Depth Estimation using Shifted Digital Still Camera

Iva Nikolova, Atanas Nikolov and Georgi Zapryanov

Abstract: *The present work investigates a simple method for determining the distance to objects in a scene using the principles of the canonical stereo vision systems. The objective is to prove by the physical experiments that using conventional digital still camera in combination with image analysis techniques relying on binocular cues it is possible to effectively determine the distance to particular objects in a given scene. The main request is that the camera should have precise horizontal movement, high resolution and possibilities of adjusting the parameters of the optical system. Experimental results with structured scenes and camera shifted on various distances demonstrate the effectiveness of the method in providing a reliable estimation of the depth of a scene, and also outline some of its limitations and shortcomings.*

Key words: *Image Analysis, Depth Estimation, Canonical Stereovision System, Digital Still Camera*

1. INTRODUCTION

The human activity field is a 3D world, where the location of each point is represented by *x,y,z* coordinates. Therefore, it is highly demanded to be able to get all the three coordinates for the interested point in the field. However, cameras can only capture a two-dimensional image where each point is represented by x and *y* coordinates.

There are many methods providing mechanisms for acquiring the z coordinate (referred to as the depth). In principle depth can be recovered either from monocular cues (shading, shape, texture, motion) or from binocular cues (stereo correspondences). Conventional methods for depth estimation have relied on multiple images. Stereo vision [1, 3] measures disparities between a pair of images of the same scene taken from two different viewpoints and uses the disparities to recover the depth. Structure from motion (SFM) [7] computes the 2D motion field of corresponding scene points in order to recover the 3D motion and the depth of the scene. Depth from focus (DFF) [5] captures a set of images using multiple focus settings and measures the sharpness of image at each pixel locations. The sharpest pixel is then selected to form an all-in-focus image and the depth of each pixel depends on which image the pixel is selected from. Depth from defocus (DFD) [6] requires a pair of images of the same scene with different focus setting. It estimates the degree of defocus blur, by which the depth of scene can be recovered providing the camera setting.

The objective of this paper is to investigate the possibilities of a simple method for acquiring the depth using the principles of the canonical stereo vision systems. The aim is to prove by the physical experiments that in the case when a real stereo visual system is not available, using conventional digital still camera it is possible to effectively determine the z coordinates to particular object points in a given scene. The main request is that the camera should have precise horizontal movement, high resolution and possibilities for adjusting the parameters of its optical system (focal length, zoom).

The paper is structured as follows. Section II briefly discusses the main principles of canonical stereo vision systems. In Section III are described the basic steps of the realized algorithm for determining the distance to objects in a scene. In Section IV are presented and discussed the results of experimental studies conducted with real scene images. In the last section are given some concluding remarks and directions of further research work.

2. THEORETICAL BASIS OF THE CANONICAL STEREO CONFIGURATION

Canonical stereo configuration is an ideal stereo camera configuration, where a scene point, P, is being projected into the image planes of a pair of stereo cameras. The images planes are co-planar, parallel to the baseline, and the horizontal axis of each image plane's coordinate systems are coincident each other.

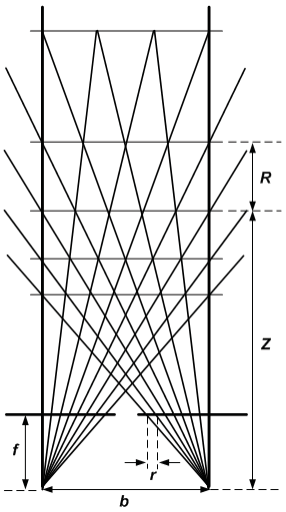
The problem of determining the distance to objects within the scene from pair stereo images obtained by such a canonical stereo configuration is to find the length of the horizontal displacement (disparity) between the two matched image points. For a given disparity match, the basic cameras distance, the focal length of the cameras and the pixel size of the cameras, it is easy to apply triangulation to find the distance to the scene points in world coordinates:

