HIGH-RESOLUTION DEPTH MAP GENERATION BY APPLYING STEREO MATCHING BASED ON INITIAL DEPTH INFORMATON

[1]*Eun-Kyung Lee,* [1]*Sung-Yeol Kim,* [2]*Young-Ki Jung, and* [1]*Yo-Sung Ho*

[1]Gwangju Institute of Science and Technology (GIST), Korea
[2]Honam University, Korea

ABSTRACT

In this paper, we propose a new algorithm to generate a high definition (HD) depth map using a standard definition (SD) depth camera and HD stereoscopic cameras. In order to obtain the initial depth information for the left camera, we perform the three-dimensional (3-D) warping technique using depth information acquired by a depth camera. After depth values from a depth camera are converted into initial disparity values for stereo matching, we search only a small neighboring region of the initial disparity to find a more accurate disparity value. Then, we generate 3-D edge segments to refine the depth map in boundary areas of objects. Experimental results show that the proposed scheme generates a more accurate depth map than traditional stereo matching algorithms.

Index Terms— depth map generation, 3-D video, stereo matching

1. INTRODUCTION

As natural and realistic services are expected to become available in near future, we are very interested in the three-dimensional (3-D) video as high-quality visual media. ISO/IEC JTC1/SC29/WG11 Moving Picture Experts Group (MPEG) has recognized the importance of multi-view video with depth (MVD) for 3-D video (3DV) applications [1]. In MPEG, several state-of-the-art techniques have been proposed to generate accurate depth information [2, 3].

We can classify 3-D depth sensor techniques into two categories: active depth sensor methods and passive depth sensor methods. In order to obtain 3-D depth information from real scenes directly, active depth sensor methods employ optical sensors, such as a laser sensor, an infrared ray sensor, or a light pattern sensor. Examples of active depth sensor methods are the structured light pattern approach [4] and the depth camera approach [5]. Although the active depth sensor methods only support low-resolution depth maps and need much cost to get them, they can produce accurate depth information.

Passive depth sensor methods indirectly obtain depth information of real scenes using images captured by cameras. Examples of passive depth sensor methods can be stereo matching [6] and shape from motion [7]. The main advantage of passive depth sensor methods is that we can get high-resolution depth maps without paying much cost. However, the accuracy of depth information generated by the passive depth sensing is relatively lower than the other.

In order to generate more accurate 3-D information, hybrid depth sensor methods that combine active and passive depth sensor methods have been tried. Most hybrid depth sensor methods usually refine 3-D models or scenes with images after constructing them with the 3-D information acquired by a 3-D scanner [8]. However, there exists the limitation in the previous hybrid works for scanning real scenes in real time.

In 2005, Um et al. proposed a hybrid method that combines multi-view camera and a depth camera [9]. They works could get the depth information of out of the capturing range of a depth camera by compensated depth information acquired from stereo matching. However, this hybrid camera system could not support high-resolution depth maps because it was not dependent on the multi-view camera system but the depth camera system.

In other words, the output of the previous work was standard definition (SD) depth maps that the current depth camera only supports. We can upgrade the current depth camera and develop a new depth camera to support high definition (HD) or higher resolution depth maps. Therefore, it is reasonable to develop a new scheme to generate larger size depth maps using the existing depth camera.

In this paper, we propose a new scheme to generate high-resolution and high-quality depth maps by combining high definition (HD) stereo cameras and a SD depth camera. In this work, we regard the depth information acquired by carrying out 3-D warping with the data from the depth camera as the initial depth values during stereo matching. Then, we refine the estimated depth information obtained from a stereo matching algorithm using 3-D edge segments.

Our main contribution of this work is that we find out the generation method for high-resolution depth maps using the proposed hybrid camera system. Instead of developing a new depth camera with tremendous cost and strenuous efforts, the hybrid camera system can produce high-resolution depth maps for the real scenes.

