Generation of Multi-View Video Using a Fusion Camera System for 3D Displays

Eun-Kyung Lee and Yo-Sung Ho, *Senior Member*, IEEE

Abstract — *In this paper, we present a fusion camera system combining one time-of-flight depth camera and two video cameras to generate multi-view video sequences. In order to obtain the multi-view video using the fusion camera system for 3D displays, we capture a stereo video using a pair of video cameras and a single view depth video with the depth camera. After performing a 3D warping operation for the depth video to obtain an initial depth map at each viewpoint, we refine it using segment-based stereo matching. To reduce mismatched depth values along object boundaries, we detect moving objects using color difference between frames. Finally, we recompute the depth value of each pixel in every segment using stereo matching with a new cost function. Experimental results show that the proposed fusion system produces multi-view video sequences with accurate depth maps, especially along the boundaries of objects. Therefore, it is suitable for generating more natural 3D views for 3D displays than previous works*[1].

Index Terms — 3D display, depth camera, depth estimation, multi-view image generation

I. INTRODUCTION

As 3D videos become attractive in a variety of 3D multimedia applications, it is essential to obtain multi-view video sequences with corresponding depth maps, which are often called as multi-view video-plus-depth data [1]. In the near future, consumers will be able to experience 3D depth impression and choose their own viewpoints in the immersive visual scenes created by 3D videos. Recently, the ISO/IEC JTC1/SC29/WG11 Moving Picture Experts Group (MPEG) has recognized the importance of the multi-view video-plus-depth data for free-viewpoint TV (FTV) or 3DTV [2], and has investigated needs for standardization on 3D video coding [3], [4]. Moreover, 3D video systems have been studied to represent 3D scenes for 3D displays [5], [6].

With respect to the current 3D research activities, it is important to estimate accurate depth information from real world scenes. Although various depth estimation methods have been developed in the field of computer vision, accurate measurement of depth information from natural scenes is still time-consuming and problematic.

In general, depth estimation methods can be classified into two categories: passive depth sensing and active depth sensing. The former calculates depth information indirectly from 2D images captured by two or more video cameras. Typical examples include shape from focus [7] and stereo matching [8]. The advantage of indirect depth estimation is low price because we can create depth maps using cheap off-the-shelf video cameras. However, accuracy of the depth maps is relatively lower than those produced from active approaches in occlusion and textureless regions.

On the other hand, active depth sensing methods usually employ physical sensors, such as laser, infrared ray (IR), or light pattern, to obtain depth information from natural scenes directly. Structured light patterns [9] and depth cameras [10], [11] are major examples of these approaches. Nevertheless, these direct depth estimation tools and systems are quite expensive for consumers. Therefore, time-of-flight (TOF) depth cameras with low price and small size have been introduced and applied for 3D home game and multimedia environment [12]. While they can capture depth values directly in real-time, their crucial disadvantages are that they produce only low-quality depth maps with optical noises.

In recent year, fusion camera systems composed of multiple video cameras and one or more TOF cameras have been proposed [13], [14]. Zhu *et al*. [15] presented a calibration method to improve depth quality using a TOF depth sensor. They used the probability distribution function of the depth information measured by the TOF depth sensor and provided a more reliable depth map. Lee *et al*. [16] enhanced the depth resolution and accuracy by combining the actual distance information measured by the depth camera with the disparity map estimated by the passive depth sensing method. However, the previous fusion systems have produced only low-resolution depth maps and focused on generating depth maps of static 3D scenes.

Nowadays, many research institutes and companies are interested in development of a fusion camera system for 3D consumer devices such as 3D cellphones, 3D tablet PCs, 3D laptops, 3D game consoles, etc. Since forthcoming 3-D multimedia applications running on those devices are expected to use high quality 3-D videos, we need to create multi-view video data with high quality depth information.

In this paper, we devise a fusion camera system with one depth camera and stereo video cameras. The proposed 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 propose a practical 3D video generation solution for dynamic 3D scenes, which can be applicable to 3D consumer devices.

[1] 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-2010-(C1090-1011-0003)).

E.K. Lee and Y.S. Ho are with the Department of Information and Communications, Gwangju Institute of Science and Technology (GIST) (e-mail: {eklee78, hoyo}@gist.ac.kr).

Contributed Paper
Manuscript received 10/15/10
Current version published 12/23/10
Electronic version published 12/30/10.

