2D to 3D Conversion in 3DTV Using Depth Map Generation and Virtual View Synthesis

Cheolkon Jung[1], Xiaohua Zhu[1], Lei Wang[1], Tian Sun[1], Mingchen Han[2], Biao Hou[1], and Licheng Jiao[1]

Abstract. 2D to 3D conversion is an important task in 3DTV due to the lack of 3D contents. In this paper, we propose a novel framework of the 2D to 3D video conversion. The proposed framework consists of two main stages: depth map generation and virtual view synthesis. In the depth map generation, motion and relative-height cues are effectively used to generate depth maps. In the virtual view synthesis, depth-image-based-rendering (DIBR) is adopted to generate the left and right virtual views from the depth maps. Experimental results demonstrate that the proposed 2D to 3D conversion is very effective in generating depth maps and providing realistic 3D effects.

Keywords: 2D to 3D conversion, 3DTV, depth-image-based-rendering, motion parallax, depth map generation, relative height, virtual view synthesis.

1 Introduction

3DTV provides realistic 3D effects to viewers by employing stereoscopic contents compared with 2D videos. This technology can be used in various applications, including games, education, films, etc. Hence, 3DTV is expected to have the dominant market of the next generation digital TV. However, the promotion of 3DTV is constrained by the lack of stereoscopic contents. There are several approaches for generating stereoscopic contents. It is a common way that the 3D videos are captured by stereoscopic cameras which is a type of camera with two or more lens. Stereoscopic contents are also generated from monocular 2D videos. It is well-known that there is no direct depth information in the conventional 2D contents. The aim of this work is to convert 2D videos into 3D videos. The 2D to 3D conversion technology can enrich stereoscopic contents, and promote the de-

[1] C. Jung (✉), X. H. Zhu, L. Wang, T. Sun, B. Hou, L. C. Jiao
Xidian University, Xi'an 710071, China
e-mail: zhengzk@xidian.edu.cn

[2] M. C. Han
Huawei Technologies, Shenzhen 518129,China

velopment of the stereoscopic display industry. Thus, it attracts a lot of researchers and technical companies' attention now. To meet the consumer's demands, many companies are developing 2D to 3D conversion systems. For example, there are 2D to 3D conversion products such as DDD's TriDef, 3D player, and 2D to 3D software embedded in LG's 3DTV. The visual ability to perceive 3D scenes is related to human depth perception, and arises from a lot of depth cues. The depth cues are mainly classified into binocular and monocular ones. Binocular cues require that the input is from both eyes including stereopsis while monocular cues provide depth information by one eye using focus/defocus and geometric cues such as linear perspective, known size, relative size or height in picture, interposition, and texture gradient. In depth from motion cues, video sequences provide motion parallax for perceiving depth information [1]. Motion parallax is one of the common depth cues in the 2D to 3D video conversion technology. Depth from motion information is useful for generating depth maps. Kim et al. proposed a stereoscopic video generation method using stereoscopic display characterization and motion analysis [2]. Motion vectors are calculated by feature tracking and mean-shift image segmentation. In the method, three cues are used to convert motion vectors into disparity. The three cues are magnitude of motion, camera movements, and scene complexity. Pourazad et al. proposed a H.264-based scheme for the 2D to 3D video conversion [3]. This method characterized that motion vectors are directly extracted by decoding videos in compression domain, and a non-linear model is adopted between the motion of the objects and their distance from camera. It is very suitable for real time applications. Lai and Xu proposed a 2D to 3D conversion method based on object extraction [4, 5]. They first separated the image into foreground objects and the background, and then refined the foreground object by using gradient vector flows. Depth values were assigned to the foreground objects according to the motion analysis. In the background, depth values were assigned by using the linear perspective. Relative-height was one type of well-known depth recovery cues, especially in landscape scene. Jung et al. proposed a 2D to 3D conversion technique based on the relative-height depth cue [6]. The perceptual depth information from monocular images was estimated by the optimal use of relative-height. They tracked the line according to the strong edge, and then separated the image into several height layers. Depth assignment operation was followed to generate the initial depth map. The advantage of relative-height was that it could be utilized in the majority of the scene and did not need large computation. In the 3D visualization, many methods have been proposed. The European Information Society Technologies (IST) project 'Advanced Three-Dimensional Television System Technologies' (ATTEST) proposed depth-image-based-rendering (DIBR) technology. One or more virtual views of the real world could be generated by utilizing the color image with the corresponding depth map. DIBR was adopted by many global companies and organizations because of its advantage. Inspired by the previous work, we propose a novel 2D to 3D conversion framework which consists of two stages: depth map generation and virtual view synthesis.

