International Conference on Intelligent Information Hiding and Multimedia Signal Processing

Segment-based foreground object disparity estimation using Zcam and multiple-view stereo

Tae Hoon Kim, Hoyub Jung, Kyoung Mu Lee, and Sang Uk Lee
School of EECS, ASRI
Seoul National University,151-742, Seoul, Korea
thkim@diehard.snu.ac.kr, hoyub@diehard.snu.ac.kr, kyoungmu@snu.ac.kr, sanguk@ipl.snu.ac.kr

Abstract

3D videos play an important role in adoption of 3DTV display modules for the masses because creating realistic contents for 3DTV is a hard and time-consuming process. In this paper, we consider the problem of generating three-view 3D video depth using a segment-based stereo algorithm and a depth camera (ZCam), and propose a new foreground/background video segmentation algorithm as well as a new segment-based stereo algorithm. By combining the depth camera (ZCam) and the stereo algorithm simultaneously, we can obtain better results compared to employing only a single method.

1. Introduction

We believe that three-dimensional television (3DTV) will be the next logical betterment of display technology for more natural and realistic visualization. 3DTV supports a natural viewing experience in depth domain and allows viewers to experience more naturalistic scenes. However, difficulties of 3DTV lie not only in the display techniques but also in the technology of generating multi-view depth video contents. Therefore, 3D depth generation cameras have been often proposed for efficient creation of 3DTV contents.

In general, 3D video generation can be obtained by using two kinds of videos: multi-view texture videos and their depth videos. In this paper, we describe a system for depth generation according to each viewpoints. There are many methods for estimating a depth at different viewpoints. Generally, two methods are used: the depth sensing camera or the stereo algorithm. Recently, the depth camera such as ZCam [?] provides very accurate depth information at high speed and high depth resolution. At the same time, it provides synchronized and synthesized color (RGB) video. However it can only be used for measuring the small range of depth of foreground objects, because of Zcam's very limited range. Otherwise, the stereo algorithm compares all pixels between two textured images, and theocratically we can obtain the depth map of the entire image using two typical RGB cameras. However, there are many difficulties in using two cameras for video stereo vision. Usually two cameras do not provide reliable color matching. Also, lighting variation at each camera usually hinders correct correspondence between stereo images. Overall, it is very hard to obtain accurate depth map based on the corresponding points between two texture images.

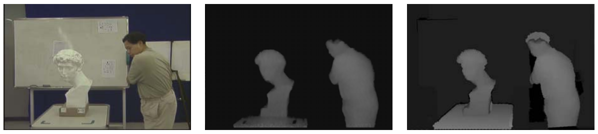

(a) texture image (b) ZCam depth (c) Our result

Figure 1. An example of depth images: ZCam depth and our result (ZCam + Stereo).

In this paper, we deal with a three-view depth estimation system by simultaneously using both single depth camera (ZCam) and stereo algorithm. There are three main contributions. First, we propose the foreground/background video segmentation approach using ZCam depth data. Also, we formulate new segment-based stereo algorithm. Finally, we introduce the efficient fusion of ZCam depth and stereo disparity. Fig. 1 shows that compared with ZCam depth, our fused depth produces better performance in each frame.

2 Proposed algorithm

An overview of our algorithm is shown in Fig. 2. We used three cameras: two RGB Camera and one ZCam as in Fig. 3(a). Therefore, we initially obtain three textured images and one ZCam depth image. Since ZCam can only

978-0-7695-3278-3/08 $25.00 © 2008 IEEE
DOI 10.1109/IIH-MSP.2008.343

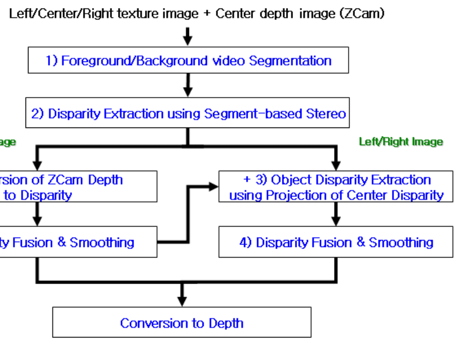
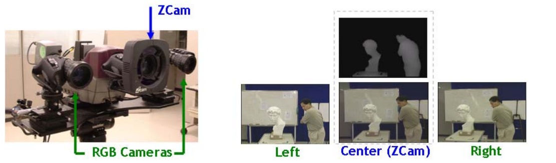

