AUTOMATIC 2D-TO-3D VIDEO CONVERSION TECHNIQUE BASED ON DEPTH-FROM-MOTION AND COLOR SEGMENTATION

Lai-Man Po[1], *Xuyuan Xu*[2], *Yuesheng Zhu*[1,2], *Shihang Zhang*[1,2], *Kwok-Wai Cheung*[1,3] *and Chi-Wang Ting*[1]

[1]Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR, China
[2] Communication and Information Security Lab, Shenzhen Graduate School, Peking University, China
[3]Department of Computer Science, Chu Hai College of Higher Education, Hong Kong SAR, China

Abstract— Most of the TV manufacturers have released 3DTVs in the summer of 2010 using shutter-glasses technology. 3D video applications are becoming popular in our daily life, especially at home entertainment. Although more and more 3D movies are being made, 3D video contents are still not rich enough to satisfy the future 3D video market. There is a rising demand on new techniques for automatically converting 2D video content to stereoscopic 3D video displays. In this paper, an automatic monoscopic video to stereoscopic 3D video conversion scheme is presented using block-based depth from motion estimation and color segmentation for depth map enhancement. The color based region segmentation provides good region boundary information, which is used to fuse with block-based depth map for eliminating the staircase effect and assigning good depth value in each segmented region. The experimental results show that this scheme can achieve relatively high quality 3D stereoscopic video output.

***Keywords* - Depth from Motion, 3D-TV, Stereo vision, Color Segmentation.**

I. INTRODUCTION

In 2010, 3DTV is widely regarded as one of the next big things and many well-known TV brands such as Sony and Samsung were released 3D-enabled TV sets using shutter-glasses based 3D flat panel display technology. This commercialization of 3DTV [1] is another revolution in the history of television after color TV and high-definition digital TV. Basically, this revolution should be starting from 2005 after Disney's release of 3D version of *Chicken Little* in movie theaters, the industry rediscovered huge business potential of 3D video content. At the same time, the technologies of 3D displays and digital video processing have reached a technical maturity that possible for making cost effective 3DTV sets. However, the successful adoption of 3DTV by the general public will not only depend on technological advances, it is also significantly depend on the availability of 3D video contents. Although high quality 3D content does exist, this is generally not directly usable on a home 3DTV. This is simply because these content were designed to be viewed on a large screen and when viewed on a much smaller screen the left/right pixel disparities become too small that most of the 3D effect is lost. We believe that the conversion of monoscopic 2D videos to stereoscopic 3D videos is one way to alleviate the predicted lack of 3D content in the early stages of 3DTV rollout. If this conversion process can operate economically, and at acceptable quality, it could provide almost unlimited 3D content.

Basically, generation of 3D video from monoscopic 2D video input source [2-10] have been investigated for many years. Most of them are based on an estimated depth map of each frame and then using DIBR (Depth Image Based Rendering) [12] to synthesized the additional views. To estimate the depth maps, there are a number of manual techniques that are currently used such as hand drawn object outlines manually associated with an artistically chosen depth value; and semi-automatic outlining with corrections made manually by an operator. Such manual and semi-automatic methods could produces high quality depth maps but they are very time consuming and expensive. As a result, automatic 2D-to-3D video conversion techniques that can achieve acceptable quality are highly interested by both academic and industrial communities. Automatic solution can be easily implemented in a number of hardware platforms, such as notebook PCs and TVs.

In this paper, an automatic scheme using block-matching based depth from motion and color segmentation techniques is presented for synthesizing stereoscopic video from monoscopic video. The design principle and system structure will be presented in section II. The depth map generation and DIBR processes are described in sections III and IV, respectively. Experimental results are provided in section IV. Finally, a conclusion is given in section V.

II. 2D-TO-3D CONVERSION SYSTEM STRUCTURE

Stereoscopic video is relied on the illusion effect of the human eye. Due to small spatial displacement between the right-eye and left-eye views (horizontal disparities), the 3D perception is created in our brain. Thus, the main purpose of the 2D-to-3D stereoscopic video conversion system is to generate additional views from monoscopic video input. The basic structure of the proposed automatic 2D-to-3D video conversion system using block-matching based depth from motion estimation [7] and color based region segmentation is shown in Fig. 1.

