Bulletin of the *Transilvania* University of Braşov
Series I: Engineering Sciences • Vol. 4 (53) No. 1 - 2011

AN OBJECT DETECTION AND 3D RECONSTRUCTION APPROACH FOR REAL-TIME SCENE UNDERSTANDING

G. MĂCEŞANU[1] S. GRIGORESCU[1] T.T. COCIAŞ[1]
F. MOLDOVEANU[1]

Abstract: *This paper present the theoretical and practical aspects of a 3D object reconstruction approach. The reconstruction process involves the usage of a pair of stereo rectifies images acquired from two digital cameras. The imaged object is rebuilt in a virtual 3D space with the help of internal and external camera parameters obtained from the camera calibration algorithm. For the 2D image detection of the object of interest a color segmentation method is applied, followed by triangulation for estimating the object-camera distance.*

Key words: *object recognition, stereo vision, depth estimation.*

1. Introduction

3D object reconstruction is the task of generating a 3D model of an object given multiple 2D images taken of a scene. Such algorithms can be found in applications such as robotics, virtual reality and entertainment.

The most common approach to 3D reconstruction, or depth sensation, is through stereo vision. Early work, conducted in the 1970s and early 1980s, was primarily conducted by the image understanding community [3]. Barnard and Fischler [1] firstly reviewed stereo reconstruction methods in 1981, with focus on fundamental algorithms and criterias for performance evaluation. Recently, books by Hartley and Zisserman [9] and Cyganek and Siebert [4] provide a wealth of information on the geometric aspects of multiple view stereo geometry.

In literature, there are a number of 3D object reconstruction methods, which can be classified into two different groups [10]: *active* and *passive* methods. The *active methods* use laser, Time-of-Flight (ToF), or structured light systems to obtain 3D data. Still, they remain expensive and require special skills for the acquisition process itself. The *passive methods* approach use digital cameras which acquire images from different points of view. The 3D information is then extracted from the sequence of 2D colour images by using different techniques [10], [13]. One of the most used techniques is based on recovering information regarding the structure of a 3D space directly from depth measurements. The depth is usually obtained from computing stereo matching between pairs of images [4]. This technique, known as triangulation, represent the process of finding

[1] Dept. of Automatics, *Transilvania* University of Braşov.

coordinates of a 3D point based on its corresponding stereo image points, as well as with the parameters of the cameras (e.g. focal length, optical centre etc.) [4], [11].

The rest of the paper is organized as follows. In Section 2, an object detection approach based on color processing is presented, followed in Section 3 by a detailed description of the 3D reconstruction steps. In Section 4, an evaluation of the obtained results is given. Finally, conclusions are presented in Section 5.

2. Object of Interest Detection

In order to extract the coordinates of the object of interest in successive images, three steps must be accomplished. Those steps are composed of *image enhancement*, *segmentation* and *object detection.*

Usually a color image $f(x, y)$ is represented in and RGB (*Red, Green, Blue*). On this RGB image a median filter is applied. The process involves the shifting of a filter mask, $w(i, j)$, over the input image $f(x, y)$. At each point (x, y), the response of the filter at that point is calculated using a predefined relationship [6]. In our implementation a 3 x 3 mask was used.

The second stage is represented by object segmentation, a process based on the HSI (*Hue, Saturation, Intensity*) color model. The representation of the model is given in Figure 1. The HSI model of an image is obtained from the RGB one using the following color transformation system [6]:

$$H = \begin{cases} \theta, \ if \ B \le G, \\ 360 - \theta, \ if \ B \ge G, \end{cases} \tag{1}$$

where:

$$\theta = \cos^{-1}\left\{ \frac{\frac{1}{2}\cdot[(R-G)+(R-B)]}{[(R-G)^2+(R-B)(G-B)]^{1/2}} \right\}. \tag{2}$$

The saturation component is given by:

$$S = 1 - \frac{3}{R+G+B}\cdot \min(R,G,B). \tag{3}$$

Finally, the intensity component is given by:

$$I = \frac{1}{3}(R+G+B). \tag{4}$$

The usage of the HSI systems simplifies the object segmentation via color information, since in this case the color is represented only on the hue image plane, in comparison to the RGB system, where the color is distributed over all three channels.