Left/Center/Right texture image + Center depth image (ZCam)
1) Foreground/Background video Segmentation
2) Disparity Extraction using Segment-based Stereo
Center Image
Left/Right Image
+ Conversion of ZCam Depth to Disparity
+ 3) Object Disparity Extraction using Projection of Center Disparity
4) Disparity Fusion & Smoothing
4) Disparity Fusion & Smoothing
Conversion to Depth

Figure 2. A overview of our algorithm.



(a) Camera system (b) Acquired images from (a)

Figure 3. An example of acquired images from our camera system.

acquire the depth within a limited range, we first adjust the range to cover the foreground objects. An example of acquired images from this camera system is shown in Fig. 3(b). Since the ZCam is at the center, the left/right projection of the center depth can be easily obtained.

2.1 Foreground/Background video segmentation

First, we produce the foreground/background video segmentation. In this work, we propose to apply the Random Walker image segmentation algorithm [**?**] to the spatiotemporal domain with ZCam depth. Fig. 4 represents a description of video segmentation using Random Walker. Assuming that the segment of $n-1$ frame is initially known, we can construct a graph model such as Fig. 4. Since we know the segment of $(n-1)$-th frame and the foreground pixels of n-th frame from ZCam depth (or the left-/right-projected depth from ZCam depth), the segment of n-th frame can be computed using the Random Walker algorithm. An example of our segmentation result is shown in Fig. 5.

If the quality of the extracted object boundaries is poor, an user can add boundary refining step using Graph Cuts (GC) [**?**][**?**] using Foreground/Background color models, after generating a trimap, similar to the Grabcut algorithm [**?**].

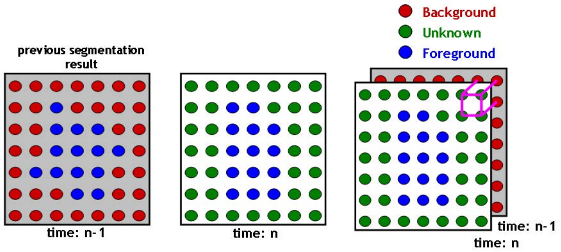


Figure 4. A description of video segmentation using random walker.

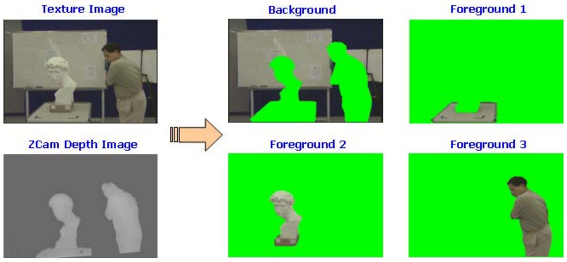


Figure 5. An example of our segmentation result. Three foregrounds are divided from an image.

2.2 Disparity extraction using segment-based stereo

Disparity is extracted using a new segment-based stereo algorithm according to each foreground/background segment. The stereo algorithm is applied twice for left/center and center/right cameras, using simple pairwise stereo model. Given a segmented image, we produce the over-segmentation results using mean-shift color segmentation algorithm [**?**]. In this process, the graph nodes become the over-segmented regions s instead of pixels and the parameters are the plane coefficients f, not pixel disparities. The plane coefficients are calculated using the iterative plane-fitting method [**?**]. Our segment-based stereo algorithm is based on the pairwise energy function that is minimized by Belief Propagation (BP) [**?**]. The proposed energy function E is defined by

$$E = \sum_{s \in \mathbf{R}} D(s) + \sum_{(s,s') \in \mathbf{N}} V(s,s') - \sum_{(s,s') \in \mathbf{N}} C(s,s'), \quad (1)$$

where $\mathbf{R}$ is the set of all over-segmented regions in a segmented image. $\mathbf{N}$ is the set of all neighboring region pair (s, s'). In E, the first term D is data term that is defined as the Sum of Absolute Difference (SAD) of the region s