$$Z_i = \frac{bf}{(x_{li} - x_{ri}) * s_h}, \tag{1}$$

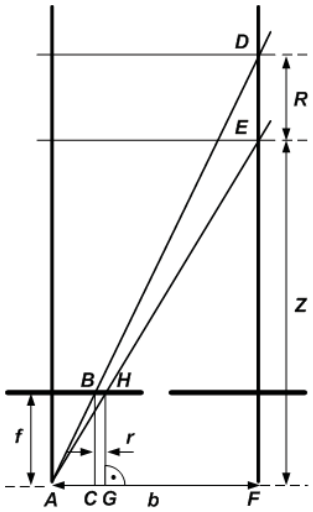
where: Z_i denotes the distance [in meters], between the line connecting the optical centres of the cameras and the scene point, P_i; s_h is the horizontal pixel size [in μm]; b is the distance [in cm], between the optical centres of the cameras; f is the focal length [in mm]; x_{li} and x_{ri} are distances [in pixels], measured from the top left corner of the image to the corresponding projections of arbitrary point, within the scene on the image planes.

The difference (disparity) $x_{li} - x_{ri}$ is always positive, since the projection of an arbitrary 3D point on the left image is located to the right of the projection of the same point on right image with respect to the origin of the image coordinate system.

The phenomenon of diminishing accuracy of depth measurement with increasing distance from the camera planes is shown on Fig. 1(a). This is a geometrical limitation since it depends exclusively on geometrical parameters of a stereo system [3].



(a) Phenomenon of a limited accuracy of depth measurement with increasing distance from the camera



(b) Relation of depth measurement accuracy in respect to camera resolution

Fig. 1: Depth resolution of a canonical stereo setup

The dependence of the depth accuracy versus camera resolution and distance to the observed scene can be found analysing Fig. 1(b). Observing the similarity of triangles ΔABC and ΔADF, as well as triangles ΔAEF and ΔAHG, the following formula can be derived to calculate the depth measurement resolution value, R

$$R = \frac{rZ^2}{fb - rZ}$$

Assuming that fb/Z is much larger than the pixel resolution r the following approximation can be established, which is justified for relatively small values of Z (Fig. 1(a)):

$$R \approx \frac{rZ^2}{fb} \tag{2}$$

For most image acquisition systems, the values of r, b and f are constant, at least for a single acquisition. This means that there is such a value Z for which it is not possible to measure the depth of the observed scene due to geometrical limitations of the stereo camera setup.

3. DEPTH ESTIMATION SEQUENCE AND METHODS USED IN THE STUDY

In the standard stereo case the two camera images just have a simple horizontal shift between them. The distance to objects is then determined calculating the resulting disparity of the corresponding pixels in both images. This disparity corresponds to the depth of objects in the space.

The depth estimation process performs the following basic steps:

***Step 1*:** Detection of distinctive invariant image features to perform reliable matching between the two camera views of an object or scene. [2, 4];

***Step 2*:** Searching for matches between detected image features [3, 8].

Step 3: Computation of the horizontal distance between each pair of corresponding feature points in the images and determine the disparity values.

***Step 4*:** Depth estimation from disparity using the relationship between disparity and depth obtained on the base of the geometrical model of canonical stereo configuration [3].

In the present study a feature-based approach, which relies on specifying so called cornerness measure. The most popular one is that defined by Harris and Stephens [4]:

$$C(x,y) = \det(M) - k\left(trace(M)\right)^2 \tag{3}$$

where $M = \sum_{x,y} W(x,y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}$ is 2x2 matrix from image derivatives, and $W(x,y)$ is a Gaussian window function; k empirical coefficient ($k = 0.04 \div 0.06$)..

The use of this metric for practical purposes is often related with troubles with specifying the appropriate value of *k*. To overcome this problem, an alternative cornerness measure [2] is adopted in our study:

$$C(x,y) = \det(M) / trace(M) \tag{4}$$

The next step of the depth estimation algorithm is to find the correspondent corner pair in the images. A feature-based matching (as part of the local stereo matching methods) is adopted for this stage. It consists of measuring the degree of correlation between beforehand detected corners in matched images. For best match evaluation, instead of comparing single pixels, groups of pixels (neighborhood N) are taken

simultaneously for comparison. In the case of horizontally shifted images, the search can be restricted to horizontal displacements only.

