3-D Video Generation using Hybrid Camera System

Eun-Kyung Lee, Yun-Suk Kang, Yo-Sung Ho
Gwangju Institute of Science and Technology (GIST)
1 Oryong-dong, Buk-gu, Gwangju, 500-712, Korea
{eklee78, yunsuk, hoyo}@gist.ac.kr

Young-Kee Jung
Honam University
Eodeungno 330, Gwangsan-gu, Gwangju, Korea
ykjung@honam.ac.kr

ABSTRACT
In this paper, we present a new camera system combining a time-of-flight depth camera and multiple video cameras to generate a multiview video-plus-depth. In order to get the 3-D video using the hybrid camera system, we first obtain a multiview image from the multiview camera and a depth map from the depth camera. Then, initial depths of each view image are estimated by performing 3-D warping with the depth map. Thereafter, multiview depth estimation using the initial depths is carried out to get each view initial disparity map. Finally, we refine the initial disparity map using a belief propagation algorithm so that we can generate the high-quality multiview disparity map. Experimental results show that the proposed hybrid camera system produces a 3-D video with more accurate multiview depths and supports more natural 3-D views than the previous works.

Keywords
3-D video generation, depth camera, depth estimation, multiview camera

1. INTRODUCTION
As 3-D video becomes attractive in a variety of 3-D multimedia applications, it is essential to obtain a multiview video enriched with its associated depth map, which is often called as a multiview video-plus-depth [1]. In the near future, consumers will be able to not only experience 3-D depth impression but also choose their own viewpoints in the immersive visual scene created by the 3-D video. Recently, the ISO/IEC JTC1/SC29/WG11 Moving Picture Experts Group (MPEG) has also been acknowledged the importance of the multiview video-plus-depth for free-viewpoint TV or 3DTV [2], and has been investigating the needs of standardization about 3-D video coding [3, 4].

In order to represent the 3-D video, it is necessary to obtain accurate depth information. In general, depth estimation methods are categorized into two major classes: active depth sensors [5, 6] and passive depth sensors [7, 8]. Passive depth sensors estimate depth information indirectly from 2-D images captured by two or more cameras. One of the well-known indirect depth sensors is stereo matching [9, 10]. On the other hand, active depth sensors obtain depth information from the natural scenes directly using physical sensors, such as laser or infrared ray (IR) sensors.

In general, passive depth sensors are limited to estimate an accurate depth map due to the failure of correspondence point matching on the textureless and occluded regions. On the other hand, active depth sensors can only generate depths of nearby objects in a lower resolution.

Recently, fusion methods that combine video cameras and a time-of-flight (TOF) depth camera have been introduced [11, 12]. The depth camera produces accurate depths from real scenes by integrating a high-speed pulsed IR light source with a conventional broadcast TV camera [13, 14]. These hybrid camera systems enhance depths estimated by a passive depth sensing by compensating them with depths obtained from the depth camera. In addition, Zhu et al. have also presented an effective calibration method to improve depth quality using a TOF depth sensor [15]. They used the probability distribution function on depths from the TOF depth sensor to produce more reliable depth maps. However, the previous hybrid camera systems have only produced low-resolution depth maps and focused on generating depth maps for static scenes.

Since the future 3-D applications are expected to use high-quality and high-resolution 3-D videos, we need to create such a multiview video-plus-depth. In this paper, we propose a new camera system consisting of a depth camera and multiple video cameras such as the one shown in Fig. 1. The proposed hybrid camera system produces a high-resolution multiview depth map of a dynamic scene using a low-resolution depth map obtained from the depth camera. The main contribution of this work is that we provide a practical solution to generate a high-quality 3-D video using 3-D warping in the hybrid camera system.

2. HYBRID CAMERA SYSTEM
The proposed hybrid camera system is composed of one depth camera and *n* video cameras. The *n* video cameras are allocated in array to construct a multiview camera. In addition, there is a clock generator sending a synchronization signal constantly. The synchronization signal is distributed into each camera and its corresponding personal computer equipped with a video capture board. Basically, the proposed hybrid camera system provides *n* images from the *n* video cameras and a depth map from the depth camera for each frame. Figure 1 is the overall architecture of the proposed 3-D video generation.