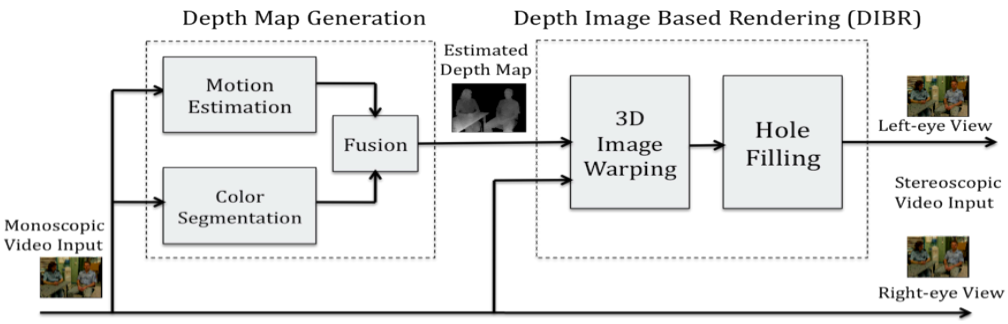


Fig. 1: System Structure of the Automatic 2D-to-3D Stereoscopic Video Conversion.

2.1 Synthesis View Selection

One of the main features of this system structure is that the input monoscopic video is used as the output right-eye view video of the synthesized stereoscopic 3D video and the left-eye view video is generated based on input video and the estimated depth map by the DIBR. This selection is mainly based on the 3D video quality and eye dominance characteristic of human perception. It has been known that human have a preference of one eye over the other and about 70% are right eyed, 20% left eyed, and 10% exhibit no eye preference. A recent study [11] found that the role of eye dominance have significant implications on the asymmetric view encoding of stereo views. These results suggest that the right-eye dominant population does not experience poor 3D perception in stereoscopic video with a relatively low quality left-eye view while right-eye view can provide sufficient good quality. On the other hand, the synthesized view of the 2D-to-3D video conversion based on DIBR always introduce distortion during the hole-fill process due to the disocclusion problem which lower the visual quality. Making use of about 70% right-eye dominance population, the proposed system therefore uses the original input video as the right-eye view and generates the left-eye view using DIBR for maintaining high quality right-eye view video.

III. DEPTH MAP GENERATION

To generate the left-eye view video, two key processes are involved: (1) Depth Map Generation and (2) DIBR as shown in Fig. 1. The depth map generation process is first introduced in this section.

3.1 Block-Matching Based Depth Map Estimation

Basically, depth map is an 8-bit grey scale image as shown in Fig. 2(b) for a 2D image frame of Fig. 2(a), in which grey level 0 indicates that furthest distance from camera and the grey level 255 specifying the nearest distance. To achieve good depth map quality in the proposed system, the depth map of each frame is first estimated by block-matching based motion estimation [7] and then fused with color based region segmented image. The basic principle is underlying on the motion parallax, near objects move faster than far objects, and thus relative motion can be used to estimate the depth map.

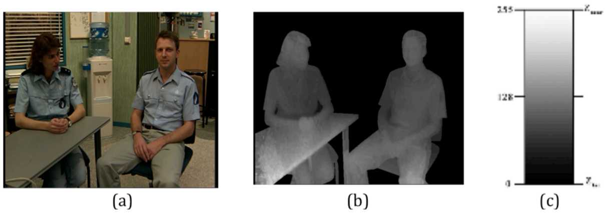
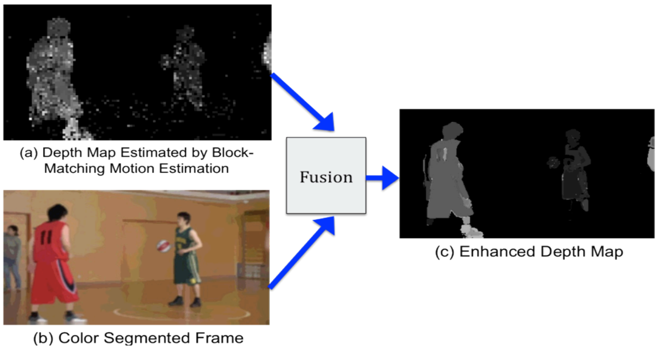


Fig. 2: (a) A frame of a monoscopic video, (b) the corresponding true depth map, (c) the grey-level of the depth values.