The metric used here is that based on Normalized Cross-Correlation (NCC):

$$NCC = \frac{\sum_{(i,j)\in U} I_1(x+i, y+j) \,.\, I_2(x+d_x+i, y+d_y+j)}{\sqrt{\sum_{(i,j)\in U} I_1(x+i, y+j)^2 .\, \sum_{(i,j)\in U} I_2(x+d_x+i, y+d_y+j)^2}}, \quad (5)$$

where: I_1 and I_2 are two image regions being compared. The region I_1 is built around a reference point (x,y), and the region I_2 - around point $(x+d_x, y+d_y)$, where with d_x and d_y are denoted the relative horizontal and vertical displacements of the two image blocks being compared. The matching regions are defined by a set U of offset values, measured from their reference points, i.e. (x,y) and $(x+d_x, y+d_y)$, respectively.

The last steps of the distance estimation process require determining the disparity using the corners correspondences and then performing disparity-to-depth conversion, based on the relationship defined by Eq. 1. The displacement length varies depending on the distance at which is a point in the 3D space. Less distance meets larger displacement and vice versa, i.e. the relationship between the distance and the horizontal displacement between the corresponding points is inversely proportional.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

Our experimental work has two goals: ***(i)*** to verify the applicability of the mathematical model to the practical camera system we use and ***(ii)*** to test the evaluation accuracy of the estimated distance in a real scene.

Experiments are conducted using digital still camera Olympus E-P2 (resolution 4032x3024, pixel size 4x4 μm) mounted on a platform that provides the opportunity for horizontal translation to specific distances (Fig. 2). For the purposes of the study, several experiments with static and partially structured scenes with many different objects of different distances are conducted. Each of the experiments was made with series of 6 photos: the first one - on chosen starting position, while for the remaining five the camera shifts to different base lengths - from 6 to 10 cm with step of 1 cm. During the experiments, besides the change of the starting position, the focal length was changed too. The images were taken at maximum camera resolution and with maximum image quality. The real distances to objects from the scene (Fig. 5) was previously measured by laser distance meter Leica Disto^{TM}D3 with measurement precision ± 1 mm (Fig. 3), in order to be used as ground truth data to determine the evaluation accuracy (Fig. 6).

Fig. 2: Experimental Platform

Fig. 3: Laser distance meter

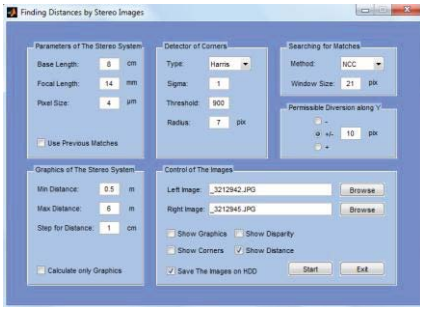


Fig. 4: Software testbed

A software testbed for the purposes of experimental research was developed (Fig. 4), which allows various parameters of the investigated algorithms to be specified (for example, the basic distance between a pair stereo images, the focal length, the pixel size,

the minimum and the maximum distance at which to look for matches, the Harris corner detector settings, the size of the correlation window, and etc.) during the various stages of the depth estimation process.

Two types of experimental images are analyzed: images obtained at maximum camera resolution and images reduced in size, twice. The aim of performing reduction in image size was to check for any improvements of the matching process (in addition to speeding up the calculations), since the most scaling algorithms, perform smoothing of the image and thus reducing the image's high-frequency components and the image noise [9]. At double less resolution, however, the step between two adjacent measured distances is greater, as shown in Fig. 1. Therefore, even differences of ± 1 pixel in the estimation of the disparity values can lead to much larger uncertainty at greater distances. For example, in the case of image reduced in size, twice (this corresponds to twice as large pixel dimensions, 8x8 μm in our case), the step between two measured distances is ± 18 cm at 5 m (Fig. 7), while at real image size this step is ± 9 cm.