The obtained hue image $f_h(x, y)$ is thus segmented in order to separate the object of interest from background and other objects present in the scene. This process uses two thresholds values, $[T_1, T_2]$. The output pixels which form the binary segmented image are defined based on the

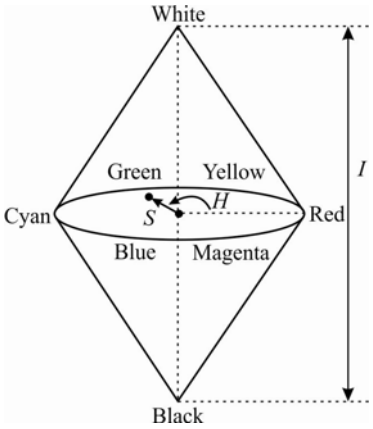

Fig. 1. *HSI Color model representation*

input hue pixels and are calculated using the following equation [6]:

$$t_h(x,y)=\begin{cases}1, if\ f_h(x,y)\in[T_1,T_2],\\ 0, if\ f_h(x,y)\notin[T_1,T_2],\end{cases} \quad (5)$$

where $t_h(x, y)$ is the binary segmented hue image.

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 [2].

These contours are specific for each segmented object. Once the contours have been detected, their image moments will be computed. The moments represent a certain particular weighted average of the image pixel's intensities, defined as [2]:

$$M_{i,j}=\sum_{k=1}^{n} I(x,y)x^i y^j, \quad (6)$$

where, $I(x, y)$ represent the x and y intensity, while $M_{i,j}$ is the (i, j) moment.

In cluttered scenes, the imaged objects usually have an altered shape in comparison to their original reference shape. To solve this problem invariant Hu moments are computed [2]. This moments use moments general to calculate a set of coefficients invariant to rotation, translation and scaling. Using this characteristic, an object can be uniquely detected, frame after frame, and its center of gravity (x_c, y_c) determined for the purpose of 3D reconstruction:

$$\begin{cases}x_c = M_{10}/M_{00},\\ y_c = M_{01}/M_{00},\end{cases} \quad (7)$$

where, $M_{i,j}$ represent the moments determined using the Equation (6).

3. 3D Object Reconstruction

The main goal of 3D reconstruction is to estimate the geometrical distances between the viewed scene and the camera. To be able to compute a distance between camera and an object, a calibrated stereo camera must be used. In order to determine the object's 3D position, the used images must be rectified. 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 [5].

The reconstruction of objects based on stereo image acquisition involves the following stages:

- stereo camera calibration and image rectification;
- 2D points matching in pairs of images;
- distance, or depth, computation.

3.1. Stereo camera geometry

The geometry of a stereo image acquisition system is entitled epipolar geometry. This geometry is illustrated in Figure 2 and behaves as follows.

Consider a real world 3D point P in homogeneous coordinates represented as $P = [X\ Y\ Z\ 1]$. This point is projected onto the left and right images planes, I_L and I_R, respectively. The projection points, on the left and right images, in homogeneous coordinates are: $p_L = [x_L\ y_L\ l]$ and $p_R = [x_R\ y_R\ l]$. O_L and O_R are the optical centres of both cameras, as show in Figure 2.

