Three-dimensional Video Capturing for Realistic Broadcasting Services

Yo-Sung Ho and Eun-Kyung Lee

Gwangju Institute of Science and Technology (GIST)
261 Cheomdan-gwagiro, Buk-gu, Gwangju, Republic of Korea

ABSTRACT

In this paper, we present a new three-dimensional (3-D) video capturing system to provide realistic broadcasting services by integrating multiple high-definition (HD) camera arrays and one standard-definition (SD) depth camera. In the proposed hybrid camera system, we first create the initial disparity for each HD cameras by applying 3-D warping operation on the depth map acquired by the depth camera. Then, the final disparity for each HD camera is obtained by a stereo matching algorithm with the initial disparity. Experimental results show that the 3-D video generated by the hybrid camera system provides reliable depth information for 3-D realistic broadcasting services. Besides, the proposed system minimizes the inherent problems of conventional depth cameras, such as limitation of measuring distance for depth information and generation of low-resolution depth maps.

INTRODUCTION

As the three-dimensional (3-D) video becomes attractive in a variety of multimedia applications, it is essential to get accurate depth information for future 3-D applications. Recently, ISO/IEC JTC1/SC29/WG11 Moving Picture Experts Group (MPEG) has been also recognized the importance of multiview video and depth information, which are often referred as a multiview video-plus-depth, for free-viewpoint TV (FTV) or 3DTV [1] [2]. However, traditional depth estimation methods, such as stereo matching, are limited to estimate accurate depth map due to the failure of correspondence point matching on the textureless and occluded regions.

In general, depth estimation methods can be categorized into two major classes: active range sensors [3] and passive range sensors [4]. Passive depth sensors estimate depth information indirectly from 2-D images captured by two or more cameras. On the other hand, active depth sensors obtain depth information from the natural scenes directly using physical sensors. However, these kinds of TOF depth cameras have some built-in problems: low resolution, noisy, and poorly calibrated.

Recently, fusion methods that combines multiview camera and a TOF depth camera have been introduced [5]. These fusion camera systems generate enhanced depth maps by applying a stereo matching algorithm to multiview image with depth information captured by the TOF depth camera. However, the previous hybrid camera systems have produced low-resolution depth maps and focused on generating depth maps for static scenes. Since most 3-D applications are expected to use high-resolution videos, it is necessary to create a high-quality multiview video-plus-depth for dynamic scenes.

To generate multiview depth map, we propose a new hybrid camera system constructed by one depth camera and multiple video cameras. The proposed system provides multiview high-definition (HD) depth map using depth information acquired from the standard-definition (SD) depth camera as a supplement. The main contribution of this work is that we provide a practical solution to create a high-resolution multiview video-plus-depth using multiple video cameras and a depth camera. In addition, the proposed system reduces the inherent problems of generating depth maps from the currently available depth camera system.

HYBRID CAMERA SYSTEM

The proposed hybrid camera system consists of six cameras; one depth camera, *Z−Cam*^{TM}, and five HD cameras. There is one clock generator sending a synchronization signal constantly. This signal is distributed to all personal computers. Figure 1 shows main components of our hybrid camera system.

We obtain test sequences with 1-D parallel camera arrangement from the hybrid camera system. Input videos consist of seven synchronized images; five HD images from the multiview camera, one SD color image and its corresponding depth map from the depth camera. Since the measureable depth range of *Z−Cam*^{TM} is up to seven meters, depth accuracy is not guaranteed in the practical environments. The depth range becomes bigger as the quality of depth map becomes lower. To obtain a more accurate depth map from the depth camera, we reduce the depth range by capturing foreground and background, respectively. We capture a color image and its depth map for background in advance.

Before obtaining synchronized multiview image from the hybrid camera system, we calibrate each camera using checkerboard patterns independently. There are two coordinate systems: depth camera coordinate system and HD camera coordinate system. Therefore, we register two camera coordinate systems using their relative camera information.