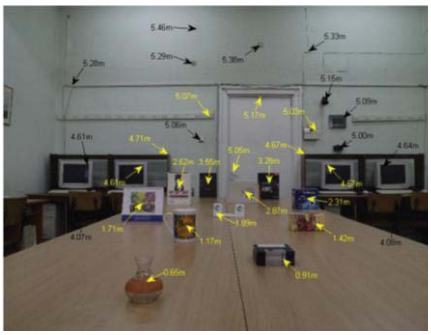


Fig. 5: Real Distances to selected objects

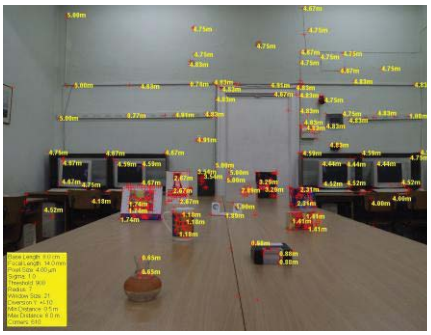


Fig. 6: Estimated distances for base 8 cm

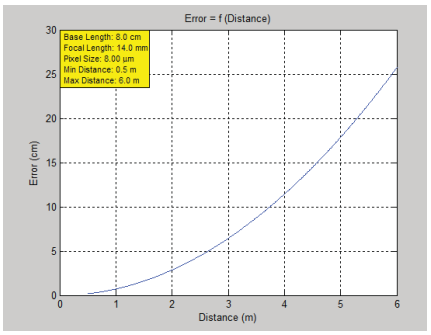


Fig. 7: Depth resolution error for base 8 cm

Table 1: Comparison between real and estimated distances for various base lengths.

Real Distance	Base length between cameras with focal length 14 mm									
	6cm		7cm		8cm		9cm		10cm	
	Est.	**Error**	**Est.**	**Error**	**Est.**	**Error**	**Est.**	**Error**	**Est.**	**Error**
0.65m	0.65m	0.00%	0.65m	0.00%	0.65m	0.00%	0.65m	0.00%	0.65m	0.00%
0.91m	0.88m	3.30%	0.88m	3.30%	0.88m	3.30%	0.88m	3.30%	0.88m	3.30%
1.17m	1.18m	0.85%	1.18m	0.85%	1.18m	0.85%	1.18m	0.85%	1.18m	0.85%
1.42m	1.42m	0.00%	1.42m	0.00%	1.41m	0.70%	1.42m	0.00%	1.42m	0.00%
1.71m	1.74m	1.75%	1.74m	1.75%	1.74m	1.75%	1.74m	1.75%	1.75m	2.34%
1.89m	1.91m	1.06%	1.90m	0.53%	1.90m	0.53%	1.90m	0.53%	1.90m	0.53%
2.31m	2.31m	0.00%	2.33m	0.87%	2.31m	0.00%	2.32m	0.43%	2.30m	0.43%
2.62m	2.66m	1.53%	2.66m	1.53%	2.67m	1.91%	2.67m	1.91%	2.67m	1.91%
2.87m	2.88m	0.35%	2.88m	0.35%	2.89m	0.70%	2.89m	0.70%	2.89m	0.70%
3.28m	3.28m	0.00%	3.31m	0.91%	3.29m	0.30%	3.28m	0.00%	3.30m	0.61%
3.55m	3.56m	0.28%	3.60m	1.41%	3.54m	0.28%	3.54m	0.28%	3.57m	0.56%
4.57m	4.47m	2.19%	4.54m	0.66%	4.52m	1.09%	4.50m	1.53%	4.49m	1.75%
4.61m	4.57m	0.87%	4.62m	0.22%	4.59m	0.43%	4.57m	0.87%	4.61m	0.00%
4.67m	4.57m	1.93%	4.62m	1.07%	4.59m	1.71%	4.57m	2.14%	4.61m	1.28%
4.71m	4.67m	0.85%	4.62m	1.91%	4.67m	0.85%	4.63m	1.70%	4.67m	0.85%
5.03m	4.88m	2.98%	4.80m	4.57%	4.83m	3.98%	4.85m	3.58%	4.79m	4.77%
5.05m	5.00m	0.99%	5.00m	0.99%	5.00m	0.99%	5.00m	0.99%	5.00m	0.99%
5.07m	4.77m	5.92%	4.90m	3.35%	4.83m	4.73%	4.77m	5.92%	4.79m	5.52%
5.17m	4.77m	7.74%	4.80m	7.16%	4.83m	6.58%	4.77m	7.74%	4.79m	7.35%

