Technical Challenges for Free-viewpoint Three-dimensional Television

Yo-Sung Ho and Eun-Kyung Lee
Gwangju Institute of Science and Technology (GIST)
261 Cheomdan-gwagiro, Buk-gu, Gwangju 500-712 Korea
Email:{hoyo, eklee78}@gist.ac.kr

***Abstract*-In recent years, the demand for three-dimensional television (3DTV) is growing rapidly. Since 3DTV is considered as the next generation broadcasting service that can deliver realistic and immersive experiences, a number of advanced 3D video processing techniques have been studied. Among them, multi-view video coding is the key technology for various applications including free-viewpoint video, free-viewpoint television, 3DTV, immersive teleconference, and surveillance systems.In this paper, after reviewing the basic techniques for 3D image capturing and 3D video display systems, we are going to cover technically challenging issues of free-viewpoint 3DTV. We also explain how togenerate multi-view video sequences using a hybrid camera system combining one time-of-flight depth camera and two video cameras.**

I. INTRODUCTION

Owing to the rapid growth of various digital technologies, broadcasting services [1] has been changed from unidirectional services to bidirectional services or interactive services, such as stereoscopic TV [2], three-dimensional (3D) TV [3], and realistic broadcasting [4]. As shown in Fig. 1, the next-generation broadcasting system is supposed to provide a variety of user-friendly interactive information, as well as high-quality audio-visual broadcasting contents.

Especially, 3DTV is considered as a main theme for the future broadcasting system supporting natural viewing experience in the true three dimension. In general, 3D natural views are usually created from two 3D video representations: multi-view video [5] and video-plus-depth [6]. A multi-view video represents the 3D scene with the collection of multiple videos generated by capturing the scene at different camera locations. Since the multi-view video produces natural 3D views with a number of images at the viewing position, we can be easily immersed in the 3D content. However, we need to put more efforts to control a huge number of cameras at the same time. Moreover, since the multi-view camera system usually requires complicated coding and transmission schemes in proportional to the number of cameras, it is hard to send its data to the receiver side within limited bandwidth channel environments.

As an alternative for the 3D video representation, it is widely accepted for a monoscopic color video enriched with per-pixel depth information, which is often referred as video-plus-depth data. Since the video-plus-depth representation includes depth information as geometry data of the scene, we can generate free-viewpoint images using depth image-based rendering (DIBR) techniques for the 3D video contents service. Although the video-plus-depth approach can support narrow-viewing angle views in comparison to the multi-view video, it is considered as a suitable 3D video representation for 3DTV because it can support both backwards compatibility to the current 2D digital systems and easy adaptability to a wide range of different 2D and 3D displays. Recently, the ISO/ICE JTC/SC29/WG11 Moving Picture Experts Group (MPEG) has also been interested in multi-view video with depth (MVD), which is the combination of the multi-view video and the video-plus-depth approaches, for free-viewpoint TV (FTV) and 3DTV [7].

With respect to the current 3DTV and FTV research activities, it is very important for us to estimate accurate depth information from the natural scene. In the field of computer vision and image processing, a number of depth estimation algorithms have been proposed to generate accurate depth maps. However, accurate measurement of depth information from the natural scene still remains problematic.

In general, there are two approaches to acquire depth information: depth from active sensor depth camera system and depth estimation from stereo matching. The latter takes a longer time and is more complex. In spite of its complexity, it does not guarantee accuracy of the estimated depth. On the other hand, as sensor technologies for obtaining depth information are developed rapidly, we can capture more accurate per-pixel depth information from the real scene directly using a depth camera system. However, the depth camera system has disadvantages: high cost and limited viewing range. Therefore, we need to develop a multi-view camera system to solve these problems.

To solve these problems, hybrid camera systems have been proposed to generate enhanced depth maps by applying a stereo matching algorithm to multi-view images with depthusing the depth information captured by the depth camera. However, these systems cannot produce high-resolution depth maps, because it completely depends on the low-resolution depth camera.