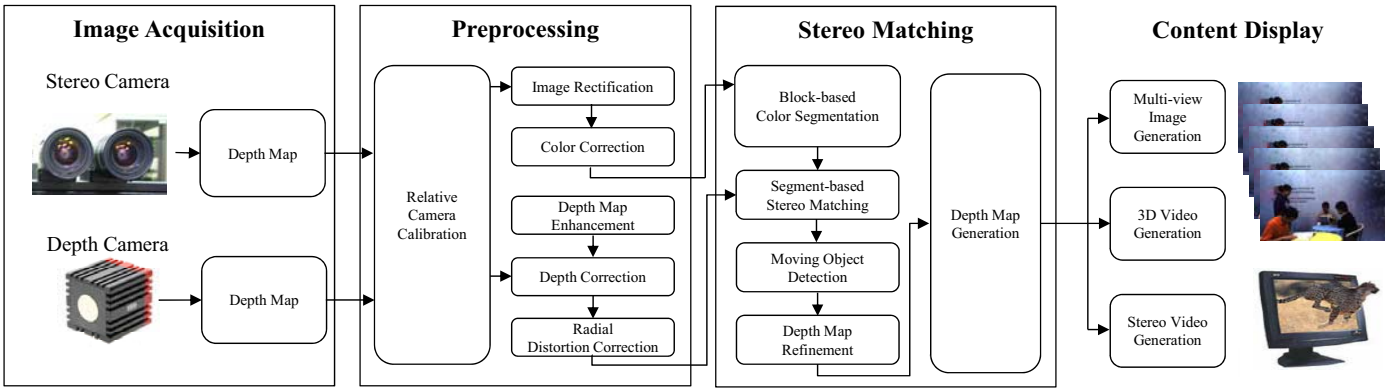


Fig. 1. Overall framework of multi-view image generation for 3D display.

The remainder of this paper is organized as follows. In Section II, we present the overall architecture of the proposed fusion camera system. Section III describes preprocessing steps for enhancing depth maps and Section IV presents how to generate the multiple video sequences with their corresponding depth maps using the proposed camera system. After showing experimental results in Section V, we draw conclusion in Section VI.

II. System Architecture

The proposed fusion system is composed of one depth camera and two video cameras. Figure 1 illustrates the overall framework to generate stereo depth map using the fusion system. After calibrating each camera independently, we perform an image rectification to adjust vertical mismatches in multiple images. Then, we apply a color correction operation to maintain color consistency among stereo images. To obtain depth maps for stereo images, we perform a 3D warping operation onto each stereo 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 stereo 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 using color difference between frames. Then, the depth of each segment is refined by a color segmentation-based stereo matching method. Finally, we obtain depth maps by conducting a pixel-based depth map refinement using a proposed cost function in each segment. Since all steps are processed twice, from left to right and right to left, we can obtain at least two depth maps in two views. From the two-view information, multi-view images can be generated from the proposed algorithm.

In this paper, we introduce a compact and minimum camera setup for multi-view image generation with two video cameras and one depth camera. However, depending on applications and device capabilities, this system can be easily extended to multi-view video and multi-view depth cameras.

III. Preprocessing Of the Fusion Camera System

If the proposed camera setup is built in 3D devices, the following steps can be skipped in practical environment. Once the camera setup is fixed in the device, parameters computed from the preprocessing stage are not changed.

A. Relative Camera Calibration

Since the proposed fusion camera system consists of two different types of cameras, a depth camera and stereo video cameras, it is essential to find out relative camera information through camera calibration [17]. For that, we apply a camera calibration algorithm to each camera in our camera system and obtain projection matrices for the depth camera and each video camera.

$$P_s = K_s[R_s \mid t_s] \tag{1}$$

$$P_k = K_k[R_k \mid t_k] \tag{2}$$

where P_s is the projection matrix of the depth camera represented by its camera intrinsic matrix K_s, rotation matrix R_s, and translation vector t_s. P_k means the projection matrices of the k^{th} video camera which consisted of its camera intrinsic matrix K_k, rotation matrix R_k, and translation vector t_k.

We then employ a rectification operation [18]. The cameras have geometric errors because they are set manually by hand. In order to minimize the geometric errors, we find the common baseline, and then apply the rectifying transformation to the stereo image. Consequently, the projection matrix of video cameras are changed as

$$P_k' = K_k'[R_k' \mid t_k'] \tag{3}$$

where K_k' and R_k' are the modified camera intrinsic matrix and rotation matrix of the k^{th} video camera, respectively. Thereafter, we convert the rotation matrix R_s of the depth camera into the identity matrix I by multiplying inverse rotation matrix R_s^{-1}. The translation vector t_s of the depth camera is also changed into the zero matrix O by subtracting the translation vector t_s. Hence, we can define new relative projection matrices for the stereo camera on the basis of the depth camera as

$$P_s' = K_s[I \mid O] \tag{4}$$

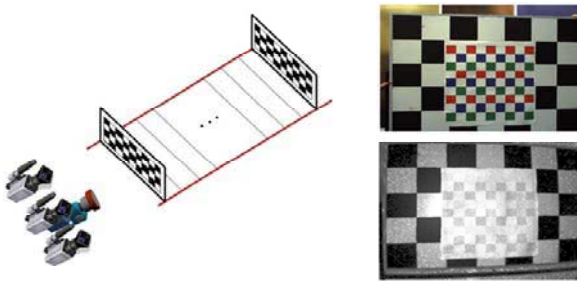
$$\widetilde{P}_k' = K_k'[R_k' R_s^{-1} \mid t_k - t_s] \tag{5}$$

where P_s' and $\widetilde{P}_k'$ are final projection matrices of the depth camera and the k^{th} video camera, respectively. After relative camera calibration, we resolve the color mismatch problem of stereo images using a color calibration method [19]. The color characteristics of captured images are usually inconsistent due to different camera properties and lighting conditions even the hardware type and specification of the multiple cameras are the same. Thereafter, we perform bilateral filtering to reduce optical noises included in the depth map acquired from the depth camera [20].