The experimental results obtained, show that at focal length of 14 mm, higher accuracy is observed at distances up to 4.5 m (the error is below 3%, see Table 1, Fig. 8), while, with increasing the distance, the error reaches to 8%. In increasing the focal length to 20 mm (Fig. 9), the error is less than 5% for shorter distances, and below 3% for those more than 4.5 m. The accuracy of distance estimation is highest for objects located in the center of the frame for both shown focal lengths, due to the lack of optical distortions.

The results obtained at the maximum camera resolution, and those obtained using images reduced in size, differ relatively small, but higher accuracy is achieved with maximal resolution, and therefore, it is preferable to operate on it.

The conducted series of experiments, varying the base length from 6 cm to 10 cm, demonstrated that there are no significant changes in the achieved accuracy (Table 1, Fig. 8-9). As the pixel size cannot be changed effectively to make estimation improvements, the parameter that can be mainly tuned is the focal length. Any other opportunities for improvement of the depth estimation accuracy are: using of a high spatial resolution sensor (a small pixel size is preferable); increasing the distance between the cameras; increasing the focal length.

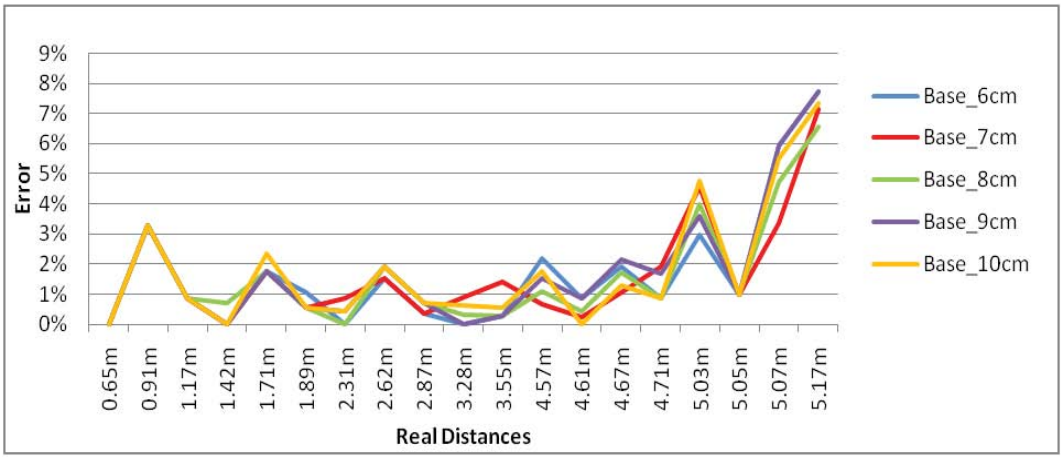


Fig. 8: Error of real vs. estimated distances for focal length equal to 14 mm

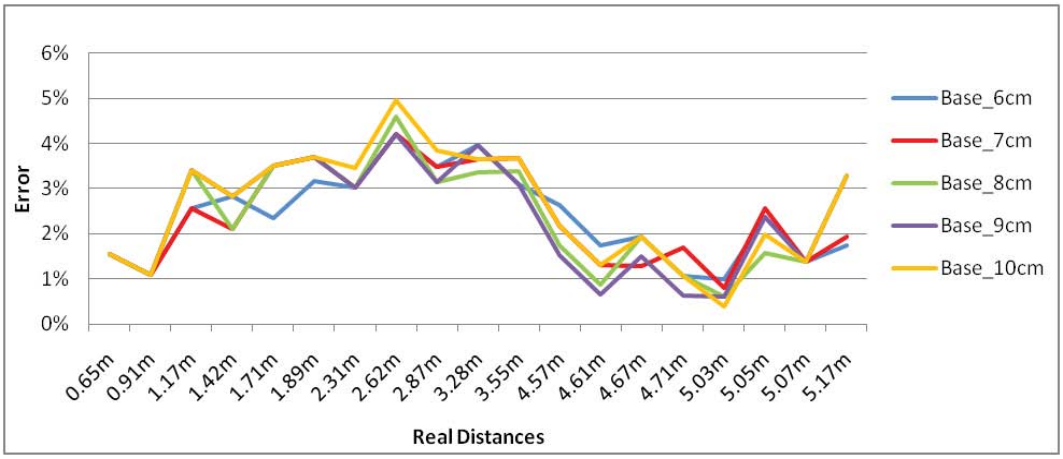


Fig. 9: Error of real vs. estimated distances for focal length equal to 20 mm

5. CONCLUSION AND FUTURE WORK

In this paper, the effectiveness of a simple approach for determining the distance to objects in a static scene, using the principles of the canonical stereo vision systems is investigated. The objective was to prove by physical experiments that despite the absence of real stereo camera, good enough results can be achieved using a conventional digital

still camera and applying classical techniques. The experimental results discussed demonstrate that the accuracy of the depth estimation is subject to the spatial quantization which is hard to control due to the restrictions on the pixel size. The large size is desirable because it leads to a high signal-to-noise ratio. In the other hand, the small size of the pixel is preferable since it results in a high spatial resolution which reflects in a high accuracy of depth estimation.