In this paper, we design a hybrid camera system with one depth camera and multiple video cameras. The proposed hybrid camera system can produce multi-view images for dynamic 3D scenes by enhancing the low-resolution depth information measured by the depth camera. The main contribution of our work is to provide a practical 3D video generation solution for dynamic 3D scenes, which is applicable to 3D consumer devices.

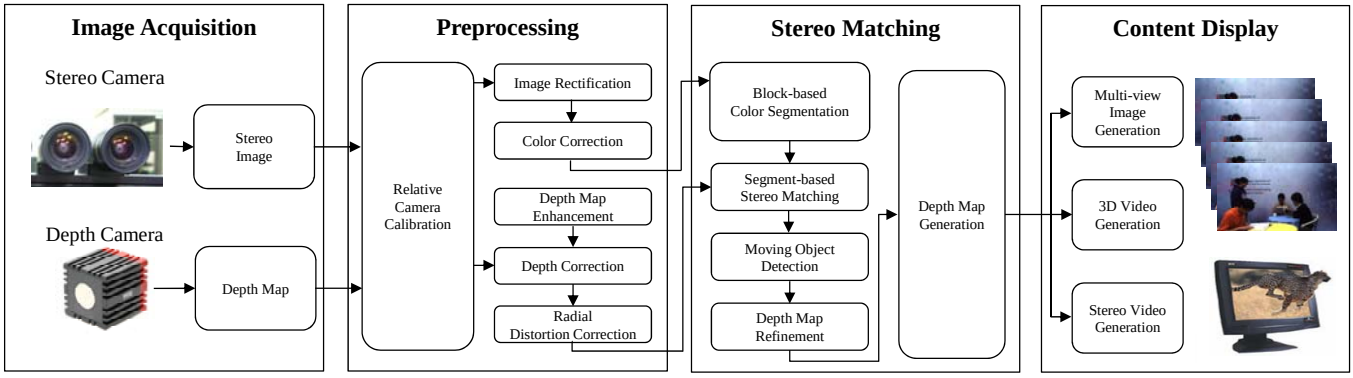


Figure 1.Overall framework of multi-view image generation

II. HYBRID CAMERA SYSTEM

The proposed hybrid camera system is composed of one depth camera and five high-definition(HD) videos. Those multiple video cameras are arranged in a one-dimensional array to construct a multi-view camera system. A clock generator sends synchronization signals constantly to each camera and its corresponding personal computer equipped with a video capture board. Basically, the proposed hybrid camera system captures multi-view images by the multiple video cameras and a depth map from the depth camera at each sampling time.

Figure 1 illustrates the overall framework to generate multi-view video sequences with their corresponding depth maps using the hybrid camera system. After calibrating each camera independently, we perform an image rectification to adjust vertical mismatches in multi-view images. Then, we apply a color correction operation to maintain color consistency among multi-view images. Figure 2 shows the proposed hybrid camera system and configuration.

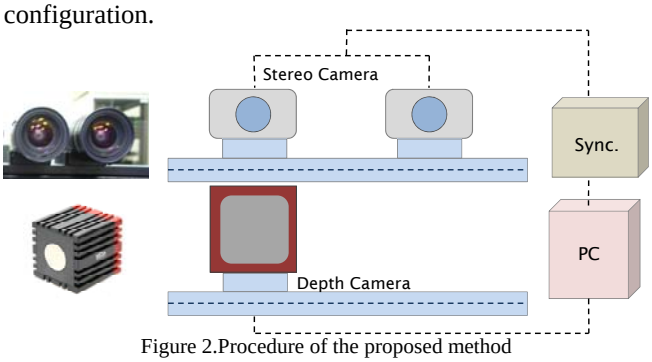


Figure 2.Procedure of the proposed method

To obtain depth maps for multi-view images, we perform a 3D warping operation onto each multi-view camera using the depth map measured by the depth camera. The warped depth data is used as an initial depth at each camera position. After we segment each multi-view image, we assign the depth value of the warped depth data in each segment as the initial depth of the segment. In order to improve the depth accuracy of object boundaries, we separate the moving objects and detect occlusion and disocclusion regions. Then, the depth of each segment is refined by a color segmentation-based stereo matching method. Finally, we obtain multi-view depth maps by conducting a pixel-based depth map refinement using a proposed cost function in each segment.