B. Depth Calibration

The depth values measured by the depth camera are very sensitive to noises. Their sources are diverse including physical limitation of hardware and specific object properties, etc. Therefore, depth data are noticeably contaminated with random and systematic measurement errors dependent on reflectance, angle of incidence, and environmental factors like temperature and lighting [21]. To reduce those errors, we employ a depth calibration method [17].

For depth calibration in indoor environments, we compute the depth of the planar checker pattern within the limited space by increasing the distance from the image pattern to the depth camera using our system as shown in Fig. 2. To extract the corresponding feature points in two different types of cameras efficiently, we use the color checker pattern. The pattern image is captured in every $5cm$ distance. The plane pattern is orthogonal to the image plane.



(a) Pattern acquisition (b) Pattern images from hybrid camera system

Fig. 2. Acquisition of the planar check pattern for depth calibration.

Thereafter, we make a four dimensional look-up table (LUT) mapping 3D positions of the multiple video cameras and the depth value from the depth camera. 3D position is constructed by x, y position of the feature point and the real depth value calculated by the multi-view image. Depth accuracy test using the acquired depth map and calibrated depth map the real depth value z calculated from the multi-view image by pairwise stereo matching. Since we have already obtained camera parameters, the real depth value is calculated by

$$d_n(p_x, p_y) = \frac{K \cdot B}{D_n(p_x, p_y)} \tag{6}$$

where K is the focal length of the left camera and B is the baseline distance between two video cameras. $D_n(p_x, p_y)$ is the real depth value corresponding to the measured disparity value $d_n(p_x, p_y)$ at the pixel position (p_x, p_y) in the checker pattern.

To check the accuracy of the calibrated depth value, we perform 3D warping to the stereo camera. Figure 3(a) is the 3D warping result using the acquired depth map and Fig. 3(b) shows that of the calibrated depth map using the LUT. While there are many mismatched depth values in Fig. 3(a), most of them are correctly matched in the boundaries of the rectangular box in Fig. 3(b). The other problem is that even though the distance from the depth camera to the object is constant, depth information from the depth camera can be different depending on the object color and lighting conditions.

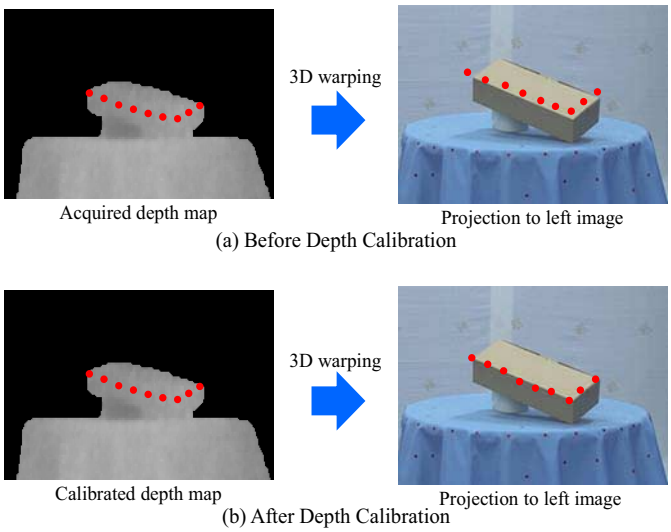


Fig. 3. Depth accuracy test using acquired depth map and calibrated depth map.

To analyze the depth sensitivity of a static object in the dynamic scene, we check the depth values of a black hair, as shown in Fig. 4. We can notice the inconsistent depth value changes of the static object caused by object movement and material properties. Especially, the depth value of the dark color region measured by the depth camera is very unstable and unreliable. The black hair has to sustain a near-constant depth in the scene; however, the acquired depth values are unpredictable and random. The reason is that dark or black colors absorb light of all frequencies and the depth camera uses near IR rays.

