3D Capturing with Monoscopic Camera

Miroslav GALABOV

Dept. of Computer Systems and Technologies, University of Veliko Turnovo, street T.Tarnovski 2, 5000 Veliko Turnovo, Bulgaria

lexcom@abv.bg

Abstract. *This article presents a new concept of using the auto-focus function of the monoscopic camera sensor to estimate depth map information, which avoids not only using auxiliary equipment or human interaction, but also the introduced computational complexity of SfM or depth analysis. The system architecture that supports both stereo image and video data capturing, processing and display is discussed. A novel stereo image pair generation algorithm by using Z-buffer-based 3D surface recovery is proposed. Based on the depth map, we are able to calculate the disparity map (the distance in pixels between the image points in both views) for the image. The presented algorithm uses a single image with depth information (e.g. z-buffer) as an input and produces two images for left and right eye.*

Keywords

3D content, multi-view camera, 3D capturing.

1. Introduction

With different types of cameras, the 3D content capturing process is completely different [1]. The stereo camera or depth camera simultaneously captures video and associated per-pixel depth or disparity information; multi-view cameras capture multiple images simultaneously from various angles, then a multi-view matching (or correspondence) process is required to generate the disparity map for each pair of cameras, and then the 3D structure can be estimated from these disparity maps. The most challenging scenario is to capture 3D content from a normal 2D (or monoscopic) camera, which lacks disparity or depth information.

Basically, typical *stereo cameras* use two cameras mounted side by side for the recording, although some variants may build them into one with two lenses.

Depth cameras refer to a class of cameras that have sensors that are able to measure the depth for each of the captured pixels using a principle called time-of-flight. It gets 3D information "by emitting pulses of infrared light to all objects in the scene and sensing the reflected light from the surface of each object." The objects in the scene are then ordered in layers in the z-axis, which gives a grayscale depth map that a game or any software application can use. The depth sensor consists of an infrared laser projector combined with a monochrome CMOS sensor, which captures video data in 3D under any ambient light conditions.

Commercial *multi-view camera* systems are rare in the market, although Honda Motor announced that it had developed a prototype multi-view camera system which displays views from multiple wide-angle CCD cameras on the vehicle's navigation screen to reduce blind spots, support smooth parallel or garage parking, and support comfortable and safe driving at a three-way intersection where there is limited visibility or on narrow roads. There are many multi-view cameras (or camera arrays) set up in labs for research efforts, where the synchronization among these cameras is conducted by gunlock devices.

2. Capturing with Monoscopic Camera

The major difference between a stereo image and a mono image is that the former provides the feel of the third dimension and the distance to objects in the scene. Human vision by nature is stereoscopic due to the binocular views seen by our left and right eyes in different perspective viewpoints. It is our brain that is capable of synthesizing an image with stereoscopic depth. In general, a stereoscopic camera with two sensors is required for producing a stereoscopic image or video. However, most of the current multimedia devices deployed are implemented within the monoscopic infrastructure.

In the past decade, stereoscopic image generation has been actively studied. In [2], the video sequence is analyzed and the 3D scene structure is estimated from the 2D geometry and motion activities (which is also called structure from motion, or SfM). This class of approaches enables the conversion of recorded 2D video clips to 3D; however, its computational complexity is rather high so that it is not feasible to use it for realtime stereo image generation. On the other hand, since SfM is a mathematically ill-posed problem, the result might contain artifacts and cause visual discomfort. Some other approaches first estimate depth information from a single-view image and then generate the stereoscopic views after that. In [3], a method for extracting relative depth information from

monoscopic cues, for example retinal sizes of objects, is proposed, which is useful for the auxiliary depth map generation. In [4], a facial-feature-based parametric depth map generation scheme is proposed to convert 2D head-and-shoulders images to 3D. In [5], an unsupervised method for depth-map generation is proposed, but some steps in the approach, for example the image classification in preprocessing, are not trivial and may be very complicated to implement, which undermines the practicality of the proposed algorithm. In [6], a real-time 2D-to-3D image conversion algorithm is proposed using motion detection and region segmentation; however, artifacts are unavoidable due to the inaccuracy of object segmentation and object depth estimation.

