VIEW-CONSISTENT MULTI-VIEW DEPTH ESTIMATION FOR THREE-DIMENSIONAL VIDEO GENERATION

Sang-Beom Lee and Yo-Sung Ho

Gwangju Institute of Science and Technology (GIST)
261 Cheomdan-gwagiro, Buk-gu, Gwangju, 500-712, Korea
Telephone: +82-62-715-2258, Fax: +82-62-715-3164
E-mail: {sblee, hoyo}@gist.ac.kr

ABSTRACT

In this paper, we propose a new algorithm for view-consistent multi-view depth estimation for 3D video generation. After we obtain depth maps at the left and right viewpoints using a conventional depth estimation method, we project them into the center viewpoint and perform an error minimization process using a multi-view graph cut algorithm. Experimental results showed that the proposed algorithm improved view consistency of depth maps as well as rendering quality of the 3D video.

Index Terms **—**view consistency, multi-view depth estimation, three-dimensional video, 3DTV

1. INTRODUCTION

Recently, 3DTV has become very attractive as one of the next-generation broadcasting services [1]. With advances of 3D display devices, such as stereoscopic or autostereoscopic displays, 3DTV can provide users with a feeling of presence from the simulation of reality. In this decade, we expect that the technology will be progressed enough to realize the 3DTV system including content generation, coding, transmission, and display.

Figure 1 shows the conceptual framework of the 3DTV system that includes the whole processes of image acquisition, processing and coding, transmission, and rendering of 3D video with both N-view color and depth information [2]. We can capture 3D video using various types of cameras, such as stereoscopic camera, depth camera, or multi-view cameras. If the depth camera is available, we can acquire the depth video directly; otherwise, we should computationally estimate the depth data.

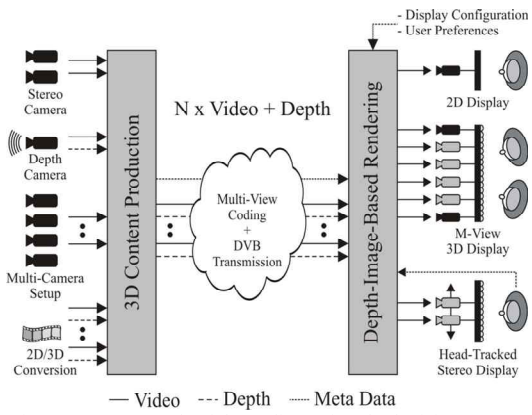


Figure 1. Framework of three-dimensional television system

The 3D video can be rendered by various types of displays, such as stereoscopic display, M-view 3D display, head-tracked stereo display, or even 2D display. It can be compatible with the conventional displays by selecting one view by user preferences. Given the increasing diversity of 3D services and displays, proper rendering techniques of 3D video, especially multi-view video, is required. If the number of multi-view video is smaller than that of the input viewpoints of 3D display, we cannot display the 3D contents properly. Furthermore, the interval between multi-view cameras is too far, viewers may feel visual discomfort while watching through the display.

In order to solve the above problems, we need to generate intermediate video at virtual viewpoints. An intermediate view is a synthesized image at the virtual viewpoint positioned between two real cameras. By synthesizing intermediate views, the natural rendering of the 3D video is possible and we can reduce the discomfort of the viewers.

In order to synthesize intermediate views at virtual viewpoints, we need depth information. Several works have been carried out for acquiring 3D depth information [3]. In particular, 3D video activities in Moving Picture Experts Group (MPEG) recognized the importance of a multi-view video and the corresponding multi-view depth video and they tried to develop depth estimation and view synthesis techniques [4]. Recently, they have distributed graph cut-based depth estimation software [5].

However, the graph cut-based depth estimation algorithm has several problems, such as boundary mismatch, textureless regions, or wide baseline. Especially, since this algorithm performs the depth estimation for each frame and viewpoint independently, the depth video is view- and temporally inconsistent. Those inconsistency problems of the depth video lead to flickering artifacts near the object boundary and degrade visual quality of synthesized views.

In this paper, we propose a new algorithm for view-consistent multi-view depth estimation. The main contribution of this paper is to exploit the total error minimization process. While the conventional algorithm estimates the depth map for each viewpoint independently, the proposed method computes depth maps at the left and right viewpoints and obtains multi-view depth maps at three viewpoints simultaneously by the total error minimization process using the graph cut algorithm.