The line between the 3D real point P and

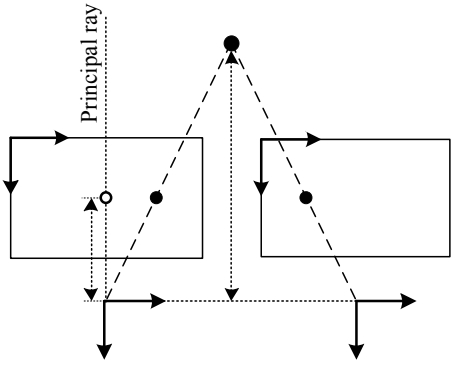

Fig. 2. *Depth estimation of a point P on a pair of rectified stereo images*

the optical centres, intersect the image plane in the projection points. The image plane is located at the distance *f*, or the focal length, from the optical centre of an each camera.

The *z* axis of each coordinate system represents the principal ray, or optical axis, whereas (c_x, c_y) is the principal point placed at intersection between the principal ray and image plane.

3.2. Stereo camera calibration and image rectification

This process represents the computation of both camera matrices. Through camera calibration the external (*extrinsic*) and internal (*intrinsic*) parameters of the cameras are computed. The intrinsic camera matrix has the following expression [4]:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{8}$$

where, f_x and f_y are the camera's focal length over the *x* and *y* axes and c_x and c_y represent the focal point.

The extrinsic matrix contains information about the relation between the left and right sensors of the stereo camera. This is expressed using a rotation matrix *R* and a translation matrix *T*. The extrinsic camera matrix, *P*, is defined as a combination of *R* and *T* and represented as:

$$P = \begin{bmatrix} R_{3x3} & T_{3x1} \\ 0_{1x3} & 1 \end{bmatrix}. \tag{9}$$

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

The rectification procedure aligns a pair of images in such a way that the corresponding points in both images reside on the same line [5]. Rectification was performed using Hartley's method [8], which only needs the computation of the fundamental matrix. This can be obtained from any matched set of seven or more points between the two views of the scene.

3.3. Stereo correspondences computation

In the stereo correspondence process-matching a 3D point in the two different camera views can be computed only over the visual areas in which the views of the two image sensors overlap [2]. 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) [2], [12]. The process of block matching using SAD can be divided into three distinct stages:

- pre-filter the input images, in order to enhance textures and to reduce lighting, these are done by using a 5x5 window;
- correspondence is computed with a sliding SAD window;
- post-filtering to eliminate bad corresponding matches.

Since in BM the correspondence point calculation is obtained based on rectified images, any match must occur on 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 [2], [4].

3.4. Depth computation

The distance between the stereo camera and the 3D point can be evaluated based on

the baseline (the distance between the two optical centres) and the projection points p_L and p_R. Knowing these parameters we can obtain the 3D position of P with respect to the camera. The 3D position of P is determined using the following equations [7]:

$$X = x_L \cdot \frac{T}{d}, \tag{10}$$

$$Y = y_L \cdot \frac{T}{d}, \tag{11}$$

$$Z = f \cdot \frac{T}{d}, \tag{12}$$

where, d represent the disparity of the projected point P and can be computed as:

$$d = x_L - x_R . \tag{13}$$

In equations (10) to (13) we can see that the distance is inversely proportional to the disparity.

4. Experimental Results

In order to test the theoretical description presented above, a practical experiment has been performed. In the experiment two Sony Evi-D70P® mono-cameras were used in a stereo manner. The considered object of interest was a green tennis ball, whose centre of gravity must be projected into a 3D virtual space. For simplicity, in Figure 3 only the left images are presented, while the process is applied to both images, that is, to the stereo images. The output of color segmentation from the input left image is presented in Figure 3a. At this, the optimal thresholds $T_1 = 42$ and $T_2 = 68$ used applied in order to get the binary segmented image. This interval corresponds to the green values of the considered object, as shown Figure 1. The detected centre of the tennis ball is presented in Figure 3b, where the red circle represents the centre of gravity. Based on the obtained centres, we are able to get the 3D position of the object using equations (10)…(13). Thus, the 3D position (x, y, z) of point P (in this case, the centre of the ball) has the coordinates (–0.144, 0.018, 1.427). Since the real ball-camera distance has a value of 1.41 m, that is, over the z axis, the estimated depth computation error is 0.017 m. Figure 3c illustrates the resulted 3D reconstructed scene.