978-1-4244-1755-1/08/$25.00 ©2008 IEEE

2. PROPOSED HYBIRD CAMERA SYSTEM

2.1. Structure of the hybrid camera system

In this paper, we focus on generating a HD depth map at the position of the left camera of the stereo cameras. Figure 1 shows the main components of the proposed hybrid camera system. Basically, the hybrid camera system consists of HD stereo cameras and a SD depth camera. Every camera is connected to a sync generator to get synchronized image sequences continuously. The hybrid camera system provides left and right images acquired from stereo cameras and a color image and a depth map acquired from a depth camera for each frame. Figure 2 shows four different images produced by the hybrid camera system.

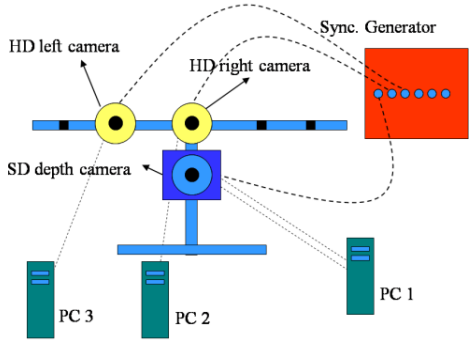
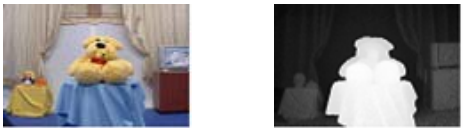


Fig. 1. Proposed hybrid camera system

(a) Left image (b) Right image



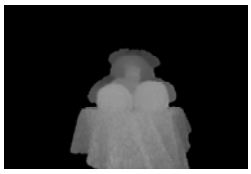
(c) Color image

(d) Depth map

Fig. 2. Input images

2.2. Depth map correction

Although the ideal depth range of the current depth camera is usually from 0.5m to 7m, the measurable distance is from 2m to 4m in practical environments. In order to increase the measuring distance, we capture depth maps for foreground and background separately, as shown in Fig. 3(a) and Fig. 3(b). Then, the foreground and background depth maps are merged into a depth map and quantized to fit the depth range from 0 to 255, as shown in Fig. 2(d).



(a) Foreground depth map

(b) Background depth map

Fig. 3. Depth data by the depth camera

3. DEPTH MAP ESTIMATION

3.1. Generation of initial disparity using 3-D warping

We regard depth information acquired from a depth camera as initial depth information for the left image. In order to match the depth information with its corresponding color value in the left image, we perform the camera calibration for the left camera and the depth camera independently. Equation (1) and Equation (2) is the projection matrix Ps and P_l of the depth camera and the left camera, respectively

$$P_s=K_s[R_s|t_s] \quad (1)$$

$$P_l = K_l[R_l|t_l] \quad (2)$$

where K_s is the intrinsic parameter, R_s and t_s are the extrinsic parameter of the depth camera. K_l is the intrinsic parameter, R_l and t_l are the extrinsic parameter of the left camera. In order to calculate the relative positions between the depth camera and the left camera, we move the depth camera to the principal point in the world coordinate by

$$R'_{ori} = R_s R_s^{-1}=I$$
$$t'_{ori}=t_s-t_s=0 \quad (3)$$

Then, we determine the new left camera position based on the depth camera position with Eq. (4). We multiply the rotation matrix R_l of the left camera by the inverse matrix R_s^{-1} of rotation matrix R_s of the depth camera. t'_l is the translational difference between t_l and t_s .

$$R'_l=R_l \cdot R_s^{-1}$$
$$t'_l=t_l-t_s \quad (4)$$

The 3-D warping matrix to move pixels from the SD depth camera to the HD left camera is run by Eq. (5)

$$p_l=P'_l \cdot P_s^{-1} \cdot p_s \quad (5)$$

where $p_l = (p_{lx}, p_{ly}, 1)$ is the image coordinate in the left image corresponding to the p_s, and the depth information $D(p_{lx}, p_{ly})$ of p_l is followed by Eq. (6).

$$D_l(p_{lx}, p_{ly})=(t_{lz} - t_{sz})+D_s(p_{sx}, p_{sy}) \quad (6)$$

The 3-D warped depth information is used to the initial depth information of the left image. Figure 4 is the result of the initial depth map of the left camera.