To use the initial disparity information for each HD camera, we perform a 3-D warping operation using camera parameters and depth information acquired from the depth camera. Pixel intensities of the warped data are then used as the initial disparity for each view.

In multiview camera, all HD images are rectified and color-segmented. The initial disparities generated from the depth camera are assigned into the corresponding segments. We separate each HD image into three different regions to detect occluded and disoccluded regions: background, foreground, and unknown regions. The disparity of each segment is independently estimated by a color segmentation-based stereo matching algorithm.

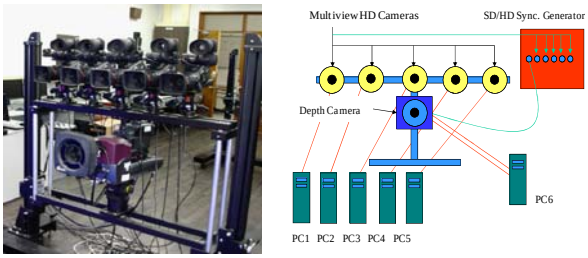


Figure 1. The proposed hybrid camera system

GENERATION OF MULTIVIEW DEPTH MAP

In the hybrid camera system, the intrinsic and extrinsic camera parameter of each camera are different, since we merge two different types of cameras. Therefore, it is essential to find out relative camera information about the camera set using a camera calibration algorithm.

Basically, the camera calibration is executed with pattern images acquired from the hybrid camera. we employ the well-known camera calibration algorithm provided by Caltech to estimate intrinsic and extrinsic parameters of each camera [11]. Since the hybrid camera is composed of six cameras, we carry out the camera calibration as many as the number of cameras independently.

Image rectification is a process that makes epipolar lines of two images captured at different position parallel each other. Vertical coordinates of all image points of two images become identical and there remain the horizontal disparities

only. For multiview image rectification, the horizontal axis of each camera is parallel to the baseline and the principal axis of each camera is perpendicular to the baseline For rectifing multiview image at the same time, we calculate the common baseline considering all camera positions and apply the rectifying transformation defined by camera rotations and camera intrinsic parameters. Then, rectified images have uniform horizontal disparities and no vertical mismatches between

adjacent views [12].

We employ both down-sampling and linear interpolation operations to reduce optical noises in the depth map. After we apply mean filtering on the depth map, the depth data enhancement algorithm executes down-sampling on the mean-filtered depth map. We then perform bilateral filtering on the downsampled depth map. Finally, we recover the depth map using a linear interpolation method.

We regard depth information acquired from the depth camera as initial disparity information for the multiview camera. For matching the depth information with its corresponding color value in the multiview image, we perform camera calibration for the multiview camera and the depth camera independently. To calculate the relative positions between the depth camera and the multiview camera, we move the depth camera to the origin of the world coordinate by

$$\mathrm{R'_{ori}} = \mathrm{R_s R^{-1}{}_s} = \mathrm{I} \quad (1)$$

$$\mathrm{t'_{ori}} = \mathrm{t_s} - \mathrm{t_s} = 0 \quad (2)$$

Then, we determine the new multiview camera position based on that of the depth camera by *Eq. (3)* and *Eq. (4)*. We multiply the rotation matrix R_{hn} of the multiview camera by the inverse matrix R_s^{-1} of the rotation matrix R_s of the depth camera. t'_{hn} is the translational difference between t_{hn} and t_s.

$$\mathrm{R'_{hn}} = \mathrm{R_{hn} R^{-1}{}_{hn}} \quad (3)$$

$$\mathrm{t'_{hn}} = \mathrm{t_{hn}} - \mathrm{t_{s.}} \quad (4)$$

The 3-D warping matrix to move pixels from the SD depth camera to the HD multiview camera is given by

$$\mathrm{p_{hn}} = \mathrm{P'_{hn} P_s^{-1} p_s} \quad (5)$$

where p_{hn} is the image coordinate in the multiview image corresponding to the ps, and the depth information $D(p_{hnx}, p_{hny})$ of p_{hn} is followed by

$$\mathrm{D_{hn}(p_{hnx}, p_{hny})} = (\mathrm{t_{hnz}} - \mathrm{t_{sz}}) + \mathrm{D_s(p_{sz}, p_{sy})} \quad (6)$$

The 3-D warped depth information is used to the initial depth information of the multiview image. To generate the initial disparity map, we apply the 3-D warping in *Eq. (7)* on the color and depth map from the depth camera. In the 3-D warping operation, we project the color and depth data to the world

coordinate, and then reproject the warped 3-D data into each HD camera. Figure 2 shows the 3-D warping results of the foreground and background, respectively.