III. MULTIVIEW DEPTH MAP GENERATION

A. 3D Warping of Depth Map

We generate initial depth of the multi-view image by performing 3D warping of the depth values obtained from the depth camera. First, we project pixels of the depth map into the 3D world coordinate using the depth values. We then reproject the 3D points into each view.

Let us assume that $D_s(p_{sx}, p_{sy})$ is the depth intensity at the pixel position (p_{sx}, p_{sy}) in the depth map. $P_s(x_{sx}, y_{sy}, z_{sz})$ is a 3D point corresponding to D_s. The backward projection for moving D_s to the world coordinate is carried out by

$$P_s = K_s^{-1} \cdot p_s \qquad (1)$$

where K_s^{-1} indicates the intrinsic matrix of the depth camera. In the backward 3D warping, since rotation and translation matrices of the depth camera are the identity matrix I and zero matrix O, we have only to consider its intrinsic matrix. Thereafter, we project the 3D points P_s into the each view to get its corresponding pixel position p_k' (u_k, v_k) of the k^{th}-view image by

$$p_k' = \tilde{P}_k' \cdot P_s \qquad (2)$$

where Pk' indicates the projection matrix of the k^{th}-view video camera. Figure 3 shows the result of 3D warping using the acquired depth maps.

Figure 3.3D warped depth map

B. Segment-based Depth Estimation

We define the initial depth of each segment as 3D warped depths in the segment; the assumption is that each segment has one depth value. However, there is one

problem to set the initial depth using warped depth value. The 3D warping is performed from the small resolution depth map to the HD image in our system. Since there are many errors such as camera calibration error and depth error acquired from the depth camera, the warped result is not exactly matched with the HD image.

To obtain the accurate initial depth value, we use the warped results as multiple initial depth values for stereo matching. However, if the given initial depth is the error value, we could find wrong areas which has local minimum. Therefore, the assignment of the correct initial depth is crucial in using the depth camera. Because there are correct initial depths around the currently warped position, which are not exactly matched with the original image, we increase the candidates of the initial depth value to resolve this problem. By using the multiple initial depths, we can set initial depth for the depthless regions in the boundary of objects.

For determining the disparity of each segment, we calculate the mean of absolute difference (MAD) values between the segment in the current view image and its matched region in the left and right view images by

$$FG\_d_i(InitDisp) = \min(\min(\sum_{j=0}^{a} MAD(j)), \min(\sum_{k=0}^{b} MAD(k)) \quad (3)$$

where i is the index of the segment, j and k means index of the multiple initial depth. a and b are the number of the initial depth in the horizontal and vertical regions, respectively. FG_di(InitDisp) is the refined initial depth value from pairwise stereo matching. Search range to estimate disparities of the current view image is from InitDisp-5 to InitDisp+5. The disparity with the minimum MAD in the search range is chosen as the refined initial disparity of the segment in the current view image.

Since the acquired depth map is only for foreground regions, there is no depth information for background areas. In estimating depth of background, we set the minimum and maximum disparity value. We then find the minimum MAD as the initial disparity of the current segment in the background by

$$BG\_d_i(InitDisp) = \min(\sum_{i=\min Disp}^{\max Disp} MAD(i)) \quad (4)$$

whereBG_di(InitDisp) is the disparity for background, minDisp and maxDisp mean minimum and maximum disparity search range for background. The disparity with the minimum MAD is chosen as the initial disparity di(Initdisp) of the segment i in the current view image n by

$$d_i(InitDisp) = \min(FG\_d_i(InitDisp), BG\_d_i(InitDisp)) \quad (5)$$

D. Depth Refinement

In stereo matching, depth refinement usually enhances depth accuracy through iteration at the cost of long processing time, lots of memory requirement, and heavy computation. However, it has challenges when our target is to generate high-resolution 3D video based on multi-view depth maps. We therefore propose a simplified depth refinement approach using the proposed cost function for the depth map refinement, which has the following features: low memory consumption, fast processing time, and no iteration steps.