Clearly, all the methods mentioned above consider only the captured monoscopic images. Some other approaches use auxiliary sources to help generate the stereo views. In [7], a low-cost auxiliary monochrome or low-resolution camera is used to capture the additional view, and it then uses a disparity estimation model to generate the depth map of the pixels. In [8], a monoscopic high resolution color camera is used to capture the luminosity and chromaticity of a scene, and an inexpensive flanking 3D-stereoscopic pair of low resolution monochrome "outrigger" cameras are used to augment luminosity and chromaticity with depth. The disparity maps generated from the obtained three views are used to synthesis of the stereoscopic pairs. In [9–11], a mixed sets of automatic and manual techniques are used to extract the depth map (sometimes the automatic method is not reliable), and then a simple smoothing filter is used to reduce the visible artifacts of the result image.

In the following text, we introduce the new concept of using the auto-focus function of the monoscopic camera sensor to estimate depth map information [12], which avoids not only using auxiliary equipment or human interaction as mentioned above, but also the introduced computational complexity of SfM or depth analysis. The whole system design is novel, and is generic for both stereo image and video capture and generation. The additional but optional motion estimation module can help to improve the accuracy of the depth map detection for stereo video generation. The approach is feasible for low-power devices due to its two-stage depth map estimation design. That is, in the first stage, a block-level depth map is detected, and an approximated image depth map is generated by using bilinear filtering in the second stage. By contrast, the proposed approach uses statistics from motion estimation, auto-focus processing, and the history data plus some heuristic rules for estimating the depth map.

In Fig. 1, the proposed system architecture that supports both stereo image and video data capturing, processing, and display is shown. In the system, an image is first captured by a monoscopic camera sensor in the video front end (VFE), and then it goes through the auto-focus process, which helps to generate a corresponding approximated depth map. The depth map is further processed either using bilinear filtering for still-image or taking into account the motion information from the video coding process for video. After that, a depth-based image pair generation algorithm is used to generate stereo views. Clearly the 3D effect can be accomplished by choosing different display technologies such as holographic, stereoscopic, volumetric, and so on. In Fig. 2, the system architecture for still images is shown, which is simpler than the generic architecture in Fig. 1.

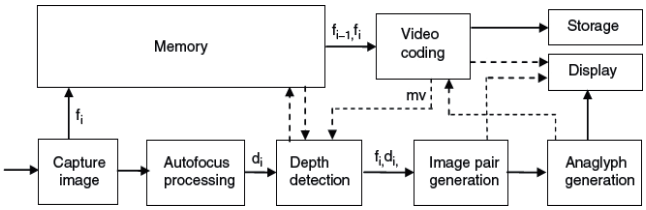

Fig.1. System architecture.

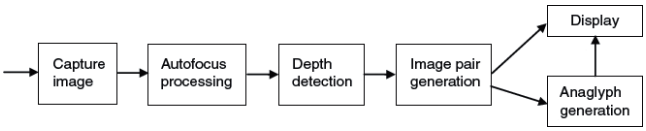

Fig. 2. System architecture for still images.

3. Depth Estimation by Autofocus Processing

In digital cameras, most focusing systems choose the best focus position by evaluating image contrast on the imager plane. Focus value (FV) is a score measured via a focus metric over a specific region of interest, and the autofocusing process of the camera normally chooses the position corresponding to the highest focus value as the best focus position of the lens. In some cameras, the high frequency content of an image is used as the focus value, for example, the high pass filter (HPF)

$$HPF = \begin{bmatrix} -1 & 0 & 0 & -1 \\ 0 & 4 & 0 & 0 \\ -1 & 0 & 0 & -1 \end{bmatrix}, \tag{1}$$

can be used to capture the high frequency components for determining the focus value.

It is important to know that there is a relationship between the lens position from the focal point and the target distance from the camera (as shown in Fig. 3), and the relationship is fixed for a specific camera sensor. Various camera sensors may have different statistics of such relationships. It means that once the autofocus process locates the best focus position of lens, based on the knowledge of the camera sensor's property, we are able to estimate the actual distance between the target object and the camera, which is also the depth of the object in the scene. Therefore, the proposed depth map detection relies on a sensor-dependent autofocus processing.