2. RELATED WORKS

The 3D video coding subgroup in Moving Picture Experts Group (MPEG) distributed a graph cut-based depth estimation software to obtain multi-view depth information. The algorithm was originally proposed to compute the disparity map for stereoscopic images. However, it is not adequate that we apply the same algorithm to test sequences that 3D video coding subgroup provided. In other words, the test sequences are extended from still image to video and from the stereoscopic view to multi-view.

978-1-4244-6378-7/10/$26.00 ©2010 IEEE

As a result, the graph cut-based depth estimation algorithm causes the view and the temporal inconsistency problem.

2.1 Temporal Consistency Enhancement

The temporal inconsistency problem is caused by independent depth estimation for each frame. Ideally, in the case of the static objects or background, depth values are identical for each frame. However, since the depth estimation algorithm is independently operated for each frame, depth values are fluctuated. The temporal inconsistency problem of the depth video propagates the synthesized views and it leads the flickering artifacts discomforting viewers. Furthermore, since it degrades the performance of the temporal prediction of the depth video coding, the coding efficiency is also decreased.

Recently, a new algorithm for improving temporal consistency was proposed [6].This algorithm adds a temporal weighting function that refers to the previous disparities when estimating the current disparities to the conventional matching function. In addition, this algorithm applies different temporal weighting function for the case of the 3D scene: static and moving area. The temporal weighting function is defined by

$$E_{temp}(x,y,d)=\begin{cases}E_{temp\_static}(x,y,d) & if \quad MAD_k < Th\\ E_{temp\_moving}(x,y,d) & else\end{cases} \tag{1}$$

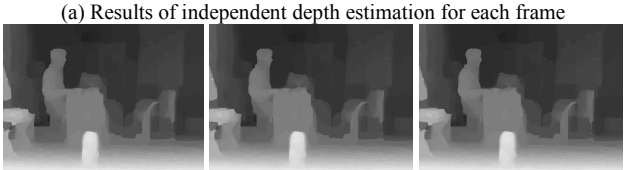
$$E_{temp\_static}(x,y,d)=\gamma\,|\,d-D_{prev}(x,y)\,| \tag{2}$$

$$E_{temp\_moving}(x,y,d)=\gamma\,|\,d-D_{prev}(x+\Delta x,y+\Delta y)\,| \tag{3}$$

where γ represents the slope of the weighting function and $D_{prev}(x,y)$ represents the previous disparity at (x,y). MAD_k represents the MAD of the k-th block including the position (x,y) and Th represents the threshold. $(\Delta x,\Delta y)$ represents the motion vector for the pixel at (x,y).

Figure 2 represents the consecutive three frames of depth video after applying temporal consistency enhancement. As shown in Fig. 2(b), we notice that the depth video becomes temporally inconsistent.

(a) Results of independent depth estimation for each frame



(b) Results of temporal consistency enhancement

Figure 2. Temporally consistent depth video

2.2 View Consistency Enhancement

The view inconsistency problem is caused by the independent depth estimation for each viewpoint. Theoretically, if the intervals of multi-view camera and the optical axes of cameras are identical, objects have the same depth values for each viewpoint. However, since the depth estimation algorithm is independently operated for each viewpoint, the depth values are different and this problem decreases quality of the synthesized views.

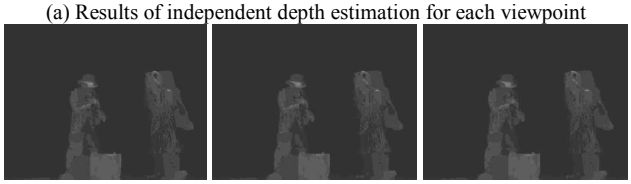
In order to solve the view inconsistency problem, a new algorithm for improving the view consistency was proposed [7]. This algorithm adds a view weighting function that refers to the disparities of the reference viewpoint when estimating the disparities of the target viewpoint to the conventional matching function. The view weighting function is defined by

$$E_{view}(x,y,d)=\left|d-D_{i'}(x,y)\right| \tag{4}$$

where $D_i(x,y)$ represents the disparity of the reference viewpoint.

Figure 3 shows the depth map of adjacent three viewpoints. As shown in Fig. 3(b), we notice that the view inconsistency problem is reduced.