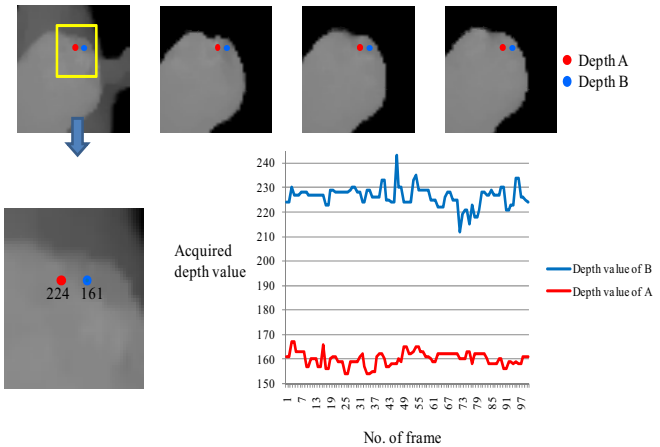


Fig. 4. Depth inconsistency for static object.

Although we perform the depth calibration to correct the acquired depth map, there are still limitations in the depth values acquired from the depth camera. To obtain the high-quality multi-view depth maps, we need to refine the acquired depth value using an efficient stereo matching algorithm.

C. Radial Distortion Correction for Depth Images

Depth map from the depth camera have a large amount of lens radial distortion. There are two types lens distortion which are barrel distortion and pincushion distortion. In this case, the barrel distortion is occurred by the intrinsic problem of the depth camera. This distortion causes not only the shape mismatch between the color image and the corresponding depth image but also the errors in the results of some feature point based processing such as camera calibration.

In order to avoid that situation, we have to perform radial distortion correction to the obtained depth images. In general, there are two main categories of radial distortion correction. Methods in the first category use the point correspondences between two or more views. The second category also has lots of approaches which are based on the distorted straight line components in the image.

In the proposed fusion camera system, we use one of the second approaches to correct the radial distortion in the depth images [22]. After finding the curved straight line component in the captured image, we estimate the distortion center and the distortion parameter. With the distortion information, we can reconstruct the image from the distorted image. Figure 5 shows the depth and intensity images before and after the correction.

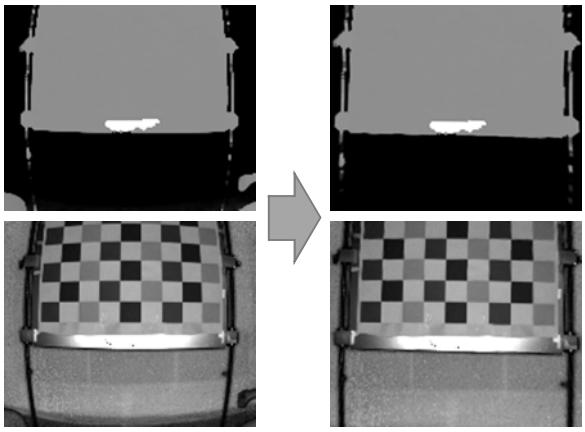


Fig. 5. Radial distortion correction.

IV. Depth Map Generation

A. 3D Warping of Depth Camera Data

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

$$P_s = K_s^{-1} \cdot p_s \tag{7}$$

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

$$p_k' = \tilde{P}_k' \cdot P_s \tag{8}$$

where P_k' indicates the projection matrix of the k^{th}-view video camera. Figure 6 shows the result of 3D warping using the acquired depth maps.

(a) Warped depth image

(b) Matched to color image

Fig. 6. 3D warped depth map.

B. Region Separation

To estimate depth maps of stereo video cameras using the warped depth information, we segment the multi-view image by a mean-shift color segmentation algorithm [23]. However, we cannot control the maximum segment size because there is no parameter to control the maximum segment size.

When we perform the segment-based stereo matching, one segment has one depth value. If the size of segment is too large, we cannot get a smooth depth map. The other way, if the size of segment is too small, it is hard to overcome textureless problem during the stereo matching. To solve this problem, we split one image into 16×16 block segments, so that we can limit the maximum segment size.

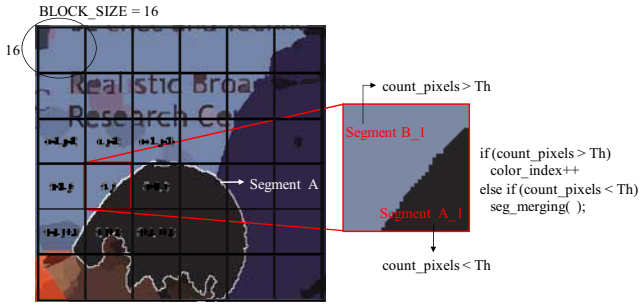


Fig. 7. Block-based segment merging.

Figure 7 shows the procedure of the segment merging. A block can have two or more color segments. Before merging the segment, we split the segmented image into block-based segment again. If each segment is smaller than half size of the block, we merge it into one segment by searching adjoined blocks to find the same indexed segment. If the size of the merged block is larger than threshold, the merging procedure is finished; otherwise we repeat the same process until merging condition is satisfied.