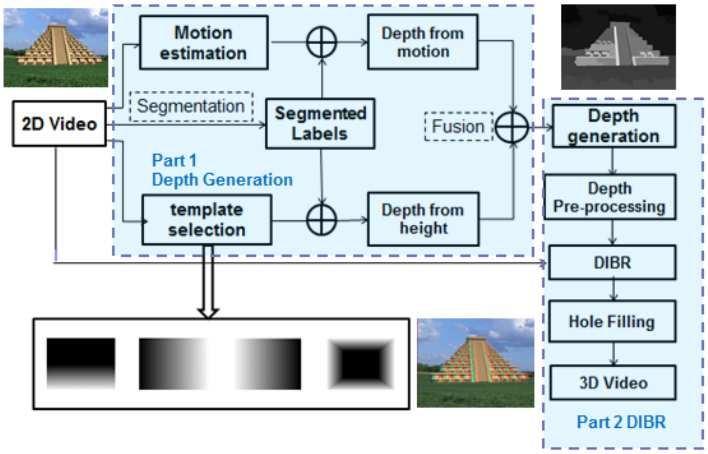

Fig. 1 The proposed 2D to 3D conversion framework.

In the first stage, motion parallax and relative-height are utilized to generate depth information from the monocular 2D videos. Two depth maps are individually generated from motion and relative-height, and then they are elaborately fused into the final depth map. In the second stage, DIBR is adopted to synthesize the left and right virtual views. The holes filling module is also added in the 2D to 3D conversion system. The framework of the proposed method is illustrated in Fig. 1.

2 Proposed Method

2.1 Depth Map Generation

First, image segmentation is conducted to separate images into several regions which are of the same color and texture. We adopt J-measure-based-segmentation (JSEG) for image segmentation. The process of JSEG considers not only colors but also textures. The process of JSEG is to separate the color image by two steps: color quantization and spatial segmentation [7]. In the first step, colors in the image are quantized to several representative classes, and the image pixel values are replaced by the corresponding color class labels, then forming a class-map of the image. In the second step, the main focus of this work is on spatial segmentation, where a criterion for good segmentation using the class-map is proposed. Then, a region growing method is used to segment the image based on the multi-scale J-images. After image segmentation, the original image is separated into several regions. We assign a label to each region, and thus both the processes of depth maps from motion parallax and relative-height are based on the label map. Motion parallax is one type of binocular depth cues. Because a video provides commonly motion parallax between the adjacent frames, motion parallax is one of the most

commonly used depth cues in 2D to 3D video conversion. Nearby things pass quickly, while far off objects appear stationary. Thus, it is reasonable that near objects move faster across the retina than far objects. This is usually called the principle of depth from motion parallax [1]. According to the principle, dense motion vectors are estimated for calculating the depth map. Generally, the motion estimation can be classification into two categories [8-11]: feature matching and block matching. In our work, the motion between two consecutive frames is estimated by feature matching based on the Kanade-lucas-tomasi (KLT) tracking. The algorithm of depth from motion parallax is summarized as follows:

Step 1: The JSEG segmentation separates the image into several similar color and texture regions;

Step 2: Each segmented region is labeled by a number. Then we can get a label map of the segmented image in the labeling stage. It is used for distinguishing the region from the other regions;

Step 3: After the above steps, the feature points are extracted by detect the change of the intensity;

Step 4: Feature tracking method are tracking the extracted feature points, and then get the motion vectors of the features;

Step 5: Convert the motion vector to dense motion maps.