(a) Depth Map Estimated by Block-Matching Motion Estimation

(b) Color Segmented Frame

(c) Enhanced Depth Map

Fig. 3: Depth map enhancement by fusion with color segmented image.

The most practical way to implement this principle is to divide the 2D image frame into non-overlapping 4x4 blocks and then perform block-matching based motion estimation using the previous frame as reference. The depth value $D(i,j)$ are estimated by the magnitude of the motion vectors as follows:

$$D(i,j) = C\sqrt{MV(i,j)_x^2 + MV(i,j)_y^2} \qquad (1)$$

where $MV(i,j)_x$ and $MV(i,j)_y$ are horizontal and vertical components of the motion vectors and C is a pre-defined constant. One of the drawbacks of this method is that the computational requirement is very high if full-search method is used in motion estimation. To tackle this problem, fast motion estimation algorithm of cross-diamond search is used in the proposed system, which can achieve very similar depth map estimation accuracy while significantly reduce the computational requirement.

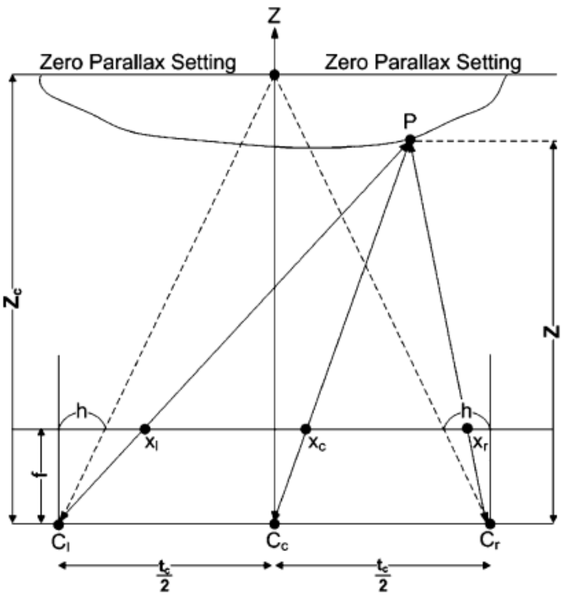


Fig. 4: Camera configuration for generation of virtual stereoscopic im ages.

3.2 Color Segmentation

The second drawback of the block-based depth estimation method is that the generated motion fields often suffer from serious staircase effect on the boundary of the objects or regions as shown in Fig. 3(a). To obtain better depth map, sophisticated region border correction technique is needed. In the proposed system, color based region segmentation is used. It is because it can provide important information of different regions that is the block-based motion depth map lacking of. Fusion with block-based depth map and color segmented image can eliminate blocking effect as well as reducing the noise. The adopted color segmentation involves two processes: (1) dominance colors generation by color quantization (CQ); and (2) regions segmentation by re-quantization. Agglomerative clustering with reducing quantization level is used for CQ which providing good trade-off on quality and computational efficiency. Based on this method, continue region with similar colors can be segmented. An example of segmented frame is shown in Fig. 3(b), that shows very smooth boundaries in difference regions and which is very effective for enhancing the blocky depth map.

3.3 Fusion

To enhance the block-based depth map as shown in the Fig. 3(a), it has to merged it with the color segmented image as shown in Fig. 3(b). This process is called fusion in this paper. The purpose of the fusion is to eliminate the staircase effort of the block-based depth map by using the good boundary information from the color segmented image. In addition, this fusion can also help on assigning better depth values in each region by using the average of the depth values within the same region. The fusion with average considerate the depth of whole part of the specify segmentation area. It takes average of the depth value from the motion estimation depth map in the area of the corresponding segmentation and assigned the value to the enhanced depth map. This process has a better estimation of the depth when there exist part of area with small or large depth value. The enhanced depth map is shown in Fig. 3(c).

IV. DEPTH IMAGE BASED RENDERING (DIBR)

To generate the stereoscopic 3D video, DIBR is used to synthesis the left-eye view video based on the estimated depth map and monoscopic video input as shown in Fig. 1. The DIBR algorithm consists of two processes: (1) 3D Image Warping and (2) Hole-filling.