5. Conclusions

In this work a theoretical and practical description of a 3D reconstruction algorithm has been presented. Using such a system, the position of an object of interest within a complex scene can be reconstructed. The accuracy of the proposed solution depends on the stereo acquisition system and the calibration process, respectively.

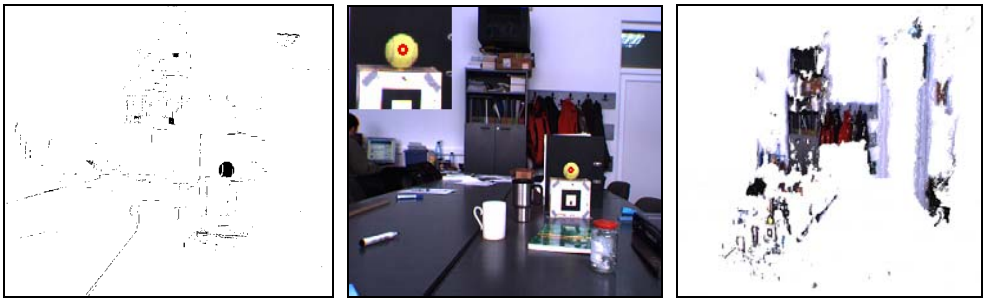

Fig. 3. *The reconstruction of an object of interest: a) color segmentation result; b) object of interest detection; c) 3D depth estimation*

Acknowledgements

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 and POSDRU/107/1.5/S/76945.

References

1. Barnard, S.T., Fischler, M.A.: *Computational Stereo.* In: ACM Computing Surveys **14** (1982), p. 553-572.
2. Bradski, G., Kaehler, A.: *Learning OpenCV*. Sebastopol, USA. O'Reilly Media, 2008.
3. Brown, M.Z., Burschka, D., Hager, G.D.: *Advances in Computational Stereo.* In: IEEE Trans. on Pattern Analysis and Machine Intelligence **25** (2003) No. 8, p. 993-1008.
4. Cyganek, B., Siebert, J.P.: *An Introduction to 3D Computer Vision Techniques and Algorithms*. Great Britain. John Wiley & Sons, 2009.
5. Fusiello, A., Trucco, E., Verri, A.: *A Compact Algorithm for Rectification of Stereo Pairs*. In: Machine Vision and Applications **12** (2000) No. 1, p. 16-22.
6. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*. 2nd Edition. New Jersey, USA. Prentice Hall, 2002.
7. Grigorescu, S.M., Moldoveanu, F.: *Controlling Depth Estimation for Robust Robotic Perception*. In: 18th World Congress of the Int. Federation of Automatic Control (2011), *to be published*.
8. Hartley, R.I.: *Theory and Practice of Projective Rectification*. In: Int. Journal of Computer Vision **35** (1998), p. 115-127.
9. Hartley, R., Zisserman, A.: *Multiple View Geometry in Computer Vision*. Cambridge. UK, Cambridge Univ. Press, 2000.
10. Hernandez, E.C., Schmitt, F.: *Multi-stereo 3D Object Reconstruction*. In: Proc. of the First Inter. Symp. on 3D Data Processing Visualization and Transmission (2002), p. 159-166.
11. Huang, Y., Duan, Z.C., Zhu, G.L., Gong, S.H.: *A Fast Triangulation Algorithm for 3D Reconstruction from Planar Contours.* In: Int. Journal of Advance Manufactory Technology **24** (2004), p. 98-101.
12. Kisacanin, B., Bhattacharyya, S.S., Chai, S.: *Embedded Computer Vision*. London. United Kingdom, Springer-Verlag, 2009.
13. Suliman, C., Moldoveanu, F.: *Video Image Processing for Mobile Robot Indoor Navigation*. In: Bulletin of the *Transilvania* University of Braşov (2008) Vol. 1 (50), Series I, p. 407-412.