(a) Results of independent depth estimation for each viewpoint



(b) Results of view consistency enhancement

Figure 3. View-consistent depth map

Finally, the matching cost for depth estimation is defined by

$$E_{data}(x,y,d)=E_{sim}(x,y,d)+E_{temp}(x,y,d)+E_{view}(x,y,d) \tag{5}$$

$$E_{sim}(x,y,d)=\min\{\left|I_C(x,y)-I_L(x+d,y)\right|,\left|I_C(x,y)-I_R(x-d,y)\right|\} \tag{6}$$

where $E_{sim}(x,y,d)$ represents the intensity difference for the center and the left or the center and the right viewpoints.

3. PROPOSED VIEW-CONSISTENT MULTI-VIEW DEPTH ESTIMATION

The view consistency enhancement algorithm refers to the disparities of the reference viewpoint when estimating the disparities of the target viewpoint. However, if there are errors in the depth map of that viewpoint, it is not desired that the whole depth values including errors are used. Practically, in the case that the whole depths are referred, the visual quality of the synthesized view is degraded since the depth errors cause near the boundaries.

(a) No depth reference (b) Depth reference (c) Different image

Figure 4. Errors of depth reference

Figure 4 shows the synthesized views obtained by the depth reference including depth errors. As shown in Fig. 4(b), the errors near the bag are caused by the depth errors. It means that the visual quality with depth reference is lower than that without it.

Therefore, we propose a view-consistent multi-view depth estimation algorithm. We first compute disparities and perform the error minimization for the left and the right viewpoints. Then, we project each pixel of those reference viewpoints onto the center viewpoint by using already obtained disparities. The projected pixels are used as additional nodes for error minimization. The total error minimization is performed in the last step using data costs of three viewpoints. Figure 5 shows the block diagram of the proposed algorithm.

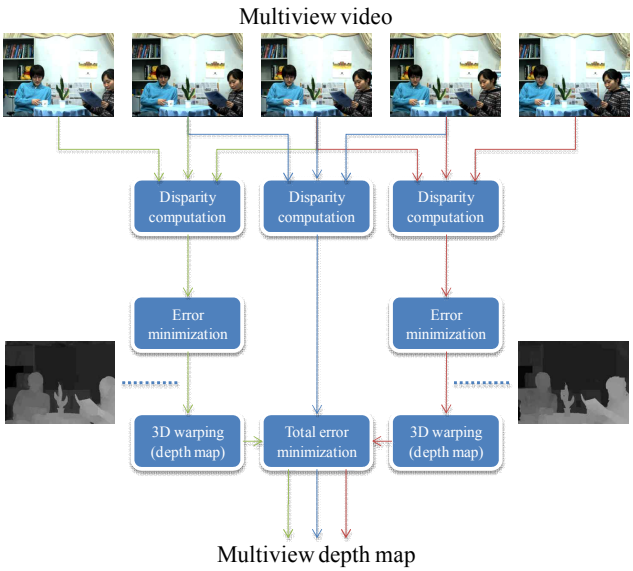


Figure 5. Block diagram of the proposed algorithm

Figure 6 shows total error minimization structure by graph cut algorithm. The 2D lattice is constructed by whole pixels of the center viewpoint as shown in Fig. 6 and the projected pixels of left and right viewpoints are linked as additional nodes. We notice that disparities that link pixels of the center viewpoint and the left or the right viewpoint are already obtained by the conventional algorithm. Finally, we perform the total error minimization using the 3D lattice.

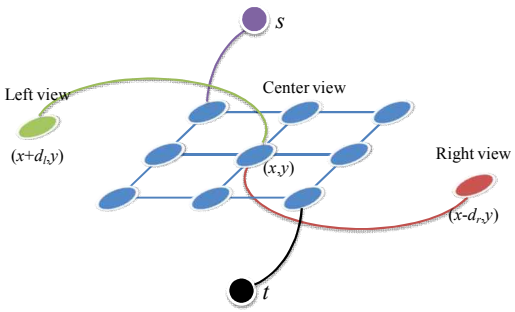


Figure 6. Total error minimization by multi-view graph cut

From the modified 3D lattice of graph, the modified matching function is defined by

$$E(x,y,d) = E_{data}(x,y,d) + E_{smooth}(x,y,d) \quad (7)$$

$$E_{smooth}(x,y,d) = \lambda \sum_{(x_i,y_i)\in N_p} \left| D(x,y,d) - D(x^i,y^i,d) \right| \quad (8)$$

where $E_{data}(x,y,d)$ is explained in Section 2 and λ represents the weighting factor. N_p represents the neighbor pixels of (x,y). While N_p is 4-neighbor pixels for the conventional algorithm, we add two more pixels projected from left and right viewpoint.