Figure 2. 3D warped depth map

For the stereo matching operation, we first calculate the average disparity value in each segment. Then, in order to refine its disparity more accurately, we examine the small neighboring area around the initial disparity. The segments of each image are obtained by a mean-shift color segmentation algorithm [13]. Since we have separate initial depth maps for foreground and background, we perform the stereo matching operation on each segmented region independently.

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

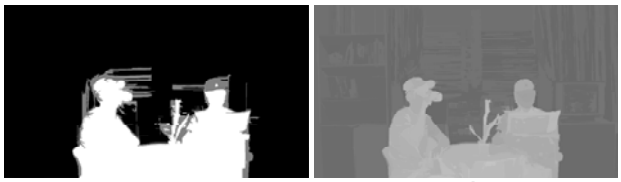
$$d_{hn}(p_{hnx}, p_{hny}) = K_{hn}B/D_{hn}(p_{hnx}, p_{hny}) \qquad (7)$$

where $d_{hn}(p_{hnx}, p_{hny})$ is the converted disparities from the position (p_{hnx}, p_{hny}) of the initial depth map $D_{hn}(p_{hnx}, p_{hny})$, B is the distance between cameras, and K_{hx} is the focal length of the camera. We calculate the average value of depth values included in color segment s_i to get the initial disparity d_{si} for each color segment by

$$d_{(si)} = \sum d_j A(s_i)/n(A(s_i)) \qquad (8)$$

where $n(A(s_i))$ is the number of pixels of each segment and $\sum d(A(s_i))$ is the sum of disparity value of each segment in the initial depth map. As shown in Fig. 5, since there are so many hole in the initial depth map, we only consider the existing disparity values in each segment. The stereo matching algorithm based on color segmentation finds the corresponding color segments using the initial disparity value in the left side and right side images. For determining the valid disparity for each segment, we calculate the sum of absolute difference (SAD) values in small search range. Corresponding pixels from each segment are compared and their differences summed. The lower the SAD the better the match and so the candidate segment with the minimum SAD should be chosen.

In general, the quality of the initial disparity map is coarse in the boundary of foreground objects due to occluded regions. To solve this inherent problem, we extract occlusion and disocclusion regions from the multiview image by searching boundary segments of foreground. These regions are as unknown regions.

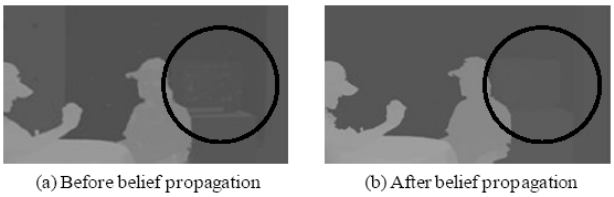


(a) (b)

Figure 3. Refined initial disparity map

Figure 3(a) shows a segmented image of foreground, background, and unknown regions. In order to correct disparities of the unknown regions, we calculate SAD with the initial disparity of the segment in the foreground, and recalculate SAD with the initial disparity of the segment in the background. We regard one of two disparities as the disparity of the segment in the unknown region by comparing their SAD values and choosing the smaller one. Figure 3(b) presents the refined disparity map after solving the occlusion problem using a region separation.

After obtaining the initial disparity map for each view image, we refine the disparity map using belief propagation (BP) [22]. Figure 8 shows the result of disparity map refinement. As shown in Fig. 4(a), there are some mismeasured disparity 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. 4(b).