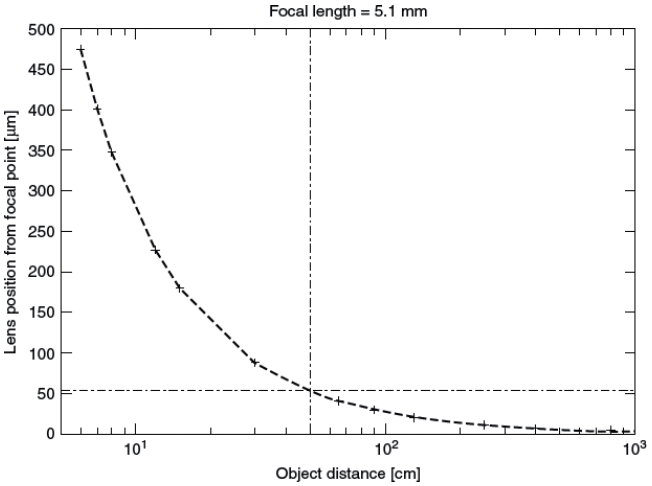

Fig. 3. Relationship between lens position from focal point and object distance.

4. Stereo Image Pair Generation

In this section, we propose a novel stereo image pair generation algorithm by using Z-buffer-based 3D surface recovery. The algorithm flowchart is shown in Fig. 4. We assume that the obtained image is the left view of the stereoscopic system; then, based on the depth map, we are able to calculate the disparity map (the distance in pixels between the image points in both views) for the image. Then a Z-buffer-based 3D interpolation process is called to construct a 3D visible surface for the scene from the right eye. Finally, the right view can be obtained by projecting the 3D surface onto the projection plane.

4.1 Disparity Map Generation

In Fig. 5, the geometry model of binocular vision is shown, where F is the focal length, L(x_L, y_L, 0) is the left eye, R(x_R, y_R, 0) is the right eye, T(x_T, y_T, z) is a 3D point in the scene, and P(x_P, y_P, F) and Q(x_Q, y_Q, F) are the projection points of the T onto the left and right projection planes. Clearly, the horizontal position of P and Q on the projection planes are ($x_P - x_L$) and ($x_Q - x_R$), and thus the disparity is $d = [(x_Q - x_R) - (x_P - x_L)]$.

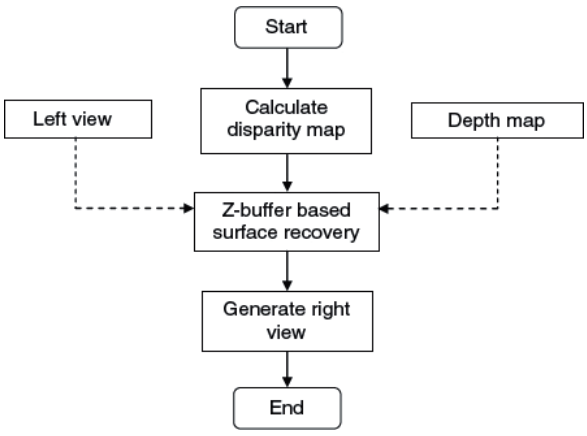

Fig. 4. Flowchart of the stereo image pair generation.

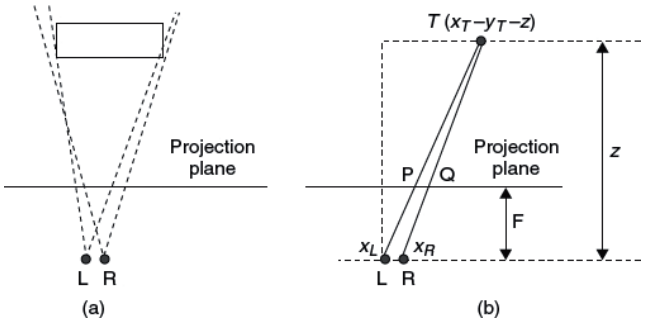

Fig. 5. Geometry model for binocular vision.

As shown in Fig. 5,

$$\frac{F}{z} = \frac{x_P - x_L}{x_T - x_L} = \frac{x_Q - x_R}{x_T - x_R}, \tag{2}$$

so

$$x_P - x_L = \frac{F}{z}(x_T - x_L), \tag{3}$$

$$x_Q - x_R = \frac{F}{z}(x_T - x_R), \tag{4}$$

and thus the disparity can be obtained as:

$$d = \frac{F}{z}(x_L - x_R). \tag{5}$$

Therefore, for every pixel in the left view, its counterpart in the right view is shifted to the left or right side by a distance of the disparity value obtained in (5). However, the mapping from left-view to right-view is not 1-to-1 mapping due to possible occlusions, therefore further process is needed to obtain the right-view image.

4.2 Z-Buffer-Based 3D Surface Recovering

We propose a Z-buffer-based 3D surface recovering algorithm for right-view generation. Since the distance between two eyes compared to the distance from eyes to the objects (as shown in Fig. 5) is very small, we can approximately think that the distance from object to the left eye is equal to the distance from itself to the right eye, which would greatly simplify the calculation.