The searching order of connected blocks is right, bottom, left, and top including the diagonal directions because left and top blocks are merged block and right and bottom block will be merged blocks. For example, *Segment A* divide into many block-based segments and *Block* (*i*, *j*) have two segments: *Segment A_1* and *Segment B_1*. Since the size of *Segment A_1* is smaller than the predefined threshold value in Fig. 8, the same indexed segment of *Segment A_1* is the blocks in (*i+1*, *j*), (*i*, *j+1*), and (*i+1*, *j+1*). We merge the current *Segment A_1* and the same indexed segment in (*i+1*, *j*) by the searching order.

Before we estimate depth maps, we separate moving object using color difference between frames. To extract the moving object in the current frame, we calculate color differences between the previous frame *n-1* and the current frame *n* by using the threshold which indicates the current position is foreground or not. We cannot directly use the segment-based moving object detection because shape of each segment can be varied in the temporal domain as shown in Fig. 8.

Fig. 8. Segmentation results in temporal domain.

Since color segmentation is performed frame by frame, it is hard to find the same segment in the temporal domain. Therefore, we use the Euclidean distance between frames to extract the moving objects as

$$E_n(x,y)=\sqrt{(R_{n-1}(x,y)-R_n(x,y))^2+(G_{n-1}(x,y)-G_n(x,y))^2+(B_{n-1}(x,y)-B_n(x,y))^2} \quad (9)$$

where R, G, and B indicate the pixel values in RGB color domain. To find the moving object, we compute the $E_n(x, y)$ at each pixel location for all pixels. If we subtract the RGB value between frames, camera noises can be mixed up. To remove them, we calculate the average RGB value for 3×3 block. If the average is larger than the threshold value, we set the center pixel of each 3×3 block as the foreground pixel. Figure 9 shows the result of moving objects for 78th frame images in the left camera.

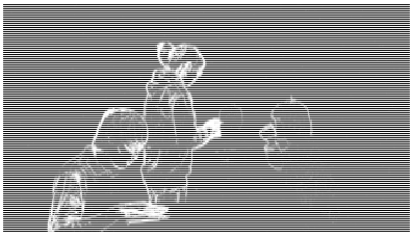


Fig. 9. Moving object detection using color difference between frames.

Segment-based Multi-view 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 [7]. 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 stereo 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 stereo image as shown in Fig. 10.

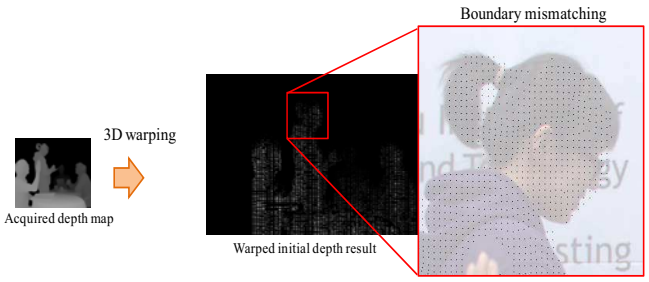


Fig. 10. Boundary mismatching problem.

To obtain the accurate initial depth value, we use the warped results as multiple initial depth values for stereo matching. If we start the stereo matching with the initial depth, we can reduce the search range for finding the matched region. In addition, depending on search range reduction, we can overcome the mismatched problem in the textureless regions. 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.

Figure 11 shows the position of the initial depth in two directional regions, horizontal and vertical regions. One or more initial depth values usually exist in a 10×10 area because of the difference of the resolution. In this case, we set the horizontal search region as 80×20 and the vertical search region as 20×80. By using the multiple initial depths, we can set initial depth for the depthless regions in the boundary of objects as shown in Fig. 11.

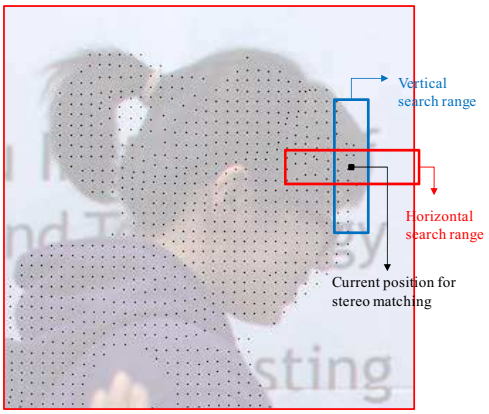


Fig. 11. Set of multiple initial depth values.

Since stereo matching measures the difference between the corresponding points of two images, called as the disparity, we convert the initial depth into its disparity for stereo matching by

$$InitDisp(x, y) = \frac{K \cdot B}{InitDepth(x, y)} \tag{10}$$

where $InitiDisp(x, y)$ is the converted disparity at the pixel position (x, y) from the corresponding initial depth $InitDepth(x, y)$. B and K are the distance between two video cameras and the focal length of the current video camera, respectively. After performing stereo matching with the initial disparity, we convert again the calculated disparity into its depth value to produce the depth map. Before performing bi-directional stereo matching, we need to set the candidate of the initial depth value.