Immerscom 2009, May 27-29, 2009, Berkley, USA.

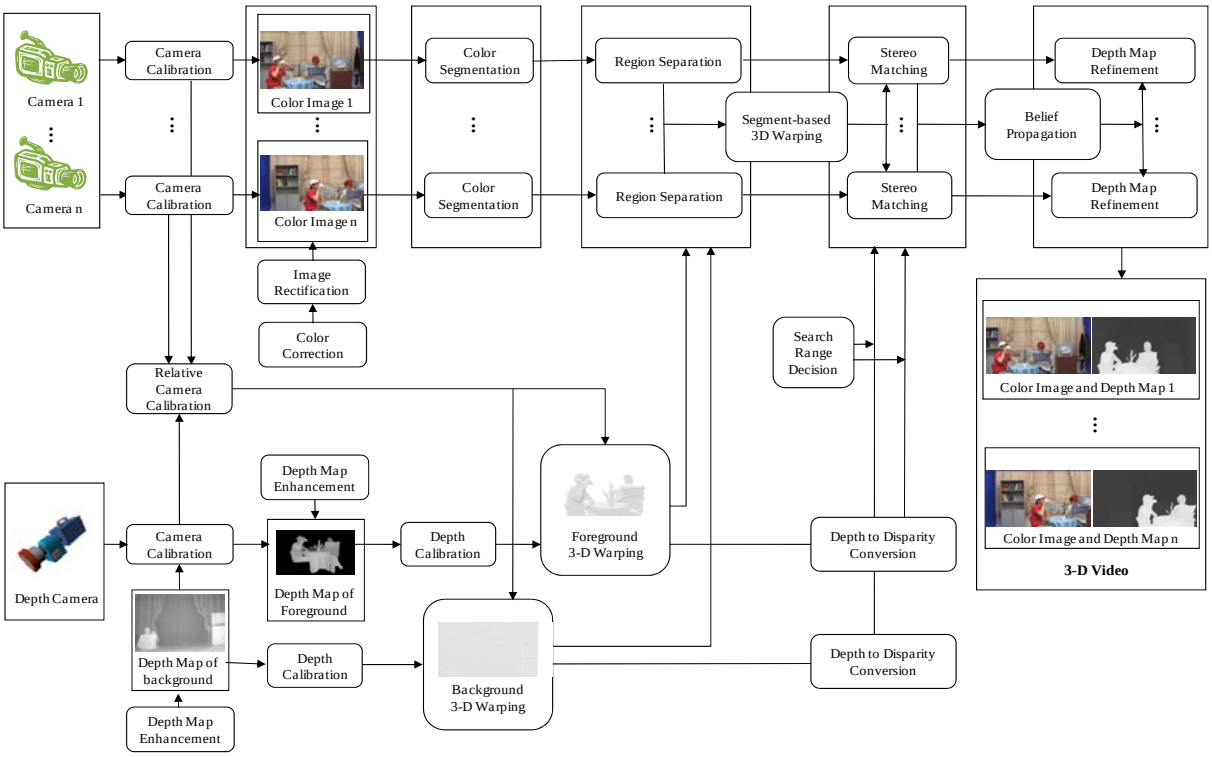


Figure 1. Overall architecture of the proposed 3-D video generation

Figure 2. The proposed hybrid camera system

Figure 2 shows the hybrid camera system that we have constructed with five high definition (HD) video cameras and a standard definition (SD) depth camera. In practice, the measureable depth range of the depth camera is up to approximately 5m, and the depth quality of the depth map becomes lower as the depth range becomes bigger. Furthermore, the depth camera is only used in the limited indoor environment, such as a virtual studio in a broadcasting station. Therefore, in order to increase depth accuracy, we obtain the background and foreground depths from the depth camera separately; we obtain the background depths from the depth camera within a small depth range in advance. Then, we obtain foreground depths from the depth camera and the multiview image from the multiview camera at the same time.

Figure 2 illustrates the overall framework to generate the 3-D video using the hybrid camera system. First, we calibrate each camera independently, and then we perform a color correction algorithm to maintain color consistency among view images. Then, the color-consistent multiview image is rectified. Before obtaining the multiview video, we capture a depth map for the background in advance.