In order to enhance the multi-view depth map along the boundary of the objects, we refine it for two regions: moving region and static region. We have already defined the moving regions using color difference between frames. If there is no variance of a pixel in the time domain, we assume that pixel is static. In that case, we can refer the previous depth value for the static pixel. Otherwise, we just use the refined disparity value without referring the previous one.

$$E(x,y,d) = \begin{cases} w_s f_s(x,y,d_s(x,y)) + w_d f_d(x,y,d_d(x,y)) & if \;\; obj\_mov(x,y) = 1 \\ w_s f_s(x,y,d_s(x,y)) + w_d f_d(x,y,d_d(x,y)) + w_t f_t(x,y,d_t(x,y)) & if \;\; obj\_mov(x,y) = 0 \end{cases} \quad (6)$$

wherew_s, w_d, w_t are the weighting factors for depth refinement. $f_s(x, y, ds(x,y)$ is the smoothness term with gradient of the refined depth value in this refinement step. $f_d(x, y, d_d(x, y)$ is the data term for the refined initial depth value in the segment-based stereo matching step and $f_t(x, y, d_t(x,y)$ is the temporal term for depth value of the previous frame for the static pixel. From our experimental, the weighting factors of the cost function w_s, w_d, w_t are 0.3, 0.5, and 0.2. obj_mov(x,y) indicates the result of the moving object detection. If obj_mov(x,y) is 0, this pixel is not moved. Then, we can refer the depth value of the previous frame.

$f_d(x, y, d_d(x, y)$ means the minimum MAD with the refined initial depth value in the search range from InitDisp-5 to InitDisp+5. $f_s(x, y, d_s(x,y)$ is the depth difference with neighborhood depth in the same segment and calculated by

$$f_s(x,y,d_s(x,y)) = med(s_a(x,y), s_b(x,y), s_c(x,y)) \quad (7)$$

We can calculate the smoothness value as shown in Fig. 4. $s_a(x,y)$ is the refined depth difference at positions between (x -1, y-1) and (x-1, y). $s_b(x, y)$ is the refined depth difference at positions between (x-1, y-1) and (x, y-1). $s_c(x, y)$ is the refined depth difference at positions between (x, y-1) and (x+1, y-1).

The function med() takes the median value among arguments to avoid the wrong depth selection, so that it maintains depth continuity along the vertical and horizontal direction. If the selected smoothness gradient is a vertical direction, this depth difference is calculated from (x, y-1). Otherwise, the depth difference is computed from (x-1, y).

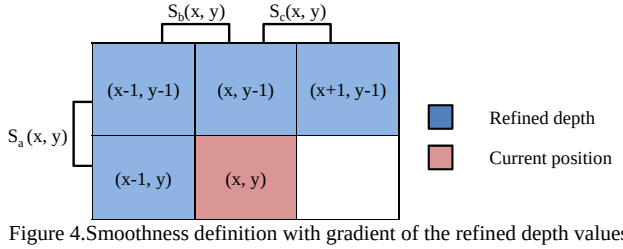


Figure 4.Smoothness definition with gradient of the refined depth values

IV. EXPERIMENTAL RESULTS

In order to generate the high-quality multi-view depth maps, we have constructed a hybrid camera system with five HD cameras and one depth camera. The measuring depth range of the depth camera is from 0.50m to 5.00m. The baseline distance among multi-view HD cameras are

6.5cm. Table 1 lists the specification of the hybrid camera system. Figure 5 shows the multi-view test sequences, Café, captured by the hybrid camera system. The resolution of the test multi-view images is 1920×1080, and that of the depth maps is 176×144.