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 or right view images by

$$FG\_d_i(InitDisp) = \min(\min(\sum_{j=0}^{a} MAD(j)), \min(\sum_{k=0}^{b} MAD(k)) \tag{11}$$

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. $FG\_d_i(InitDisp)$ 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. We define that the background has no initial depth or the number of the included initial depth in the segment is less than 10% of the size of the segment. In estimating depth of background, we set the minimum and maximum depth/disparity value. We then find the minimum MAD as the initial disparity of the current segment in the background by

$$BG\_d_i(InitDisp) = \min(\sum_{i=\min Disp}^{\max Disp} MAD(i)) \tag{12}$$

where $BG\_d_i(InitDisp)$ 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 $d_i(Initdisp)$ of the segment i in the current view image n by

$$d_i(InitDisp) = \min(FG\_d_i(InitDisp), BG\_d_i(InitDisp)) \tag{13}$$

C. Multi-view Depth Map 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-quality multi-view video based on 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 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 as shown in Fig. 9. 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.

$$E(x, y, d) = \begin{cases} w_s f_s(x, y, d_s(x, y)) + w_d f_d(x, y, d_d(x, y)) \;\; if \;\; obj\_mov(x, y) = 1 \\ w_s f_s(x, y, d_s(x, y)) + w_d f_d(x, y, d_d(x, y)) + w_t f_t(x, y, d_t(x, y)) \;\; if \;\; obj\_mov(x, y) = 0 \end{cases} \tag{14}$$

where w_s, w_d, w_t are the weighting factors for depth refinement. $f_s(x, y, d_s(x,y))$ 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. $obj\_mov(x,y)$ 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

$$f_s(x, y, d_s(x, y)) = med(s_a(x, y), s_b(x, y), s_c(x, y)) \tag{15}$$

We can calculate the smoothness value as shown in Fig. 12. $s_a(x,y)$ 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 $(x-1, y)$.

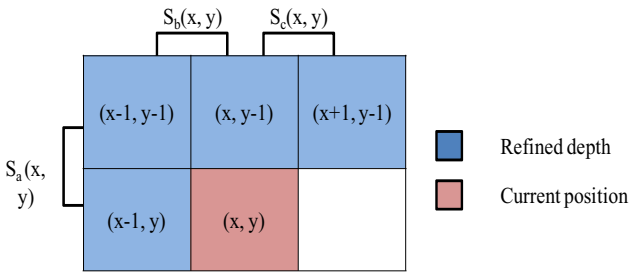
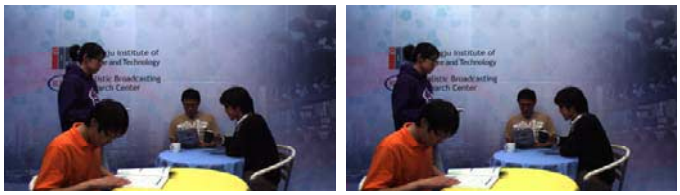


Fig. 12. Smoothness definition with gradient of the refined depth values.

V. EXPERIMENTAL RESULTS AND ANALYSIS

In order to generate the high-quality depth maps, we have constructed a fusion camera system with two cameras and one depth camera. The measuring depth range of our depth camera is from 0.50*m* to 5.00*m*. The baseline distance between two video cameras is 6.5*cm*. The proposed camera system's baseline distance depends on the physical volume of our video cameras and the depth camera. However, it is possible to reduce the baseline between cameras in other system configurations. Figure 13 shows the acquired test sequence captured by the fusion camera system. The resolution of our test stereo images is 1920×1080, and that of the depth maps is 176×144. From our experimental, the weighting factors of the cost function w_s, w_d, w_t are 0.3, 0.5, and 0.2 and the threshold value of Euclidean distance, 10 is used.



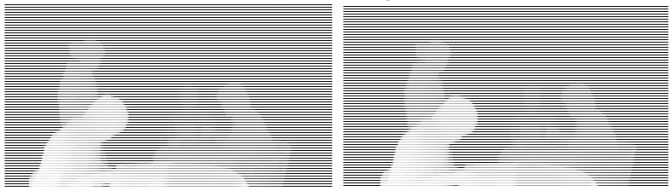
(a) Stereo Image

(b) Acquired Depth Map

Fig. 13. Test multi-view image and its depth map

Figure 14 shows the final stereo color images and their corresponding depth maps for the 1st frame. To compare the depth quality of the proposed method with previous works, we have shown the disparity map generated by Zhu's method. method for the left image of the 93rd frame as shown in Fig. 15. 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.