In order to get each view depth map of the multiview image, we perform a 3-D warping operation onto the depth information obtained from the depth camera. The warped depth data are used as initial disparities of each view depth map. After 3-D warping, we color-segment the multiview image and assign the average of warped depth data on each segment as the initial disparity of the segment. Then, we separate each view image into three different regions to improve the depth on the occluded regions: background, foreground, and unknown regions. Finally, the disparity of each segment is independently estimated and refined by color segmentation-based depth estimation for the three different regions, respectively.

3. PREPROCESSING

3.1 Relative Camera Calibration

Since the proposed hybrid camera system is constructed by merging two different types of cameras, a depth camera and a conventional video camera, it is essential to find out relative camera information for the hybrid camera set using the relative camera calibration algorithm.

In order to get relative camera information, we apply a camera calibration algorithm onto each camera in the hybrid camera system. Hence, we can get projection matrices for the depth camera and each video camera as described in Eq. 1.

$$P_s = K_s[R_s \mid t_s]$$
$$P_k = K_k[R_k \mid t_k] \quad (1)$$

where P_s is the projection matrix of the depth camera generated by its camera intrinsic matrix Ks, rotation matrix Rs, and translation vector t_s. The term P_k indicates the projection matrices of the k^{th} video camera generated by its camera intrinsic matrix K_k, rotation matrices R_k, and translation vector t_k, respectively.

Then, we perform the multiview rectification. The multi-camera arrays have geometric errors, because there are manually built. In order to minimize the geometric errors, we calculate the common baseline, and then apply the rectifying transformation to the multiview image. Hence, the projection matrices of video cameras are changed as described in Eq. 2.

$$\tilde{P}_k = K_k'[R_k' \mid t_k] \quad (2)$$

where K_k' and R_k' are the changed 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}, and we convert the translation vector t_s of the depth camera into the zero matrix O by subtracting the translation vector t_s. Hence, we can define the new relative projection matrices for the multiview camera on the basis of the depth camera as described in Eq. 3.

$$P_s' = K_s[I \mid O]$$
$$\tilde{P}_k' = K_k'[R_k' R_s^{-1} \mid t_k - t_s] \quad (3)$$

where P_s' and $\tilde{P}_k'$ indicate the modified projection matrices of the depth camera and the k^{th} video camera, respectively.

After relative camera calibration, we correct the color mismatch problem of multiview images using a color calibration method. In general, the color properties of captured images can be inconsistent due to the different camera properties of the multiview camera system. In the depth camera, we also perform bilateral filtering to reduce optical noises in the depth map.

3.2 Depth Calibration

Depth information of the depth camera is very sensitive to color and motion. Even though the distance from the depth camera to the object is constant, depth information from depth camera is different depending on the environment. Basically, the depth camera system has its own depth calibration tool. However, it is very poorly calibrated. [20]

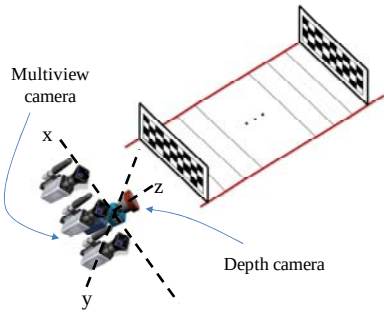


Figure 3. Depth calibration

To calibrate measured depths with their real depths, we find out a mapping curve to compensate the difference between them. In this paper, we use Zhu's algorithm to analyze the characteristics of the depth camera [15]. When objects move in the limited indoor environment, we check the depth of the planar image pattern within the limited space by increasing the distance from the image pattern to the depth camera as shown in Fig. 3. Since we already know the camera parameters of each camera, the real depth values are calculated by Eq. 4.

$$d_n(p_x, p_y) = \frac{K \cdot B}{D_n(p_x, p_y)} \quad (4)$$

where K is the focal length of the video camera, B is the baseline distance between neighboring two video cameras. $d_n(p_x, p_y)$ is the real depth value corresponding to the measured depth value $D_n(p_x, p_y)$ at pixel position (p_x, p_y) in the image pattern depth map.
Thereafter, we generate a mapping curve between real depths and measured depths from the depth camera. In order to generate the mapping curve, we find out the fitting curve using the cubic equation as described in Eq. 5.

$$y = a + bx + cx^2 + dx^3 \quad (5)$$

The cross small rectangular points on the x-y plane in Fig. 4 are formed by the measured depths x and real depths y that minimizes the sum of squared distances to these points. Figure 4 shows the mapping curve for depth calibration.