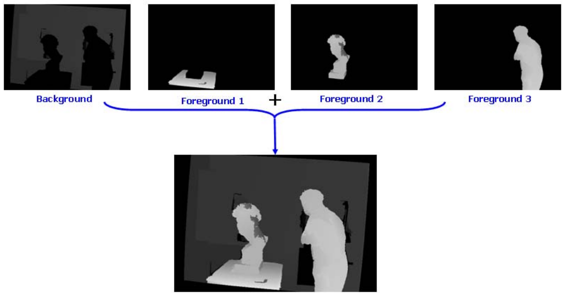


Figure 6. An example of our stereo disparity.

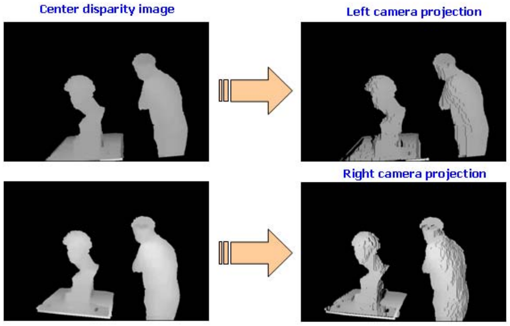


Figure 7. An example of disparity projection.

between two textured images. The second term V measures the smoothness:

$$V(s,s') = c_{ss'}\delta(f_s \neq f_{s'}), \tag{2}$$

where δ is the delta function and $c_{ss'}$ is proportional to the border length between two regions s,s'. It means that two neighboring regions with longer border length have a higher probability that they have similar disparities. The final term C is for the occlusion handling. With D term only, the sum of the costs within occluded region is included twice. By subtracting the SAD of the occluded region from D, we can solve this problem. The final parameter f is obtained by the minimization of (1) using BP. An example of our disparity extraction from segmented regions (Fig. 5) is shown in Fig. 6.

2.3 Object disparity extraction using projection of center disparity

Since the center camera (ZCam) produces the object depth map as well as texture image, the final depth of center camera has better performance than that of left/right camera. Therefore the left/right depth result is based on the projection of the foreground depth in the center image. Camera parameters can be obtained using basic Camera Calibration Toolbox [**?**]. Given this camera parameters, we can project the foreground depth of the center image into the left/right image such as Fig. 7.

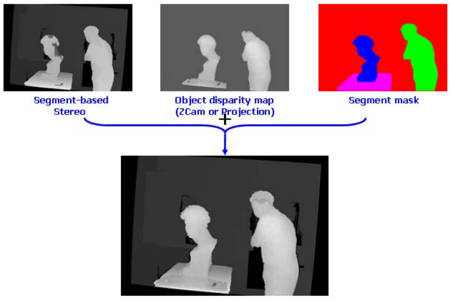
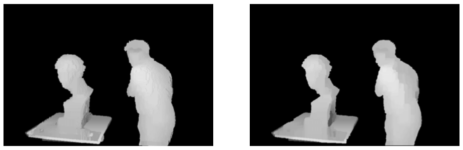


Figure 8. An example of disparity fusion.



(a) Before smoothing (b) Smoothing

Figure 9. An example of object disparity smoothing.

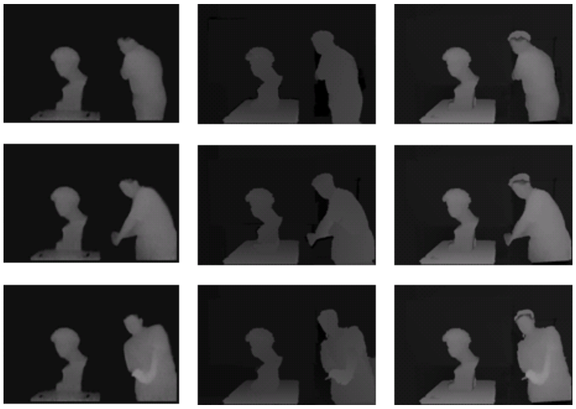
2.4 Disparity fusion and object smoothing

We assume that ZCam depth is generally better than the stereo results, because the stereo algorithm can not consider for large textureless regions. Therefore our fusion algorithm is based on the ZCam depth (or the projected depth from the results of center camera in case of left/right camera, Fig. 7). If any pixel do not have a depth from ZCam depth, we insert the depth value from the stereo results in Fig. 8. After that, we work the smoothness step using GC in Fig. 9. This step is needed for the assertion of smoothness inside the same segment after fusion process. Also in this experiment we assume that the background is unchanging. Thus, an uniform background depth for all frames is generated by averaging all background depths.