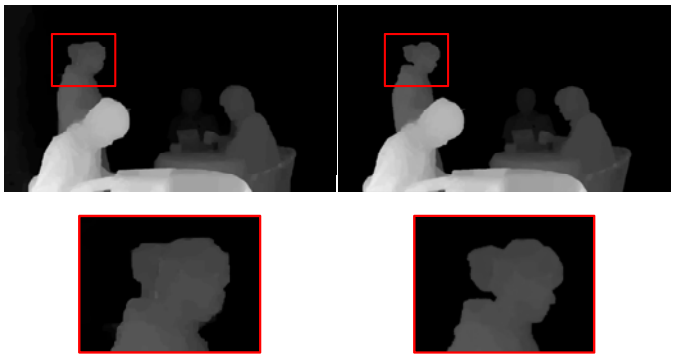
(a) Stereo Image



(b) Stereo depth map

Fig. 14. Generated depth maps

From Fig. 14 and Fig. 15, we notice that depths for the overlapped regions in foreground 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. Figure 16 presents the computed depth maps from 30th to 270th in every 30 frames.



(a) Depth map from Zhu's method (b) Depth map from the proposed system

Fig. 15. Depth comparison with the previous work.

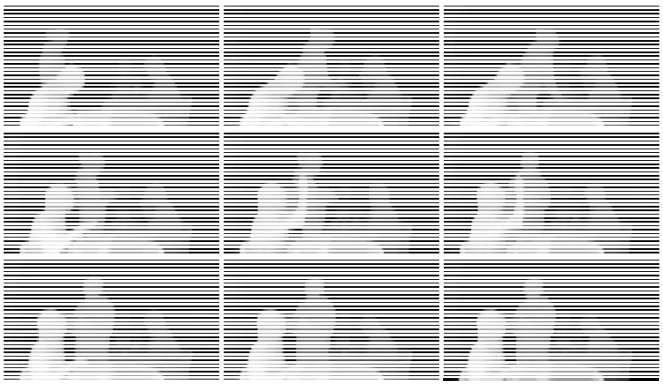
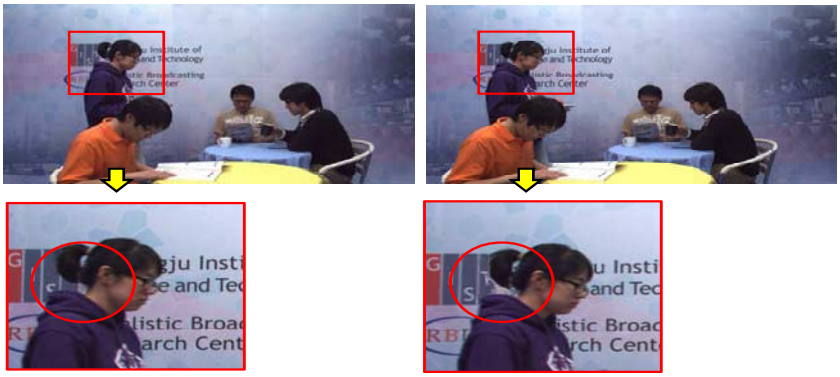


Fig. 16. Generated depth map sequences.

To evaluate the subjective quality of the proposed method, we have synthesized intermediate views with the computed depth map using VSRS software [24]. As shown in Fig. 17, the generated intermediate views using depth maps obtained by the proposed method are reasonable in the aspect of subjective quality. From the aforementioned results, the proposed approach outperforms the previous method.



(a) Zhu's method (b) Proposed system

Fig. 17. Intermediate images comparasion.

Fig. 18. Generated intermediate images using generated depth maps.

Figure 18 shows the generated multi-view images usign the generated depth maps. Table I shows the comparison of the processing time in the depth refinement step. Since each algorithm have different processing step to generate the depth map, it is hard to measure the exact processing time in the same condition. Therefore, we compare the processing time for the depth map refinement step. As shown in Table I, the proposed method is faster than others without the accuracy reduction for depth map generation. From the result, it is useful for the high-quality multi-view video generation.

TABLE I
COMPARISON OF THE PROCESSING TIME

SEQUENCE	Processing time (sec)	
	Zhu's method	Proposed method
Café	836.26	337.21

VI. CONCLUSION

In this paper, we have presented a new approach to generate depth maps corresponding to color images using the proposed fusion 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. With the proposed fusion camera system, we could solve the two main problems in the current passive depth sensing, which is depth estimation on occluded and textureless regions. Finally, we have generated high-quality multi-view images from our system. Therefore, our proposed system could be useful for various 3D multimedia applications and displays.

REFERENCES

[1] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger, "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability," *Signal Processing Image Communication*, vol. 22, no. 2, pp. 217-234, Feb. 2007.

[2] C. Fehn, R. Barre, and S. Pastoor, "Interactive 3DTV- concepts and key technologies," *Proceedings of the IEEE*, vol. 94, no. 3, pp. 524-538, March 2006.

[3] ISO/IEC JTC1/SC29/WG11 N8944, "Preliminary FTV model and requirements," April 2007.

[4] A. Smolic and D. McCutchen, "3DAV exploration of video-based rendering technology in MPEG", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 3, .pp. 348-356, March 2004.