There are many difficulties in creating an accurate yet efficient algorithm for depth estimation (variables such as object texture, lighting, object foreshortening, and image noise), which are not considered in this study.

In our future work, we will continue to investigate other more sophisticated methods based on the stereo vision that provides a new perspective for matching problems while leading to efficient numerical solutions.

ACKNOWLEDGMENTS

This paper is partially supported within the project BG 051 PO 001-3.3.04/13 of Operational Program "Human Resources Development" 2007–2013, co-financed by the European Union (EU) through European Social Fund (ESF).

REFERENCES

[1] A. Verri, and V. Torre, "Absolute Depth Estimate in Stereopsis", *Journal of the Optical Society of America*. A 3, 1986, pp. 297-299.

[2] A. Noble, "Descriptions of Image Surfaces", *PhD thesis*, Department of Engineering Science, Oxford University, 1989.

[3] B. Cyganek, J. Siebert, "An Introduction to 3D Computer Vision Techniques and Algorithms", *John Wiley & Sons*, 2009.

[4] C. Harris and M. Stephens, "A Combined Corner and Edge Detector", *In. Proceedings of the 4th. Alvey Vision Conference*, 1988, pp. 147-151.

[5] N. Asada, H. Fujiwara, T. Matsuyama, "Edge and depth from focus", *International Journal of Computer Vision* 26 (2), 1998, pp. 153-163.

[6] P. Favaro, S. Soatto, "A Geometric Approach to Shape from Defocus"*, IEEE Trans Transactions on Pattern Analysis and Machine Intelligence* 27(3), 2005, pp. 406-417.

[7] P. H. S. Torr, A. Zisserman, "Feature Based Methods for Structure and Motion Estimation", *Lecture Notes in Computer Science*, *Vision Algorithms: Theory and Practice,* Springer Verlag, Volume 1883/2000, 2000, pp. 278-294.

[8] R. Szeliski, D. Scharstein, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms", *International Journal of Computer Vision*, Volume 47, 2002, pp. 7–42.

[9] S. Mihov, G. Zapryanov, "Interpolation Algorithms for Image Scaling", *In Proceedings of the Fourteenth International Scientific and Applied Science Conference - Electronics'2005*, September 21-23, Sozopol, Bulgaria, Volume 1, 2005, pp. 162-167.

ABOUT THE AUTHORS

Assist. Prof. Iva Nikolova, Department of Computer Systems, Technical University – Sofia, Phone: (+359 2) 965-26-80, E-mail: inni@tu-sofia.bg

Atanas Nikolov, Ph.D. student, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Tel: (+359 2) 979-29-25, E-mail: a.nikolov85@abv.bg

Assist. Prof. Georgi Zapryanov, Department of Computer Systems, Technical University – Sofia, Phone: (+359 2) 965-26-80, E-mail: gszap@tu-sofia.bg