4.1 3D Image Warping

The basic concept of 3D image warping can be considered as two steps. It first projects each pixel of the real view image into the 3D world based on the parameters of camera configuration and then re-project these pixels back to the 2D image of the virtual view for view generation. As shown in Fig. 4, left-eye and right-eye images at virtual camera positions C_l and C_r can be generated for a specific camera distance t_c with providing the information of the focal length f, and the depth Z from the depth map. The geometrical relationship as shown in Fig. 4 can be expressed as:

$$x_l = x_c + \frac{t_c f}{2Z(x_x, y)} + h \qquad (2)$$

$$x_r = x_c - \frac{t_c f}{2Z(x_x, y)} + h \qquad (3)$$

where h is equal to $h = -(t_c f)/(2Z_c)$ and Z_c is the distance between the camera and the Zero Parallax Setting (ZPS). Based on these equations, we can directly map the pixels in the right-eye view to the left-eye view in the 3D image warping process.

4.2 Hole-Filling

There are two major problems for the synthesized image by 3D image warping, which are called occlusion and disocclusion. Occlusion means that two different pixels of the real view image are warped to the same location in the virtual view. This problem is not difficult to resolve as it can use the pixels with larger depth values (closer to the camera) to generate the virtual view. The disocclusion problem is due to the occluded area in the real view may become visible in the virtual view. The disocclusion problem, however, is difficult to resolve. It is because there is no information provided to generate these pixels. As the result there are some empty pixels (holes) created in the virtual view as shown in Fig. 5. Thus, a hole-filling process is required in DIBR to fill out the area lacking of data. Linear interpolation is adopted in the proposed system but it will introduce stripe distortion as shown in Fig. 6 in large holes. To minimize the effect of stripe distortion on the generated stereoscopic video's depth perception for right-eye dominance population, the proposed system uses the input video as the right-eye view and only the left-eye view is synthesized with such distortion.

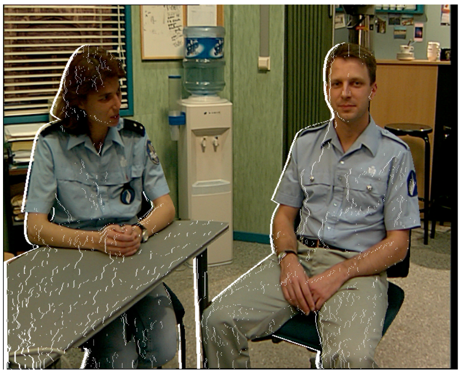


Fig. 5: Left-eye view image created by 3D image warping with holes due to disocculsion.

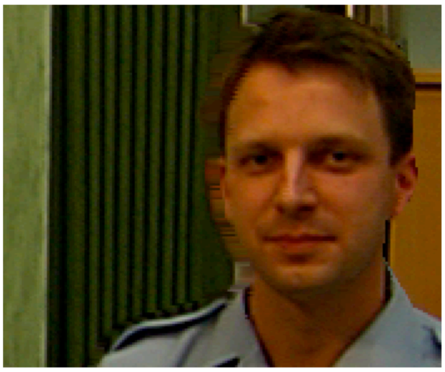


Fig. 6: Enlarged left-eye view image with stripe distortion after linear interpolation based hole-filling.

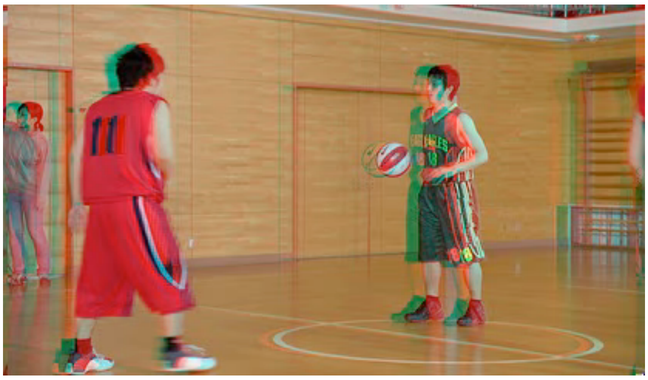


Fig. 7: Generated stereoscopic 3D video in anaglyph format.

