Three-dimensional Video Generation for Realistic Broadcasting Services

Yo-Sung Ho, Sung-Yeol Kim, and Eun-Kyung Lee
Gwangju Institute of Science and Technology (GIST)
1 Oryong-dong, Buk-gu, Gwangju, 500-712, Republic of Korea
E-mail: {hoyo, sykim75, eklee78}@gist.ac.kr

Abstract: In this paper, we propose a new scheme to generate multi-view video-plus-depth using a hybrid camera system, which is composed of one depth camera and multiple video cameras. In order to create the three-dimensional (3-D) video, we first calculate the initial disparity for each view by projecting depth camera data onto each video camera using 3-D image warping. Then, a stereo matching algorithm is applyed to estimate the final disparity from the initial disparity in each view. Finally, we convert the final disparity into depth information in each frame to generate a multi-view depth map. Experimental results show that the proposed hybrid camera system not only provides high-qualtiy depth information for 3-D realistic broadcasting services, but also minimizes inherent problems of the conventional depth camera system.

1. Introduction

Three-dimensional (3-D) video has been recognized as one of the essential parts for next-generation visual media. As one of the 3-D video representations, it is widely accepted that a monoscopic color video enriched by depth information, which is often called as *video-plus-depth*, will be used in future 3-D video applications due to backwards compatibility to the current 2-D video systems and easy adaptability to a wide range of different 2-D and 3-D video display systems. Recently, ISO/IEC JTC1/SC29/WG11 Moving Picture Experts Group (MPEG) has been working on multi-view video-plus-depth for various applications, such as interactive 3-D games and 3-D TV services [1].

In 3-D TV research activities, it is quite important to estimate the depth information of natural scenes accurately. In the fields of computer vision and image processing, the state-of-the-art 3-D depth estimation methods have been proposed to obtain accurate depth maps [2]. However, accurate measurement of the depth map from the natural scenes still remains problematic.

As sensor technologies for obtaining depth information from natural scenes are being developed, we can directly measure the depth information in real time using an active range depth camera, such as Z-Cam developed by 3DV Systems, Ltd. [3]. The depth camera integrates a high-speed pulsed infrared (IR) light source with a conventional broadcast TV camera to capture color images and their associated depth maps simultaneously. However, although the depth camera can produce useful depth information, there are some inherent problems in the currently available depth camera system.

The first problem is that the depth map generated by Z-Cam usually includes optical noises. The second problem is that the depth camera has limitations in measuring the distance of objects in the scene. In practice, the range of the depth measurement by Z-Cam is from 1m to 4m. Another problem is that the depth camera can generate only low-resolution depth maps. Typically, the resolution of the depth map acquired by Z-Cam is 720×486. One solution to produce high-quality and high-resolution depth maps is upgrading the current depth camera. However, due to various challenges in real-time distance measuring systems, upgrading and improvements of depth cameras are very slow and expensive.

As an alternative approach to develop a new depth camera, a fusion method that combines one depth camera and multi-view cameras, has been introduced recently [4]. This hybrid camera system generates enhanced depth maps by applying a stereo matching algorithm to multi-view images with depth information captured by the depth camera. However, since this fusion system completely depends on the low-resolution depth camera, it fails to produce high-resolution depth maps.

In this paper, we propose a new scheme to generate multi-view video-plus-depth using multiple high-definition (HD) video cameras and the standard-definition (SD) Z-Cam. In the hybrid camera system, the depth camera plays a role as a supplement for depth estimation. The proposed scheme generates multi-view HD depth maps by refining the depth map initially acquired by the SD depth camera.

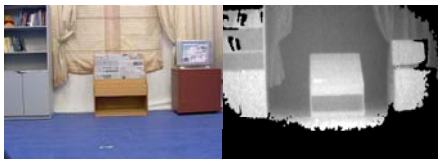
2. Hybrid Camera System

We construct a hybrid camera system combining five HD cameras with one depth camera, as shown in Fig. 1. Each camera in the hybrid camera system is connected to a personal computer equipped with a video capturing board. Besides, a common clock generator is linked to all cameras and provides them with synchronization signals constantly.