The relative-height information is an assistant depth cue for depth generation in our work. Depth from relative-height reconstructs the depth of the non-moving regions. Depth from relative-height overcomes the drawback that depth from motion parallax does not work for non-moving regions when the video is captured by the fixed camera. Relative-height in images denotes that the object locates at the bottom of the image plane is closer than the one at the top of the image [6]. Relative-height depth cue exists mostly in photographic image to enhance the depth perception [12]. In the depth from relative-height, the label map is obtained by image segmentation. In the label map, the labels of the pixels in one region are the same. Thus, both depth from motion parallax and depth from relative-height share the same label map based on image segmentation. This kind of design improves the quality of depth with low computational costs. In most cases, the bottom-to-up depth template is suitable for depth map generation. In the depth template, when the pixel locates at the bottom of the image, the closest depth is assigned. On the contrary, when the pixel locates at the highest position of the image, the farthest depth is assigned. Next, two depth maps are obtained by using motion parallax and relative-height. They are fused into one final depth map. In the depth fusion procedure, the final depth is calculated by using the following equations.

$$D(x,y) = W_m \times D_m(x,y) + W_h \times D_h(x,y) \quad (1)$$

$$W_m + W_h = 1 \quad (2)$$

where $D_m(x,y)$ is the depth value of pixel (x,y) generated from motion parallax; $D_h(x,y)$ is the depth value of pixel (x,y) generated from relative-height; W_m and W_h are the weights of depth from motion and height. Both W_m and W_h are manually set according to the video content. If the motion information is obvious, the value of W_m set larger one than W_h.

2.2 Virtual View Synthesis

To synthesize left and right virtual views from depth, we use depth-image-based rendering (DIBR). DIBR has a lot of advantages over stereoscopic video generation compared with the other 3D warping methods [13-14]. The principle of DIBR compatible with existing 2D display systems. The object is projected the position X_c in original center image plane, and it is projected the position X_l in left image plane and X_r in the right image plane [15]. The relationship between the positions in different image plane is described by the following equations:

$$X_l = X_c + \left(\frac{b}{2}\right) f/Z \tag{3}$$

$$X_r = X_c - \left(\frac{b}{2}\right) f/Z \tag{4}$$

where Z is the depth between the object and the image plan; f is the focal length; and b is the baseline between two virtual point; X_c, X_r and X_l are the position where the point is projected in the center image plane, right image plane, and left image plane, respectively. DIBR renders a virtual view of any nearby viewpoint from the pixels of the original image. However, holes, i.e., dis-occluded regions, appear in the virtual views. We simply fill the hole by using its neighboring pixels. The pixel values in the holes are calculated as follows:

$$s(x,y) = \frac{\sum_{v=-w}^{v=w}\{\sum_{\mu=-w}^{\mu=w} s(x-\mu,y-v)\times non\_hole\,(x-\mu,y-v)\}}{\sum_{v=-w}^{v=w}\{\sum_{\mu=-w}^{\mu=w} non\_hole\,(x-\mu,y-v)\}} \tag{5}$$

$$non\_hole(x,y) = \begin{cases} 0 & (x,y) \text{ is in hole} \\ 1 & (x,y) \text{ is not in hole} \end{cases} \tag{6}$$

where w is size of window.

3 Experimental Results

To verify the superiority of the proposed method, several tests are performed with several test image sequences. We use a PC with AMD Athlon X4 640 CPU and 2GB RAM. Three test image sequences of different scales and scenes are used for the experiments: 'Akko & Kayo' with 640×480, 'Mobile phone' with 720×540, and 'Temple' with 400×300. In (1), we set W_m to 0.7 and $W_{\hat{\imath}}$ to 0.3, respectively, as default values. The weights are adjusted in depth fusion step by experiments. Thus, if moving object regions are comparatively large, W_m should be set up by a large value. In the case of the 'Akko&Kayo' sequence, W_m is set to 0.8 because of the large motion. Fig. 2 shows the two depth maps generated from motion parallax and relative-height, and the final depth map fused by both depth cues. As shown in the figure, the quality of the final depth map is much better than that by single depth cue. The motion cue effectively reconstructs the depth of several moving objects, while the relative-height cue reconstructs that of the background regions. Fig. 3 shows the red/cyan stereoscopic images generated by the proposed method ( Red/cyan glasses are recommended to view this image correctly).

