2002. *Modern Control Engineering*, 4th edition. Prentice Hall, New Jersey, USA.

Shirai, Y. and H. Inoue. 1973. Guiding a Robot by Visual Feedback in Assembling Tasks. In: *Pattern Recognition*. Vol. **5**, pp. 99-108.

Welke, K., A. Tamim and R. Dillmann. 2009. Active Multi-View Object Search on a Humanoid Head. In: *2009 IEEE International Conference on Robotics and Automation*, Kobe, Japan, pp. 417-423.

Zhang, Z. 2000. A Flexible New Technique for Camera Calibration. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. **22**(11), pp.1330-1334.