4. EXPERIMENTAL RESULTS

In order to evaluate the proposed algorithm for improving view consistency, we used 'Pantomime' sequence provided by Nagoya University [8] and 'Newspaper' sequence provided by Gwangju Institute of Science and Technology (GIST) [9]. For the evaluation of the accuracy of the depth map, we performed depth estimation for three views and virtual view synthesis for the center view using the left and the right views. Then, we compare the original color image to the synthesized image in terms of PSNR. We first estimated the depth map for the center viewpoint and obtained depth maps of the left and the right viewpoint by referring to the center depth map. Table 1 represents the experiment conditions.

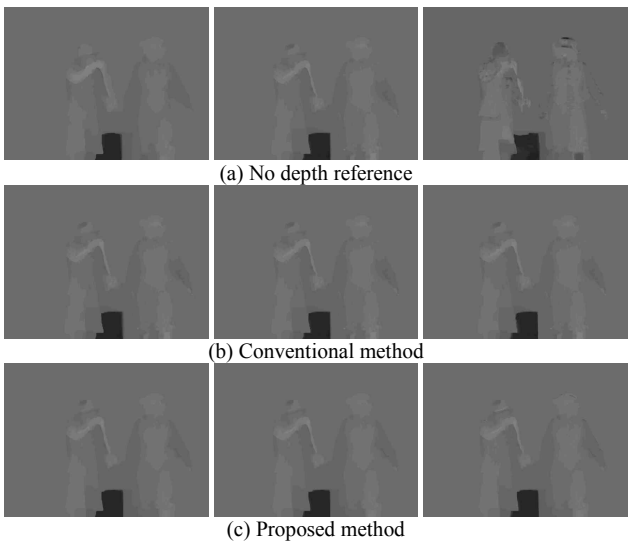
Table 1. Experiment conditions

Sequence	Pantomime	Newspaper
View numbers for depth estimation	37, 39, 41	4, 5, 6
View numbers for view synthesis	39	5
Frames	100	

Figure 7 and Figure 8 show the depth estimation results. Figure 7(a) and Figure 8(a) shows the depth map of the reference viewpoints. As shown in Fig. 7 and Fig. 8, the depth maps using depth reference gave us a higher view consistency.

Figure 9 and Figure 10 show the view synthesis results. As shown in Fig. 9(d) and Fig. 10(d), the proposed method using the total error minimization improved the visual quality compared to the previous method, while the result of the conventional method have errors near object boundaries as shown in Fig. 9(c) and Fig. 10(c).

Table 2 represents the average PSNR of the synthesized image. From the experimental results, we noticed that the proposed method improved the average PSNR of the synthesized image on average compared to the conventional method.

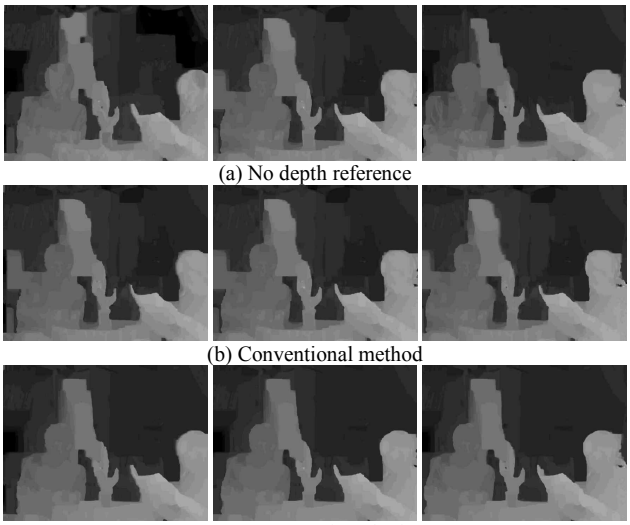


(a) No depth reference

(b) Conventional method

(c) Proposed method

Figure 7. Depth estimation results for 'Pantomime'