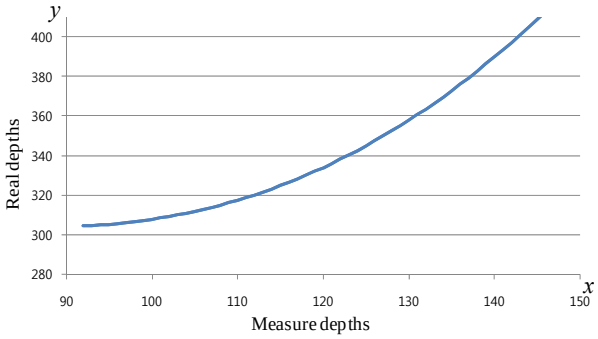


Figure 4. Depth calibration curve mapping

4. 3-D VIDEO GENERATION

4.1 3-D Warping of Depth Camera Data

We regard depths obtained from the depth camera as initial disparities of the multiview image captured by the multiview camera. To match the depth camera data with the multiview image, we project the depth information onto the world coordinate system using 3-D warping. Then, we reproject the warped 3-D data onto each view video camera.

When $D_s(p_{sx}, p_{sy})$ is the depth information at the pixel position (p_{sx}, p_{sy}) in the depth map, we can regard the pixel p_s $(p_{sx}, p_{sy}, D_s(p_{sx}, p_{sy}))$ as a 3-D point. In order to move the 3-D point onto a 3-D point $P_s(x_{sx}, y_{sy}, z_{sz})$ in the world coordinate system, the backward projection is carried out by Eq. 6.

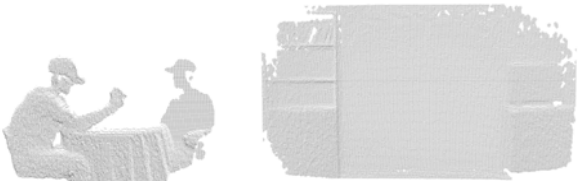
$$P_s = K_s^{-1} \cdot p_s \quad (6)$$

where K_s' indicates the intrinsic matrix of the depth camera. In the backward 3-D warping, since rotation and translation matrices of the depth camera are the identity matrix I and zero matrix O as Eq. 3, respectively, we have only to consider its intrinsic matrix.

Thereafter, we project the 3-D points P_s into each view video camera to get its corresponding pixel position $p_k'(u_k, v_k)$ of the k^{th}-view image by Eq. 7.

$$p_k' = \tilde{P}_k' \cdot P_s \tag{7}$$

where $\tilde{P}_k'$ indicates the projection matrix of the k^{th}-view video camera. In addition, depth information at the image position p_k' is equal to the calibrated depth value $D_s(p_{sx}, p_{sy})$. Figure 5 shows the result of 3-D warping using foreground and background depth maps, respectively.



(a) Foreground (b) Background

Figure 5. 3-D warped depth map

4.2 Multiview Depth Estimation

In this work, we measure depth information of each view image using the warped depths of the depth camera. In order to obtain depth maps in the multiview image, we first segment the multiview image by a mean-shift color segmentation algorithm. Then, we define the initial depth of each segment as the average of 3-D warped depths on the segment by assuming that each segment has one disparity value [8]. Thereafter, we convert the initial depth into its disparity by Eq. 8.

$$InitDisparity(x, y) = \frac{K \cdot B}{InitDepth(x, y)} \tag{8}$$

where $InitiDisparity(x, y)$ is the converted disparity at the pixel position (x, y) from the corresponding initial depth $InitDepth(x, y)$. B and K indicates the baseline distance for video cameras and the focal length of the video camera, respectively.