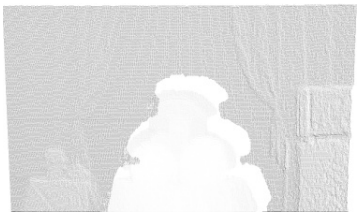
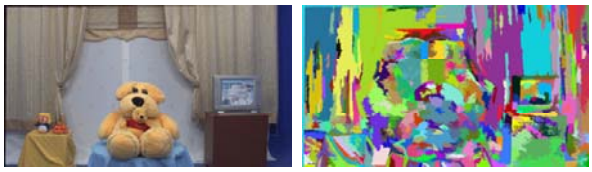


Fig. 4. 3-D Warped depth map

3.2. Stereo matching based on color segmentation

A bilateral smoothing filter is applied to left and right images to remove noises prior to color segmentation. Then, we generate color segments using the graph-cut method [10]. Figure 5 shows the results of color segmentation.



(a) Left image (b) Segmented result

Fig. 5. Color segmentation

Since each segment has smooth changes of colors, we assume that each segment has one disparity value. In order to determine the initial disparity of each segment with the 3-D warped depth information, the initial depth values are converted into initial disparities by Eq. (7)

$$d_l=\left(p_{l_x},p_{l_y}\right)=\frac{K_{l_x}\cdot B}{D_l(p_{l_x},p_{l_y})} \tag{7}$$

where $d_i(p_{lx},p_{ly})$ is the converted disparities from the position (p_{lx}, p_{ly}) of the initial depth map. We average all depth values included in color segment s_i to get the initial disparity $d(s_i)$ for each color segment using Eq. (8)

$$d(s_i)=\frac{1}{n(A(s_i))}\sum_{j=1}^{n(A(s_i))} d_j(A(s_i)) \tag{8}$$

where $n(A(s_i))$ is the number of pixels of each segment and $d_j(A(s_i))$ is the j^{th} disparity value in the initial depth map. The stereo matching algorithm based on color segmentation finds the corresponding color segments using the initial disparity value in the right image. For determining the valid disparity for each segment, we generate a disparity space distribution (DSD) [6].

In DSD, we set the disparity with the maximum value in DSD. The matching function based on histograms for each segment is defined by Eq. (9)

$$m_{ijk}(d)=\max\left(h_{l-1}+h_l+h_{l+1}+\frac{\lambda dim_{ij}}{|d(s_i)-d|+1}\right) \tag{9}$$

where dim_{ij} is the square root of the number of pixels in segment s_{ij} in the images, $d(s_i)$ is the initial disparity in the segment and h_l is the l^{th} bin in the histogram. Matching function uses a color similarity by rating color intensities of two comparative images.

3.3. Depth refinement using 3-D edge information

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 [11] in the left and right 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,.

4. EXPERIMENTAL RESULTS

For evaluating the performance of the proposed method, we construct the hybrid camera system using a pair of HD cameras and a depth camera, ZCam^{TM} [5]. Figure 6 shows the constructed hybrid camera system, and Table 1 shows its specification. The depth range of the depth camera was from 164 cm to 606 cm. The resolution of the final depth map was 1920×1080.

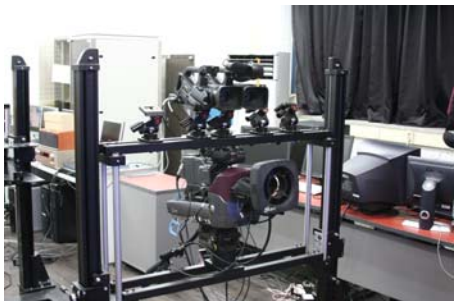


Fig. 6. Construction of hybrid camera system

Table 1. Specification of hybrid camera system

	Description
HD Camera (Canon XL-H1)	NTSC or PAL (16:9 ratio, High Definition)
Depth Camera (ZCam.^{TM})	0.5 to 7.0m
	40 degrees
	NTSC or PAL (4:3 ratio, Standard Definition)
Sync. Generator	SD/HD Video Generation

In order to calculate error statistics with respect to the known ground truth data, we compute the following three quality measures.