In this method, we maintain a depth map $Z(x, y)$ for the right view where x, y are pixel positions in the view. Here the purpose is to reconstruct the 3D visible surface for the right view. At the beginning, the depth map is initialized as infinity. Then, for every pixel (x_0, y_0) in the left view with depth z_0 and disparity value d_0, we update the depth map for its corresponding pixel in the right view as:

$$Z(x_0 + d_0, y_0) = \min[Z(x_0 + d_0, y_0), z_0]. \tag{6}$$

After all the pixels in the left view are processed, we check the reconstructed depth map and search for the pixels with values equal to infinity (the pixels without a valid map on the left view). For such pixels, we first calculate its depth by 2D interpolation based on its neighbor pixels with

available depth value. After that, we find the disparity value following the computing using (5) and then inversely find its corresponding pixel in the left view. If the corresponding pixel is available, then the corresponding intensity value can be used on the right-view pixel; otherwise, we use interpolation to calculate the intensity value based on its neighbor pixels in the right view with available intensity values. It is important to point out that the benefit of using the proposed algorithm over the direct intensity interpolation method is that it considers the 3D continuity of the object shape which results in better realism for the stereo effect.

Clearly, the problem of recovering the invisible area of the left view is an ill-posed problem. In the solution of [13], [14], the depth of the missing pixel is recovered by using its neighbor pixel in the horizontal direction corresponding to a further surface with an assumption that there are no other visible surfaces behind in the scene. For some cases, the assumption might be invalid. To consider more possible cases, in the proposed solution, the surface recovering considers depths of all the neighbor pixels in all directions, which will reduce the chances of invalid assumption and will result in better 3D continuity of the recovered surface.

5. Experimental Results

In these experiments, we used stereoscopic display to demonstrate the resulting 3D effect. We calculate the stereo image pairs using different kinds of image depth map and generate the corresponding images. As shown in Fig. 6b, this is generated by using the approximated image depth map shown in Fig. 6a, and Fig. 6d is generated by using the accurate image depth map shown in Fig. 6c. The approximate time of an execution of the algorithm for image from Fig. 6 is the following (CPU: Intel Core i3 4130 3.4 GHz 512 kb, resolution of images: 320 × 200): source ray-traced frame 2.5 s, both warped frames 0.15 s.

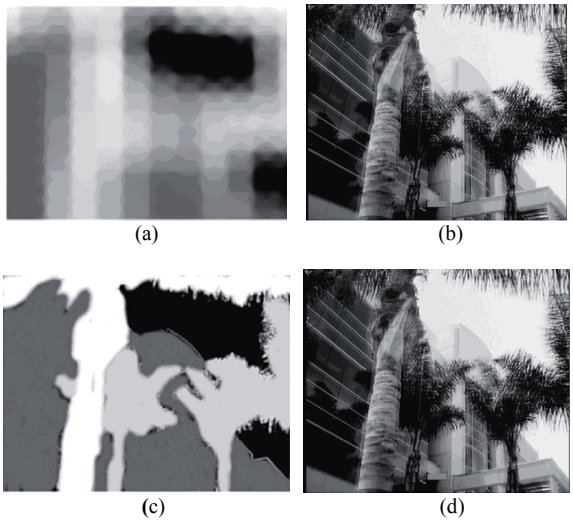

Fig. 6. Examples of the resulting stereoscopic image generated by using the different depth map.

The time of the ray-tracing depends directly on the complexity of the scene and on the contrary the time of the warping algorithm is independent of the scene complexity. Thus in the case of more complex scenes the efficiency gain is bigger. Because the process of generating of the stereo images involves a computation of a source frame with the original algorithm the time gain is at most 50 % of the original computation minus the time of processing of presented algorithm. In the case of ray-tracing the gain is just almost 50 %.

6. Conclusions

We propose a novel stereo image pair generation algorithm by using Z-buffer-based 3D surface recovery. We make experiments, in which we used stereoscopic display to demonstrate the resulting 3D effect. The results indicate that the approximated image depth map results in a similar image quality as using the accurate depth map, which proves the good performance of the proposed algorithm. The advantages of the presented method are the following: simplicity and speed that do not depend on the scene complexity.

All camera parameters need not to be known. It is sufficient to know only the resolution of images, horizontal angle of view and the distance from the camera to the plane of projection.