[5] S. Kim, S. Lee, and Y. Ho, "Three-dimensional Natural Video System based on Layered Representation of Depth Maps," *IEEE Trans. Consumer Electronics,* vol. 52, no. 3, pp. 1035-1042, August 2006.
[6] H. Shin, Y. Kim, H. Parl, and J. Park, "Fast View Synthesis using GPU for 3D Display," *IEEE Trans. Consumer Electronics,* vol. 54, no. 4, pp. 2068-2076, November 2006.
[7] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *Proc. of ACM SIGGRAPH*, pp. 600–608, August 2004.
[8] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *Proc. of ACM SIGGRAPH*, pp. 600–608, August 2004
[9] D. Scharstein, and R. Szeliski, "High-accuracy stereo depth maps using structured light," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 195-202, June 2003.
[10] G. Iddan and G. Yahav, "3D imaging in the studio and elsewhese," *Proc. of SPIE Vidometrics and Optical Methods for 3D Shape Measurements*, pp. 48–55, Jan. 2001.
[11] M. Kawakita, T. Kurita, H. Kikuchi, and S. Inoue, "HDTV axi-vision camera," *Proc. of International Broadcasting Conference*, pp. 397–404, Sept. 2002.
[12] S. Gokturk, H. Yalcin, C. Bamji, "A time-of-flight depth sensor - system description, issues and solutions," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshop*, vol. 3, pp. 35, June 2004.
[13] Y. Kim, D. Chan, C. Theobalt, and S. Thrun, "Design and calibration of a multi-view TOF sensor fusion system," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 539–546, June 2008.
[14] G. Um, K. Kim, C. Ahn, and K. Lee, "Three-dimensional scene reconstruction using multi-view images and depth camera," *Proc. of SPIE Stereoscopic Displays and Virtual Reality Systems XII*, vol. 5664, pp. 271–280, Jan. 2005.
[15] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," *Proc. of Advances in Neural Information Processing systems*, pp. 291-298, Dec. 2005.
[16] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 231–236, June 2008.
[17] E. Lee, Y. Kang,Y. Jung, Y. Ho, "3D video generation using hybrid camera system," *Proc. of International Conference on Immersive Telecommunications (IMMERSCOM)* , pp. T5(1-6), June 2009.
[18] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1330-1334, Nov. 2000.
[19] Y. Kang, and Y. Ho, "Geometrical compensation for multi-view video in multiple camera array," *Proc. of International Symposium ELMAR*, vol. 1, pp 83-86, Sept. 2008.
[20] N. Joshi, B. Wilburn, V. Vaish, M. Levoy, and M. Horowitz, "Automatic color calibration for large camera arrays," in UCSD CSE Technical. Report. CS2005-0821, May 2005.
[21] J. Cho, I. Chang, S. Kim, and K. Lee, "Depth image processing technique for representing human actors in 3DTV using single depth camera," *Proc. of 3DTV conference*, paper no. 15, May 2007.
[22] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "High-quality scanning using time-of-flight depth superresolution," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition on Time-of-Flight Computer Vision*, pp.1-8 2008.
[23] A. Wang, T. Qiu, and L. Shao, "A simple method of radial distortion correction with centre of distortion estimation," *Journal of Mathematic Imaging and Vision*, vol. 35, no. 3, pp. 165-172 Nov. 2009.
[24] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 603–619, May 2002.
[25] ISO/IEC JTC1/SC29/WG11 M15377, "Reference softwares for depth estimation and view synthesis," April 2008.

BIOGRAPHIES

Eun-Kyung Lee received both B.S. and M. S. degree in computer engineering from Honam University (HU), Korea, in 2002 and 2004, respectively. She is currently working towards her Ph.D. degree in the Information and Communications Department at the Gwangju Institute of Science and Technology (GIST), Korea. Her research interests include digital signal processing, multi-view video coding algorithms and systems, multi-view depth map generation, 3D television, and realistic broadcasting

Yo-Sung Ho received both B.S. and M.S. degrees in electronic engineering from Seoul National University, Korea, in 1981 and 1983, respectively, and Ph.D. degree in Electrical and Computer Engineering from the University of California, Santa Barbara, in 1990. He joined the Electronics and Telecommunications Research Institute (ETRI), Korea, in 1983. From 1990 to 1993, he was with Philips Laboratories, Briarcliff Manor, New York, where he was involved in development of the advanced digital high-definition television (AD-HDTV) system. In 1993, he rejoined the technical staff of ETRI and was involved in development of the Korea direct broadcast satellite (DBS) digital television and high-definition television systems. Since 1995, he has been with the Gwangju Institute of Science and Technology (GIST), where he is currently a professor in the Information and Communications Department. His research interests include digital image and video coding, image analysis and image restoration, advanced coding techniques, digital video and audio broadcasting, 3D television, and realistic broadcasting.