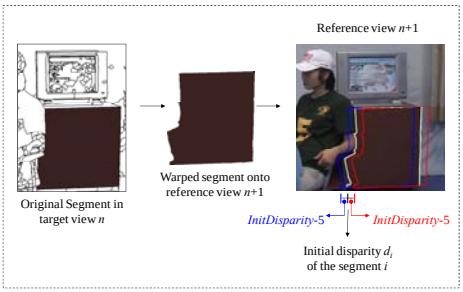
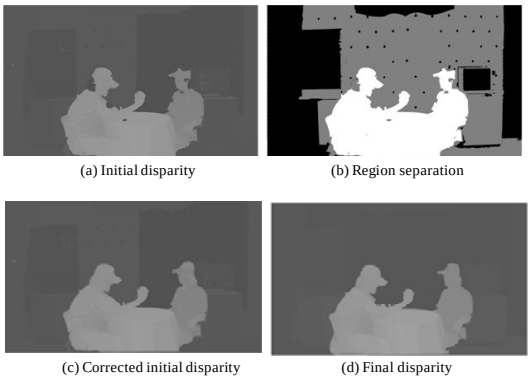


Figure 6. Color segment-based depth estimation

As shown in Fig. 6, since the shape of warped segment generated by its initial depth in the target view image n is similar to the shape of its corresponding segment in reference view image n+1, we find out the corresponding segment by moving the warped segment on the reference view image n+1. The search range to estimate disparities of target view image n is from $InitDisparity$-5 to $InitDisparity$+5.

For determining the disparity of each segment, we calculate the sum of absolute difference (SAD) values between the warped segment of the target view image n and its matched region in the reference view image n+1. The disparity with the minimum SAD is chosen as the initial disparity d_i of the segment i in the target view image n.

In practice, since the depth quality of the initial disparity map generated by the depth estimation method is usually low, it is hard to use it as the multiview disparity map for the multiview image. In order to enhance the initial disparity map, we refine it according to regions: foreground, background, and unknown region. The region of foreground and background are the set of segments in the region warped by the foreground and background depth maps captured by the depth camera, respectively. The unknown region is defined as the set of segments on the boundary of both foreground and background.



(a) Initial disparity (b) Region separation

(c) Corrected initial disparity (d) Final disparity

Figure 7. Region separation for disparity correction

4.3 Depth Map Refinement

After obtaining the initial disparity map for each view image, we refine the disparity map using belief propagation (BP). Figure 7(d) shows the final disparity map. Figure 8 shows the result of disparity map refinement. As shown in Fig. 8(a), there are some mismeasured disparities in the black circle. After disparity map refinement using BP with consideration of the initial disparity generated from the depth camera data, we can notice that the disparity errors are minimized as shown in Fig. 8(b).

(a) Before belief propagation (b) After belief propagation

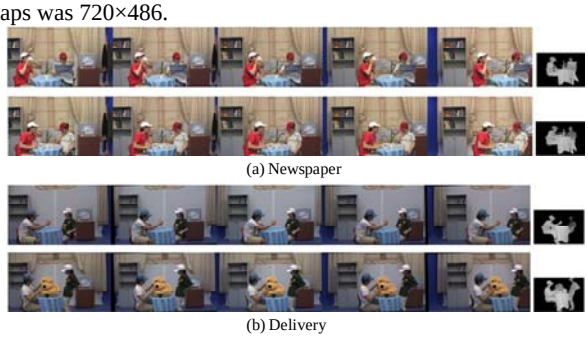
Figure 8. Refinement of initial disparity map

Conventional stereo matching algorithms are especially useful to estimate depth information for the continuous region. However, they are inappropriate for discontinuous regions due to the disocclusion problem. Feature-based estimations are comparably more effective for boundary areas than the area-based estimations like stereo matching, because the matching process is based on features, such as corner points, edges, and lines. In this paper, we generate the edge segments using scale space in multiview images,

and then make the 3-D edge segments using stereo matching based on the edge segments. In the step of depth refinement, we utilize the depth information of the 3-D edge segments to enhance the quality of the depth map.