3 Experimental results

Our test video was provided by the Electronics and Telecommunications Research Institute (ETRI). This video has 120 frames and each frame has a resolution of $720 \times 486@30fps$.

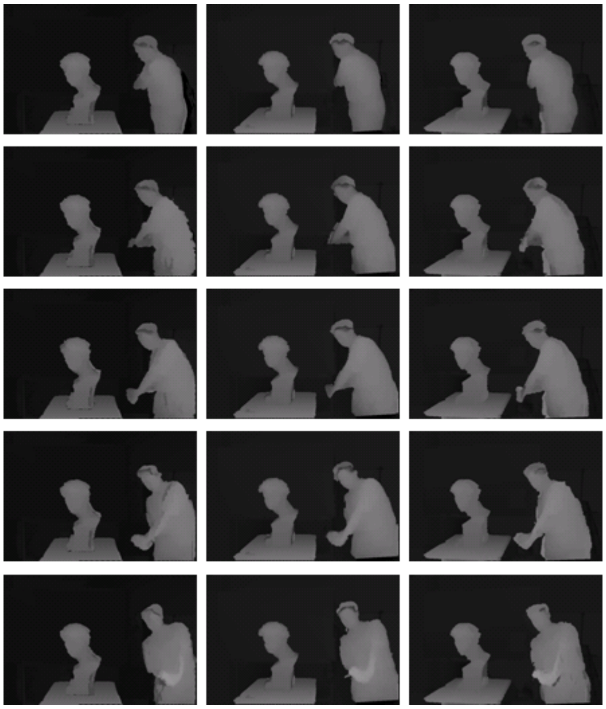
We compared our fusion results with the ZCam depth and our stereo results in Fig. 10. The ZCam produces a detailed depth of the foregrounds without the background.



(a) ZCam (b) Stereo (b) Our results

Figure 10. A comparison of our results with ZCam depth and Stereo results. We display the results from 1/60/120th images of the center camera.

The stereo result has a depth of the entire image, but it is difficult to represent the detailed foregrounds such as face and status. We show that our fusion result is a complement to these individual results. The final results of three-view system are shown in Fig. 11. Although the left/right camera is not depth sensing one, the results have good performance.



(a) Left (b) Center (b) Right

Figure 11. An sequence of resulted depth images. We display 1/30/60/90/120th results.

4 Conclusion

We have presented the three-view depth generation system that utilizes the foreground/background segmentation and the proposed stereo algorithm with occlusion constraints. Experiment results confirmed that our fusion results have the advantages of the ZCam and the stereo results. However our algorithm has the problem when the ZCam returns wrong depth. Therefore we need accurate comparison between the ZCam depth and the stereo results. This is one of our future works. Also we are researching more reliable stereo and fusion algorithms for more cluttered scenes with varying backgrounds.

Acknowledgments

This work was supported in part by the ETRI project by Electronics and Telecommunications Research Institute, Korea.

References

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[2] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[3] L. Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006.

[4] R. Gvili, A. Kaplan, E. Ofek, and G. Yahav. Depth key. In *In Proc. SPIE Electronic Imaging*, 2003.

[5] A. Klaus, M. Sormann, and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *In Proc. International Conference on Pattern Recognition*, pages 15–18, 2006.

[6] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.

[7] H. Tao, H. S. Sawhney, and R. Kumar. A global matching framework for stereo computation. In *In Proc. 9th International Conference on Computer Vision*, pages 532–539, 2001.

[8] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *In Proc. 7th International Conference on Computer Vision*, pages 666–673, 1999.