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

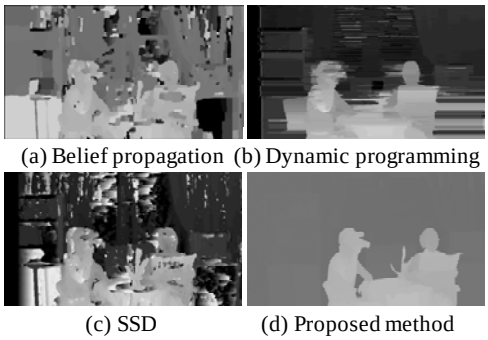
Figure 4. Refined initial disparity map

EXPERIMENTAL RESULT

To generate the multiview depth map, we have constructed a hybrid camera system with five HD cameras and one *Z−Cam*^{TM} as the depth camera. The measuring distance for depth information of the *Z−Cam*^{TM} is from 1.75m to 6.05m. The baseline distances among five HD cameras are 20cm. We

have tested with the '*newspaper*' sequence captured by the proposed hybrid camera system.

Figure 5 shows the results of the final multiview 3-D video using the proposed method. In this experiment, we have generated depth maps using our method. From our experiments, we can observe that the depth map obtained by the proposed method is more reliable than those of the previous stereo matching algorithms.



(a) Belief propagation (b) Dynamic programming

(c) SSD (d) Proposed method

Figure 5. Generated 3-D video for the '*newspaper*'

We obtain test sequences with 1-D parallel camera arrangement from the hybrid camera system. Input videos consist of seven synchronized images; five HD images from the multiview camera, one SD color image and its corresponding depth map from the depth camera. Since the measureable depth range of *Z−Cam*^{TM} is up to seven meters, depth accuracy is not guaranteed in the practical environments. The depth range becomes bigger as the quality of depth map becomes lower. To obtain a more accurate depth map from the depth camera, we reduce the depth range by capturing foreground and background, respectively. We capture a color image and its depth map for background in advance.

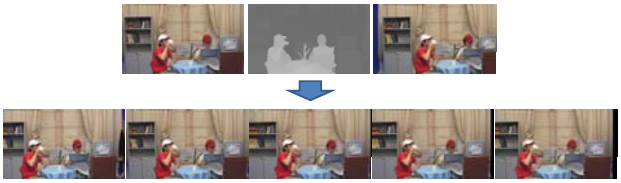


Figure 6. Result of interview images

To evaluate the subjective quality of the proposed method, we construct the 3-D scene with the generated depth map. Figure 3 shows the results of 3-D scene construction for the third view of the 'newspaper' images. We use hierarchical decomposition of depth maps for 3-D scene construction [6]. We also generate intermediate views using the generated depth map. Figure 6 show the results of the generated intermediate view images. As shown in Fig. 7, the generated 3-D scenes and intermediate views have continuous results like natural scene and video.

Figure 7. Results of 3D scene reconstruction

CONCLUSION

In this paper, we have presented a new approach to generate multi-view 3-D video 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 algorithm. Experimental results have shown that our scheme have produced more reliable depth maps compared with previous methods. We have generated high-resolution multiview depth map and natural intermediate views from our system. Therefore, our proposed system could be useful for various 3-D multimedia applications.

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 IITA (Institute for Information Technology Advancement) (IITA-2009 - C1090-0902-0017).

REFERENCES

1. C. Fehn, R. De La Barre, and S. Pastoor, "Interactive 3DTV concepts and key technologies," Processing Image Communication, Vol. 94, No. 3, 524–538 (2006)
2. ISO/IEC JTC1/SC29/WG11 N8944, "Preliminary FTV Model and Requirements," (2007)
3. C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," Proc. of ACM SIGGRAPH, 600–608 (2004)
4. M. Kawakita, T. Kurita, H. Kikuchi, and S. Inoue, "HDTV axi-vision camera," Proc. of International Broadcasting Conference, 397–404 (2002)
5. J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 231–236 (2008)
6. 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, (2006) 1035–1042