Figure 1. Hybrid camera system

(a) Multi-view images

(b) Foreground



(c) Background

Figure 2. Images from the hybrid camera system

With the hybrid camera system, we can capture seven synchronized images: five HD images from the multi-view cameras, one SD color image and its corresponding depth map from the depth camera. Since the depth camera is only avaliable in indoor environments and its range of distance measurement is limitted, we obtain a color image and its depth map for background separately. Figure 2 shows all the nine images captured by the hybrid camera system.

Figure 3 describes the overall framework of the hybrid camera system that can generate multi-view video-plus-depth information. At the preprocessing stage, after we calibrate HD cameras independently, we calculate their relative camera information to the position of the depth camera. We also apply a 3-D image warping operation onto the depth information obtained by the depth camera to project it into the world coordinate. Then, we re-project the warped depth information into the position of each HD camera one by one.

After HD images captured by the multi-view camera system are rectified and color-segmented, we assign the initial disparity values that are obtained by the depth camera into the corresponding image segments. Then, we partition each HD image into three disjoint areas: background, foreground, and unknown regions, in order to minimize the disocclusion problem. Finally, the disparity of each image segment is estimated by a stereo matching algorithm independently.

3. Mulit-view Depth Map Generation

Since we are employing two different types of cameras, we need to calculate relative camera information. For the relative camera calibration, we utilize a camera calibration toolbox [5]. After the camera calibration is completed, five HD images are rectified and color-segmented.