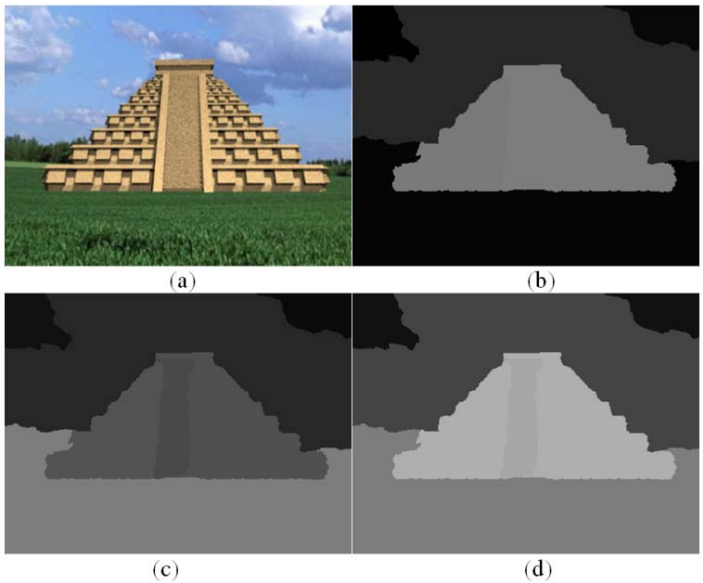

Fig. 2 Depth map generation results in the 'Temple' sequence. (a) Original image. (b) Depth map from motion parallax. (c) Depth map from relative-height. (d) Final fusion depth map.

Fig. 3 Red-cyan stereoscopic images (Red/cyan glasses are recommended to view this image correctly). (a) 'Mobile Phone'. (b) 'Akko & Kayo'. (c) 'Temple'.

As shown in the figure, the proposed method can generate demonstrable 3D stereoscopic views from 2D image sequences by the proposed method. It is well known that there are various objective quality measures for 2D images including peak signal to noise ratio (PSNR), structural similarity (SSIM), etc. However, the 2D quality measures are not suitable for the evaluation of the 3D stereoscopic views. Thus, the subjective quality assessment is conducted to evaluate the performance. We evaluate the three kinds of 3D stereoscopic videos generated by original color and depth images, by the 2D to 3D conversion software embedded in LG 3DTV 6700, and by the proposed method. The subjective evaluation is performed by 10 persons. The participants have watched all the stereoscopic videos twice. Then, they give scores considering three evaluation terms such as depth perception, sharpness distortion & color distortion, and visual comfort. The higher the score is, the better the quality of the video is. We calculate the average of the values of each term.

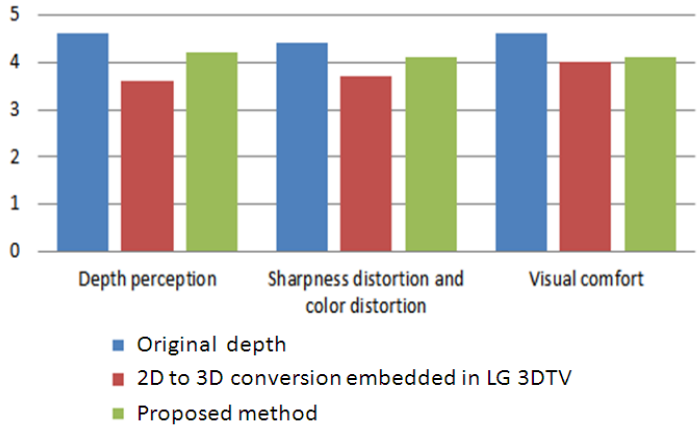

Fig. 4 Subjective evaluation results on the 3D stereoscopic videos.