5. EXPERIMENTAL RESULTS

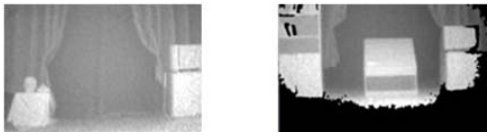
In order to evaluate the performance of the proposed method, we have constructed a hybrid camera system with five HD cameras and one depth camera as shown in Fig. 1. The measuring distance for depth information of the depth camera was from 1.75m to 6.05m. The baseline distances for the HD cameras was 20cm. Figure 9 shows the test multiview image and depth map sequences, 'Newspaper' and 'Delivery', captured by the hybrid camera system, respectively. In addition, Figure 10 shows the depth map for background of the test image. The resolution of the test multiview images was 1920×1080, and the resolution of the depth maps was 720×486.



(a) Newspaper

(b) Delivery

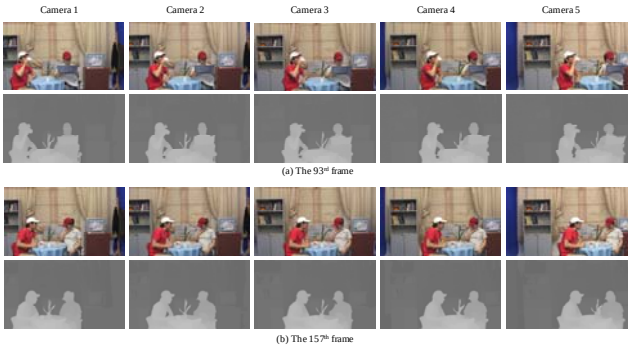
Figure 9. Test multiview image and its depth map



(a) Newspaper (b) Delivery

Figure 10. Background depth map

Figure 11 and Figure 12 show the final multiview disparity maps for the 93rd, 157th frames of 'Newspaper' and 87th, 149th frames of 'Delivery'. As shown in Fig. 11, we could notice that depths for the orchid in the flowerpot in the scene of 'Newspaper' were generated successfully, although the boundary of the orchid was sharp.

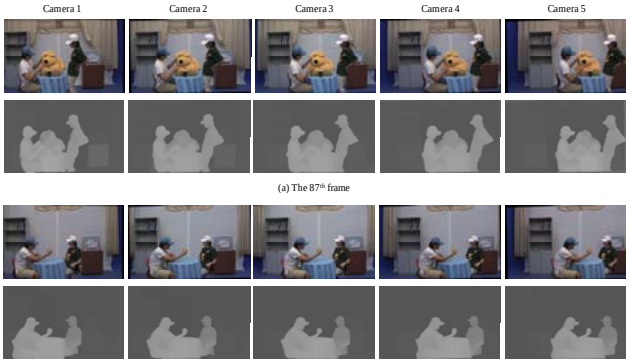


(a) The 93rd frame

(b) The 157th frame

Figure 11. Results of multiview disparity maps for 'Newspaper'

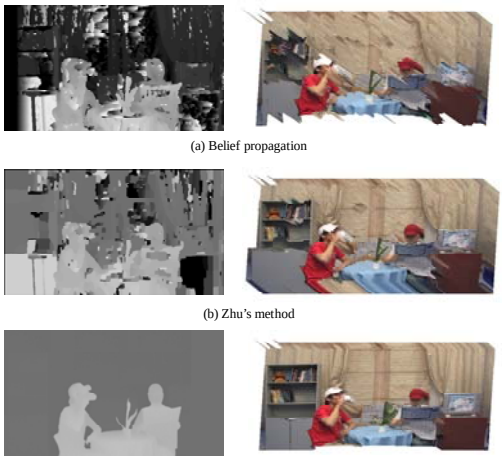
In addition, as shown in Fig. 12, the depth quality of the yellow bear doll was good, although the color of the bear was monotonous. As a result, we could overcome the two main problems of passive depth sensing, depth estimation on the occluded and textureless regions, using the depth camera data as a supplement. To compare the depth quality of the proposed method with previous works, we have compared the disparity map generated by the conventional BP algorithm and Zhu's method [15] with the 3rd view image of the 93th frame in 'Newspaper'.