Acknowledgements

The presented article is part of research work carried out in the “Analysis, research and creation of multimedia tools and scenarios for e-learning” project - Contract No: RD - 09-590-12/10.04.2013, which is financially supported by the St. Cyril and St. Methodius University of Veliko Turnovo, Bulgaria.

References

[1] GUAN-MING SU, YU-CHI LAI, KWASINSKI, A., HAOHONG WANG. *3D Visual Communications.* John Wiley & Sons, 2013.

[2] JEBARA, T., AZARBAYEJANI, A., PENTLAND, A. 3D structure from 2D motion. *IEEE Signal Processing Magazine*, May 1999, vol. 16, no. 3, p. 66–83.

[3] XU, S. B. Qualitative depth from monoscopic cues. In *Proc. of Int. Conf. on Image Processing and its Applications*, Maastricht (The Netherlands), 1992, p. 437–440.

[4] WEERASINGHE, C., OGUNBONA, P., LI, W. 2D to pseudo-3D conversion of head and shoulder images using feature based parametric disparity maps. In *Proc. International Conference on Image Processing*. Thessaloniki (Greece), 2001, p. 963–966.

[5] BATTIATO, S., CURTI, S., CASCIA, M. L., TORTORA, M., SCORDATO, E. Depth map generation by image classification. *Proc. SPIE*, April 2004, vol. 5302, p. 95–104.

[6] CHOI, C., KWON, B., CHOI, M. A real-time field-sequential stereoscopic image converter. *IEEE Trans. Consumer Electronics*, August 2004, vol. 50, no. 3, p. 903–910.

[7] SETHURAMAN, S., SIEGEL, M. W. The video Z-buffer: a concept for facilitating monoscopic image compression by exploiting the 3D stereoscopic depth map. In *Proc. SMPTE International Workshop on HDTV'96*. Los Angeles (USA), 1996, p. 8–9.

[8] KIM, K., SIEGEL, M., SON, J. Y. Synthesis of a high-resolution 3D-stereoscopic image pair from a high-resolution monoscopic image and a low-resolution depth map. In *Proc. SPIE/IS&T Conference*, January 1998, vol. 3295A, p. 76–86.

[9] WANG, H., LI, H., MANJUNATH, S. Real-time capturing and generating stereo images and videos with a monoscopic low power mobile device. *US Patent*, 2012.

[10] KAMENCAY, P., BREZNAN, M., JARINA, R, LUKAC, P., ZACHARIASOVA, M. Improved depth map estimation from stereo images based on hybrid method. *Radioengineering,* 2012, vol. 21, no. 1, p. 70-78.

[11] KALLER, O., BOLEČEK, L., KRATOCHVÍL, T. Profilometry scaning for correction of 3D images depth map estimation. In *Proceedings of the 53rd International Symposium ELMAR- 2011.* Zadar (Croatia), 2011, p. 119-122. ISBN: 978-953-7044-12- 1.

[12] CURTI, S., SIRTORI, D., VELLA, F. 3D effect generation from monocular view. In *Proc. First International Symp. on 3D Data Processing Visualization and Transmission (3DPVT 2002)*. Padua (Italy), 2002, p. 550–553. DOI: 10.1109/TDPVT.2002.1024116.

[13] KOZANKIEWICZ, P. Fast algorithm for creating image-based stereo images. In *Proc. 10th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*. Plzen (Czech Republic), 2002.

[14] BATTIATO, S., CAPRA, A., CURTI, S., CASCIA, M. L. 3D stereoscopic image pairs by depth-map generation. In *Proc. 2nd International Symp. on 3D Data Processing, Visualization and Transmission*. Thessaloniki (Greece), 2004, p. 124–131. DOI: 10.1109/TDPVT.2004.1335185.

About Author...

Miroslav GALABOV was born in Veliko Turnovo. He received his M.S.E degree in Radio Television Engineering from the Higher Naval School N. Vapcarov, Varna, Bulgaria, in 1989. After that he worked as a design engineer for the Institute of Radio Electronics, Veliko Turnovo. From 1992 to 2001 he was an assistant professor at the Higher Military University, Veliko Turnovo. He received his Ph.D. degree in Automation Systems for Processing of Information and Control from the Higher Military University, in 1999. Since 2002 he has been an assistant professor and from 2005 he has been an associate professor in the Computer Systems and Technologies Department, St. Cyril and St. Methodius University of Veliko Turnovo. He is the author of ten textbooks, and over 40 papers. His current interests are in signal processing, 3D technologies and multimedia.