IV. EXPERIMENTAL RESULTS

The proposed 2D-to-3D stereoscopic video conversion scheme is implemented on the MS-Windows platform for off-line automatic conversion. Several test sequences are used to evaluate the quality of the generated stereoscopic 3D videos. The subjective evaluation was performed and it is found that the 3D perception of the generated video is relatively good especially for video with a lot of object motions. Fig. 7 shows one of the test video sequences for basketball in anaglyph format, which achieve very good 3D video quality in terms of senses of stereo, reality, and comfortability. However, the major drawback of this scheme is that the computational requirement is very high which is not suitable for real-time applications.

V. CONCLUSION

This paper presents a robust 2D-to-3D stereoscopic video conversion system for off-line automatic conversion application. To make use of the right-eye dominance population and reduce the impact of the stripe distortion introduced in hole-fill of the DIBR, the input video is used as the right-eye view of the output stereoscopic video and the left-eye view is generated by block-matching based depth from motion estimation with color segmentation enhancement. The experimental results show that the proposed conversion scheme can yield satisfactory results.

ACKNOWLEDEMENT

The work described in this paper was substantially supported by a GRF grant with project number of 9041501 (CityU 119909) from City University of Hong Kong, Hong Kong SAR, China.

REFERENCES

[1] M. Op de Beeck, and A. Redert, “Three Dimensional video for the Home,” Proceedings of International Conference on Augmented, Virtual Environments and Three-Dimensional Image, May - June 2001, pp. 188-191.

[2] M. Kim and et al., “Stereoscopic conversion of monoscopic video by the transformation of vertical-to-horizontal disparity,” SPIE, vol. 3295, Photonic West, pp. 65-75, Jan. 1990.

[3] K. Man Bae, N. Jeho, b. woonhak, S. Jungwha, H. Jinwoo, “The adaptation of 3D stereoscopic video in MPEG-21 DIA,” *Signal Processing: Image Communication*, vol. 18(8), pp. 685-697, 2003.

[4] K. Manbae, P. Sanghoon, and c. Youngran, “Object-Based Stereoscopic Conversion of MPEG-4 Encoded Data,” *Lecture Notes in Computer Science*, vol. 3, pp. 491-498, Dec. 2004.

[5] P. Harman, J. Flack, S. Fox, M. Dowley,” Rapid 2D to 3D Conversion,” *Proceedings of SPIE*, vol. 4660, pp. 78-86, 2002.

[6] L. Zhang, J. Tam, D. Wang, “Stereoscopic image generation based on depth images,” *IEEE Conference on Image Processing*, pp. 2993-2996, Singapore, Oct. 2004.

[7] I. Ideses, L.P. Yaroslavsky, B. Fishbain, “Real-time 2D to 3D video conversion,” *Journal of Real-Time Image Processing*, vol. 2(1), pp. 2-9, 2007.

[8] M. T. Pourazad, P.Nasiopoulos, and R. K. Ward, “Converting H.264-Derived Motion Information into Depth Map,” *Advances in Multimedia Modeling*, volume 5371, pp. 108-118, 2008.

[9] Y.L. Chang, C.Y. Fang, L.F. Ding, S.Y. Chen and L.G. Chen, “Depth Map Generation for 2D-to-3D conversion by Short-Term Motion Assisted Color Segmentation,” *Proceeding of 2007 International Conference on Multimedia and Expo*, pp. 1958-1961, July 2007.

[10]F. Xu, G. Fr, X. Xie, Q. Dai, “2D-to-3D Conversion Based on Motion and Color Mergence,” *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, pp. 205-208, May 2008.

[11] Hari Kalva, Lakis Christodoulou, Liam M. Mayron, Oge Marques, and Borko Furht, "Design and evaluation of a 3D video system based on H.264 view coding," Proceeding of the 2006 International Workshop on Network and Operating System Support for Digital Audio and Video, Newport, Rhode Ishland, Nov., 2006.

[12] W. J. Tam, f. Speranza, L. Zhang, R. Renaud, J. Chan, C. Vazquez, “Depth Image Based Rendering for Multiview Stereoscopic Displays: Role of Information at Object Boundaries,” *Three-Dimensional TV, Video, and Display IV*, vol. 6016, pp. 75-85, 2005.