(a) The 87th frame

(b) The 149th frame

Figure 12. Results of multiview disparity maps for 'Delivery'



(a) Belief propagation

(b) Zhu's method

(c) Proposed method

Figure 13. Depth comparison with the previous works

The generated disparity maps using previous methods and the proposed method are shown in the first column from Fig. 13(a) to Fig. 13(c). We could notice that some regions of the disparity maps generated by the previous methods had mismeasured depths. On the other hand, the mismatched disparities were notably reduced by the proposed method. As a result, the proposed method outperformed the previous methods.

To evaluate the subjective quality of the proposed method, we constructed the 3-D scene with the generated disparity map. The second column from Fig. 13(a) to Fig. 13(c) shows the results of 3-D scene construction with disparity maps of the first column from Fig. 13(a) to Fig. 13(c). We employed hierarchical decomposition of depth maps for 3-D scene construction [16].

As shown in the result of 3-D scene reconstruction, we could subjectively notice that the disparity map obtained by the

proposed method had more reliable depth data than the previous method. In addition, the 3-D surfaces generated by the proposed method were smoother than ones generated by the other methods.

We have also generated intermediate views using the disparity maps and images of camera 3 and camera 4 view of the 87th frame in 'Delivery'. In order to construct virtual views with the estimated multiview video-plus-depth, we employed a view synthesis algorithm [17]. In this experiment, we have generated 15 view images between camera 3 and camera 4 by moving a virtual camera with 1 degree interval. As shown in Fig. 14, we could generate intermediate views successfully and provide natural 3-D video service to potential consumers.

(a) Newspaper

(b) Delivery

Figure 14. Intermediate views using generated depth maps

6. CONCLUSION

In this paper, we have presented a new approach to generate HD depth maps corresponding multiview HD color image using a hybrid camera system. We have used depth information acquired by a depth camera to generate the initial disparity maps by 3-D warping and generated the final disparity maps using a segmentation-based stereo matching and the belief propagation algorithm. Experimental results have shown that our scheme has produced more reliable depth maps compared than previous methods. With the hybrid camera system, we could solve the two main problems in the current passive depth sensing and depth estimation on the occluded and textureless regions. We have generated high-resolution and high-quality 3-D video from our system. Therefore, our proposed system could be useful for various 3-D multimedia applications.

7. ACKNOWLEDGMENTS

This work was supported in part by ITRC through RBRC at GIST (IITA-2009-C1090-0902-0017) and in part by the Basic Research Program of the Korea Science & Engineering Foundation (R01-2007-000-20330-0).

8. 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, 2007.

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

[3] ISO/IEC JTC1/SC29/WG11 N8944, "Preliminary FTV Model and Requirements," 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] Canestavision^{TM} electronic perception development kit, canesta inc. http://www.canesta.com/

[6] C. S. Swiss Ranger SR-2. The swiss center for electronics and microtechnology. http://www.csem.ch/fs/imaging.htm.

[7] M. Okutimi and T. Kanade, "A multiple-baseline stereo," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 4, pp. 353–363, 1993.

[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, 2004.

[9] 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, 2001.

[10] M. Kawakita, T. Kurita, H. Kikuchi, and S. Inoue, "HDTV axi-vision camera," *Proc. of International Broadcasting Conference*, pp. 397–404, 2002.

[11] G. Um, K. Kim, C. Ahn, and K. Lee, "Three-dimensional scene reconstruction using multiview images and depth camera," *Proc. of 3D Digital Imaging and Modeling*, pp. 271–280, 2005.

[12] 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, 2005.

[13] 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, 2001.

[14] M. Kawakita, T. Kurita, H. Kikuchi, and S. Inoue, "HDTV axi-vision camera," *Proc. of International Broadcasting Conference*, pp. 397–404, 2002.

[15] 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, 2008.

[16] S. Kim, S. Lee, and Y. Ho, "Three-dimensional natural video system based on layered representation of depth maps," *IEEE Trans. on Consumer Electronics*, vol. 52, no. 3, pp. 1035–1042, 2006.

[17] C. Lee, K. Oh, S. Kim, and Y. Ho, "An efficient view interpolation scheme and coding method for multiview video coding," *Proc. of International Conference on Systems, Signals and Image Processing*, pp. 107–110, 2007.