Table 1.Specification of hybrid camera system

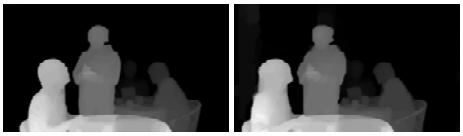
Devices	Specifications	Details
Multi-view cameras (pcA1900-32gc)	Output format	NTSC or PAL (16:9 ratio, HD)
Depth camera (SR400)	Measured depth range	0.50m to 5.00m
	Pixel Array Size	QCIF (176 (h) x 144 (v))
Sync. Generator (NI Trigger Box)	Output format	SD/HD Video Generation

(a) Stereo Image

(b) Acquired Depth Map

Figure 5.Test multi-view image and its depth map

(a) Stereo Image



(b) Stereo Depth Map

Figure 6.Generated multi-view depth video

Figure 6 shows the final multi-view color images and their corresponding depth maps for the 200st frame of Café. To compare the depth quality of the proposed method with previous works, we have shown the disparity map generated by the DERS software for the 3rd view image of the 200st frame of Café as shown in Fig. 7. We can observe that some regions of the depth maps generated by the previous method have noticeable errors in concave areas. Furthermore, the mismatched disparities in black hair were remarkably reduced by the proposed method.

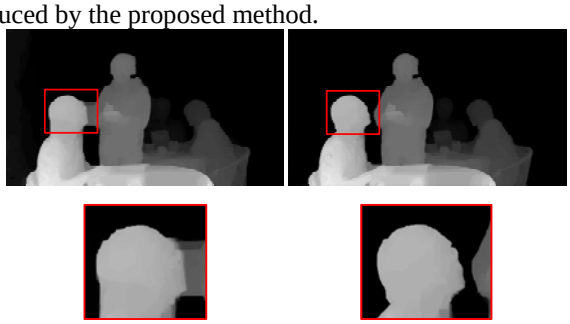


Figure 7.Depth comparison with the previous work

From Fig. 6 and Fig. 7, we notice that depths for the overlapped regions in foreground of Café were generated successfully, though the boundaries of the black hair were noisy. In addition, the yellow table expresses gradual depth difference despite the monotonous color of the table. As a result, we could overcome the two main problems of passive depth sensing efficiently, depth estimation on the occluded and textureless regions, using the depth camera data as the supplementary information.

To evaluate the subjective quality of the proposed method, we list the PSNR result of the synthesized view images using the previous method and the proposed one in Table 2.

Table 2.Average PSNR of synthesized images for CAMERA3

SEQUENCE	Average PSNR	
	DERS	Proposed method
Café	33.95	34.87

V. CONCLUSION

In this paper, we have presented a new approach to generate depth maps corresponding to color images using the proposed hybrid camera system. We have used depth information acquired by a depth camera to generate the initial depth maps for stereo matching. We then have generated the final depth maps using segmentation-based stereo matching and the proposed cost functions. Experimental results have shown that our scheme produced more reliable depth maps and multi-view images compared with previous methods. Therefore, our proposed system could be useful for various 3D multimedia applications and displays.

ACKNOWLEDGMENT

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2011-(C1090-1011-0003)).

REFERENCES

[1] Ministry of Science and Technology of Korea, National Technology Roadmap, 2003.

[2] K. Balasubramanian, “On the Realization of Constraint-free Stereo Television,” IEEE Trans. on Consumer Electronics, vol. 50, no. 3, pp. 895-902, 2004.

[3] C. Fehn, E. Barre, and S. Pastoor, “Interactive 3DTVConcepts and Key Technologies,” Proceedings of the IEEE, vol. 94, no. 3, pp. 524-538, 2006.

[4] J. Cha, S.M. Kim, S.Y. Kim, S. Kim, I. Oakley, J. Ryu, K.H. Lee, W. Woo, and Y.S. Ho, “Client System for Realistic Broadcasting: a First Prototype,” Lecture Notes in Computer Science, vol. 3768, pp. 176-186, 2005.

[5] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, “Efficient Prediction Structures for Multiview Video Coding,” IEEE Trans. on Circuit and Systems for Video Technology, vol. 17, no. 11, pp. 1461-1473, 2007.

[6] C. Fehn, “Depth-image-based Rendering (DIBR), Compression and Transmission for a New Approach on 3D TV,” Proc. of SPIE Stereoscopic Displays and Virtual Reality Systems, vol. 5291, pp. 93-104, 2004.

[7] ISO/IEC JTC1/SC29/WG11 N8944, “Preliminary FTV model and requirements,” April 2007.