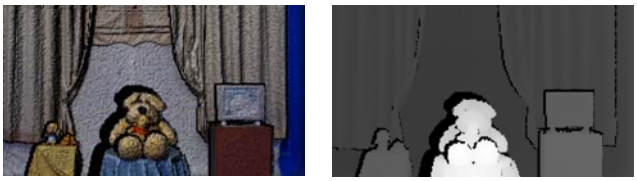
$$B_O = \sum_{s \in O} (|d(s)-d_T(s)|>\delta_d) \quad (10)$$

$$B_T = \sum_{s \in T} (|d(s)-d_T(s)|>\delta_d) \quad (11)$$

$$B_D = \sum_{s \notin D} (|d(s)-d_T(s)|>\delta_d) \quad (12)$$

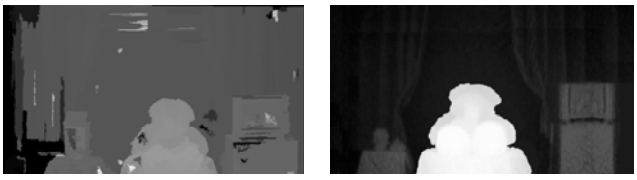
Performance of our depth generation algorithm is confirmed by quality measures proposed by Scharstein and Szeliski based on the known ground truth data using from Eq. (10) to Eq. (12). B_T represents performance in textureless regions and B_D shows performance in depth discontinuity regions. In particular, B_O presents overall performance of the proposed algorithm. For quantitative evaluation, we use the percentage of wrong pixels. Error percentages are computed for four different image regions.

Figure 7 shows the 3-D scene modeling and the ground truth from 2-D scan data. Figure 8 shows comparison of result of the stereo matching method and result of the proposed method. Table 2 shows quantitative results for the proposed method. The proposed method with the hybrid camera system using the initial depth information shows better results than the stereo matching algorithm.



(a) 3-D scene modeling (b) Real depth map

Fig. 7. 3-D scan data



(a) Stereo matching (b) Proposed method

Fig. 8. HD depth map

Table 2. Performance using quantitative statistics

	B_O	B_T	B_D
Stereo matching	1.25	1.62	6.68
Proposed	0.88	1.29	4.76

5. CONCLUSION

In this paper, we have solved two problems of previous depth map generation methods using the proposed hybrid camera system. In order to increase the accuracy and reduce the processing time, we have performed the stereo matching using the initial depth map. As a result, we have generated high-quality depth map with image resolution 1920×1080. We have checked the accuracy of the generated depth map in terms of percentage of wrong pixels. The proposed scheme has produced more accurate depth map than conventional algorithms. Therefore, the proposed method can be used in future 3-D multimedia applications.

ACKNOWLEDGMENTS

This work was supported in part by ITRC through RBRC at GIST (IITA-2008-C1090-0801-0017) and in part by the Basic Research Program of the Korea Science & Engineering Foundation (R01-2007-000-20330-0).

REFERENCES

[1] ISO/IEC JTC1/SC29/WG11 N8944, "Preliminary FTV model and requirements," July 2007.
[2] ISO/IEC JTC1/SC29/WG11 M15090, "Improvement of depth map estimation and view synthesis," January 2008.
[3] ISO/IEC JTC1/SC29/WG11 M15119, "Segmented-based multi-view depth map estimation for FTV," 2008.
[4] M. Waschbüsch, S. Würmlin, D. Cotting, and M. Gross, "Point-sampled 3D video of real-world scenes," *Signal Processing Image Communication*, vol. 22, no. 2, pp. 203-216, February 2007.
[5] 3DV systems, http://www.3dvssytems.com, 2005.
[6] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *Proc. of SIGGRAPH*, pp. 600-608, August 2004.
[7] A. Hengel, A. Dick, T. Thormahlen, B. Ward, and P. Torr, "VideoTrace: rapid interactive scene modeling from video," *Proc. of SIGGRAPH*, article 86, July 2007.
[8] P. Dias, V. Sequeira, F. Vaz, and J. Gonçalves, "Registration and fusion of intensity and range data for 3D modeling of real world scenes," *Proc. of 3-D Digital Imaging and Modeling*, pp. 418-426, October 2003.
[9] 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, January 2005.
[10] Efficient Graph-Based Image Segmentation, http://people.cs.uchicago.edu/~pff/segment/
[11] W. Grimson, "Computational experiments with feature based stereo algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 7, no.1, pp. 17-34, January 1985.