As shown in Fig. 4, the 3D stereoscopic video generated by original color and depth images provides the highest score in the comparison. The stereoscopic video generated by proposed method produces better performance than the one generated by the 2D to 3D conversion software embedded in 3DTV. That is, compared to the stereoscopic videos generated by the embedded conversion software in 3DTV, the proposed method offers better depth perception rates.

4 Conclusions

We have proposed a novel 2D to 3D video conversion system for generating 3D contents from 2D videos. The proposed system utilizes motion parallax and relative-height as the depth cues. Depth from motion parallax provides obvious depth maps for the fast moving objects while depth from relative-height offers general depth maps of the scenes. The two depth maps are elaborately combined into the final depth map. Moreover, DIBR is employed for generating 3D stereoscopic videos from the depth maps. We have compared the performance of the proposed method with the embedded 2D to 3D conversion software in 3DTV by the subjective quality assessment. Experimental results show that the proposed method effectively produces 3D stereoscopic videos from 2D ones.

5 Acknowledgements

This work was supported in part by the Huawei innovation research program under Grant IRP-2011-03-04. This work was also supported by the National Basic

Research Program (973 Program) of China (No. 2013CB329402), the National Natural Science Foundation of China (Nos. 61271298 and 61050110144), the Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) (No. B07048), and the Program for Cheung Kong Scholars and Innovative Research Team in University (No. IRT1170).

6 References

1. Zhang L, Vazquez C, Knorr S (2011) 3D-TV content creation: automatic 2D-to-3D video conversion. IEEE Transactions on Broadcasting 99:1-12
2. Kim D, Min D, Sohn K (2008) A stereoscopic video generation method using stereoscopic display characterization and motion analysis. IEEE Transactions on Broadcasting 54:188-197
3. Pourazad MT, Nasiopoulos P, Ward RK (2008) An H. 264-based scheme for 2D to 3D video conversion. IEEE Transactions on Consumer Electronics 55:742-748
4. Lai YK, Lai YF, Chen YC (2012) An effective hybrid depth-perception algorithm for 2D-to-3D conversion in 3D display systems. In: Proc. IEEE ICCE, pp.612-613
5. Yu F, Liu J, Ren Y, Sun J, Gao Y, Liu W (2011) Depth generation method for 2D to 3D conversion. In: Proc. 3DTV-Con
6. Jung YJ, Baik A, Kim J, Park D (2009) A novel 2D-to-3D conversion technique based on relative height depth cue. In: Proc. SPIE Electronics Imaging, Stereoscopic Displays and Applications
7. Deng Y, Manjunath BS (2001) Unsupervised segmentation of color-texture regions in images and video. IEEE Transactions on Pattern Analysis and Machine Intelligence 23:800-810
8. Zhang L, Lawrence B, Wang D, Vincent A (2005) Comparison study on feature matching and block matching for automatic 2D to 3D video conversion. In: Proc. IEE European Conference on Visual Media Production, pp.1869-1872
9. Harris C, Stephens M (1998) A combined corner and edge detector. In: Proc. Alvey Vision Conference, pp. 147-152
10. Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: Proc. International Joint Conference on Artificial intelligence
11. Tomasi C, Kanade T (1991) Detection and tracking of point features. School of Computer Science, Carnegie Mellon Univ
12. Ostnes R, Abbott V, Lavender S (2004) Visualization techniques: An overview—Part 1. Hydrographic Journal 113:4–7
13. Fehn C (2004) Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In: Proc. SPIE
14. Park YK, Jung K, Oh Y, Lee S, Kim JK, Lee G, Lee H, Yun K, Hur N, Kim J, (2009) Depth-image-based rendering for 3DTV service over T-DMB. Signal processing: Image communication 24:122-136
15. Hong YR, Tseng YC, Chang TS (2010) Stereoscopic images generation with directional Gaussian filter. In: Proc. IEEE ISCAS, pp.2650-2653