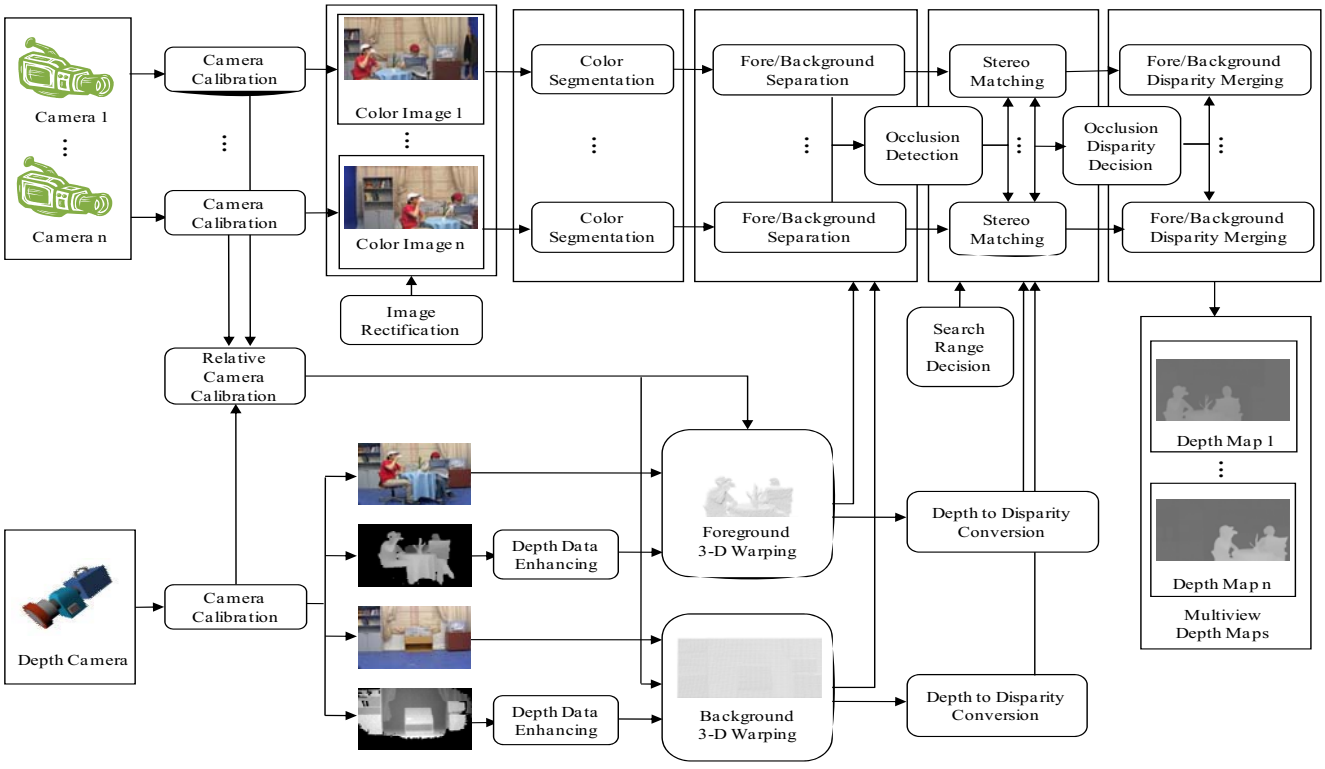


Figure 3. Overall framework for the generation of multi-view depth maps

In multi-view image rectification, we find a common baseline that minimizes the sum of squared distances to the camera centers, and transform multi-view images into virtual camera images that are located on the baseline [6]. In addition, we apply a depth data enhancement algorithm to reduce optical noises in the depth map captured by the depth camera [7].

In order to generate the initial disparity map, we apply a 3-D image warping algorithm on the depth camera data. When $D_s(p_{sx}, p_{sy})$ is the depth information at the pixel position (p_{sx}, p_{sy}) of the depth map, we can regard the pixel p_s $(p_{sx}, p_{sy}, D_s(p_{sx}, p_{sy}))$ as a point in the 3-D space. The corresponding point p_n of the n^{th} HD image is calculated by

$$p_n = \widetilde{P}_n' \cdot P_s^{-1} \cdot p_s \tag{1}$$

where p_n $(p_{nx}, p_{ny}, 1)$ includes the corresponding pixel position (p_{nx}, p_{ny}) of the pixel p_s in the n^{th} HD image.

Then, the depth information $D_n(p_{nx}, p_{ny})$ of p_n is calculated by

$$D_n(p_{nx}, p_{ny}) = (\widetilde{t}_{n_z} - t_{x_z}) + D_s(p_{sx}, p_{sy}) \tag{2}$$

where $\widetilde{t}_{nz}$ and t_{sz} indicate the third value of the transition matrix of the n^{th} camera $\widetilde{t}_n$ and the transition matrix of the depth camera t_s, respectively.

In general, the matching failure on occluded regions is a big problem in the stereo matching operation. In this paper, each HD image is separated into foreground, background, and occluded regions, as shown in Fig. 4(a). Occlusion regions are detected by merging neighboring image segments between foreground and background.

After region separation, we calculate the sum of absolute differences (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 occlusion region by comparing their SAD values and choosing a smaller one. Figure 4(b) shows the refined disparity map after solving the occlusion problem using region separation.

(a) Region separation (b) Refined disparity map

Figure 4. Solving occlusion problems

In stereo matching, we first calculate the average disparity value in each segment with the corresponding disparity map generated by 3-D image warping. Then, we examine a small neighboring area around the initial disparity to refine its disparity more accurately. Since we have separate initial disparity maps for foreground and background, we perform the stereo matching operation on each segmented region independently.

The final step is refining the initial depth map. For each pixel in the n^{th} image, we calculate SAD values along with the search range in block by block by

$$\mathrm{SAD} = \left[\sum_{j=\left\lfloor -\frac{w}{2}-1 \right\rfloor}^{\left\lfloor \frac{w}{2}+1 \right\rfloor} \sum_{i=\left\lfloor -\frac{w}{2}-1 \right\rfloor}^{\left\lfloor \frac{w}{2}+1 \right\rfloor} \left| I_n(i,j) - I_{n+1}(i,j) \right| \right] \tag{3}$$

where w is the block size and d means the refined disparity value with a small search range.

Then, we find the block having the smallest SAD value. Thereafter, we check the disparity of the block and update the disparity of the pixel in the block by the computed one. We repeat this operation for each pixel in the n^{th} image. For boundary areas that form incomplete blocks, we only use pixels available inside the blocks. We can get the final disparity map through these procedures.

4. Experimental Results

In order to evaluate the proposed 3-D video generation scheme, we have constructed a hybrid camera system, as shown in Fig. 1. In the hybrid camera system, the distance range that we can measure by Z-Cam was from 1.75m to 6.05m, and the baseline distance between adjacent HD cameras was 20cm.

In our experiment, we captured five image sequences using the hybrid camera system. Once we had five-view HD images, we generated multi-view HD depth maps by the proposed scheme. We could generate multi-view HD depth maps in near real time, if we apply the stereo matching method only to refine the initial depth map captured by Z-Cam directly.