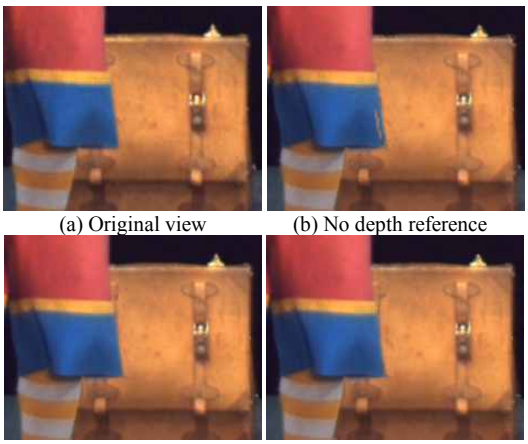


(a) No depth reference

(b) Conventional method

(c) Proposed method

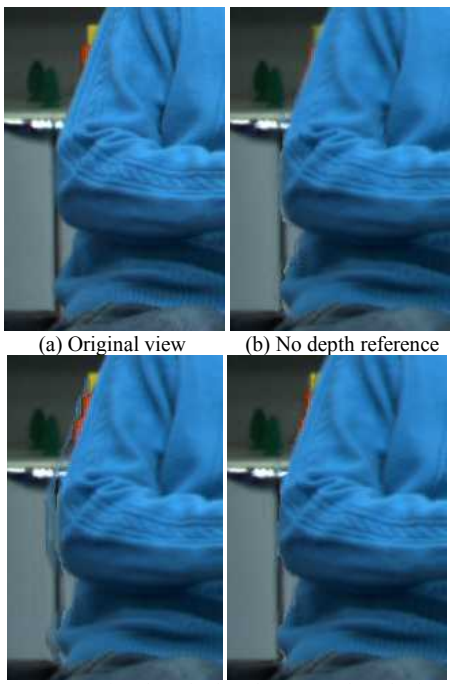
Figure 8. Depth estimation results for 'Newspaper'



(a) Original view (b) No depth reference

(c) Conventional method (d) Proposed method

Figure 9. View synthesis results for 'Pantomime'



(a) Original view (b) No depth reference

(c) Conventional method (d) Proposed method

Figure 10. View synthesis results for 'Newspaper'

Table 2. Average PSNR of synthesized views

Sequence	Average PSNR (dB)		
	No depth reference	Conventional method	Proposed method
Pantomime	34.6441	35.7680	35.3720
Newspaper	32.3933	30.3536	32.6432

5. CONCLUSIONS

In this paper we have proposed the view-consistent multi-view depth estimation. In order to improve the view consistency, we exploited the total error minimization using three viewpoints simultaneously. The total error minimization was performed by using four neighboring pixels plus two more pixels projected from left and right viewpoint. As a result, we realized the view-consistent multi-view depth map and more natural rendering of 3D video by improving visual quality of synthesized views. The rendering quality was increased by 0.4889 dB on average compared to the work without depth reference and by 0.9468 dB on average compared to the previous depth reference method.

6. ACKNOWLEDGEMENT

This research was supported in part by MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by NIPA (National IT Industry Promotion Agency) (NIPA-2010-(C1090-1011-0003))

7. REFERENCES

[1] A. Smolic, K. Muller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, T. Wiegand, "3D Video and Free Viewpoint Video - Technologies, Applications and MPEG Standards," IEEE International Conference on Multimedia and Expo (ICME), pp. 2161-2164, July 2006.

[2] ISO/IEC JTC1/SC29/WG11, "Applications and Requirements on FTV," N9466, Oct. 2007.

[3] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, R. Szeliski, "High-quality Video View Interpolation Using a Layered Representation," SIGGRAPH, pp. 600-608, Aug. 2004.

[4] ISO/IEC JTC1/SC29/WG11, "Call for Contributions on 3D Video Test Material," N9595, Jan. 2008.

[5] ISO/IEC JTC1/SC29/WG11, "Reference Software of Depth Estimation and View Synthesis for FTV/3DV," M15836, Oct. 2008.

[6] S. Lee, Y. Ho, "Temporally Consistent Depth Map Estimation Using Motion Estimation for 3DTV," International Workshop on Advanced Image Technology, pp. 149(1-6), Jan. 2010.

[7] ISO/IEC JTC1/SC29/WG11, "Depth Estimation Reference Software (DERS) 5.0," M16923, Oct. 2009.

[8] ISO/IEC JTC1/SC29/WG11, "1D Parallel Test Sequences for MPEG-FTV," M15378, April 2008.

[9] ISO/IEC JTC1/SC29/WG11, "Multi-view Video Test Sequence and Camera Parameters," M15419, April 2008.