Figure 5 shows the multi-view depth maps generated by our hybrid camera system for the NEWSPAPER images. In order to evaluate the quality of generated multi-view depth maps, we generated intermediate views using the depth map generated by the proposed scheme. As shown in Fig. 6, boundary regions in our depth maps were clearer. Especially, regions of the background curtain in the NEWSPAPER image contained accurate depth data. Therefore, we could generate intermediate views faithfully.

Figure 7 shows the result of 3-D scene modeling for the 3rd view of NEWSPAPER images. We used hierarchical decomposition for the 3-D scene reconstruction [8]. We have observed that the depth map obtained by the proposed scheme had reliable depth information subjectively.

5. Conclusions

In this paper, we proposed a new scheme that can generate multi-view video-plus-depth using a hybrid camera system. Experimental results demonstrated that we could generate high-resolution multi-view depth maps successfully. Since we refined the depth map that was initially acquired by a depth camera using various image processing techniques, the final depth information was quite accurate. We think that the proposed hybrid camera system can be exploited for 3-D video generation in 3-D multimedia applications.

Figure 5. Result of multi-view video-plus-depth generation

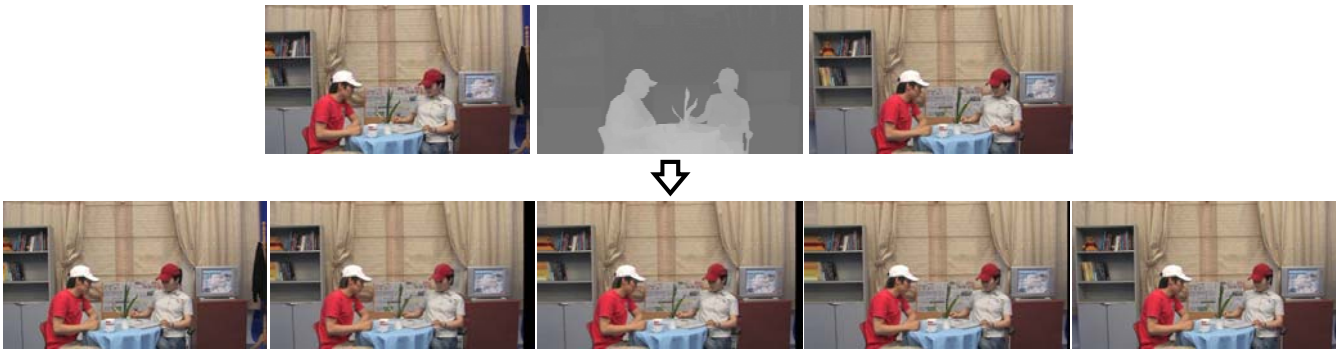


Figure 6. Intermediate view generation: from view 2 to view 3

Figure 7. 3-D scene reconstruction for view 3

Acknowledgement

This work was supported in part by ITRC through RBRC at GIST (IITA-2008-C1090-0801-0017).

References

[1] C. Fehn, R. de la Barré, and S. Pastoor, "Interactive 3-DTV- Concepts and Key Technologies," Proceedings of the IEEE, vol. 94, no. 3, pp. 524-538, 2006.

[2] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms," International Jour. of Computer Visioin, vol. 47, no. 1-3, pp. 7-42, 2002.

[3] G.J. Iddan and G. Yahav, "3D Imaging in the Studio and Elsewhere," Proc. of SPUE Videometrics and Optical Methods for 3D Shape Measurements, pp. 48-55, 2001.

[4] G. Um, K.Y. Kim, C. Ahn, and K.H. Lee, "Thee-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, 2005.

[5] Camra Calibraion Toolbox for Matlab by Caltech, http://www.vision.caltech.edu/bouguetj/calib_doc/

[6] Y.S. Kang, C. Lee, Y.S. Ho, "An Efficient Rectification Algorithm for Multi-view Images in Parallel Camera Array," 3DTV Conference, pp. 61-64, 2008.

[7] J.H. Cho, I.Y. Chang, S.M Kim, and K.H. Lee, "Depth Image Processing Technique for Representing Human Actors in 3DTV Using Single Depth Camera," Proc. of 3DTV Conference, Paper no. 15, 2007.

[8] S.Y. Kim, S.B. Lee, and Y.S. 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.